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

'You cannot open a book without learning something.'
— Confucius, 551 BC – 479 BC

Chapter 1

Introduction

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

Since twenty years ago, the high energy physics community has been considering a next-generation electron−positron collider after the Large Hadron Collider (LHC). Measurements from the LHC helped to establish a Standard Model of particle physics. Yet there are issues that Standard Model could not explain. For example, the origin of the masses of neutrinos and the particles account for cosmic dark matter are questions that need to be addressed. Precision measurements from a next-generation electron−positron collider will hopefully provide answers to some of these questions.

The advantage of an e^+e^- linear collider over a hadron collider include: few background events from photon−photon collisions; a similar rate of pairs production of all particles as photon couples to all particles equally; smaller theoretical uncertainties on the electroweak interaction; and full reconstruction of events.

The International Linear Collider (ILC) [1], and the Compact Linear Collider (CLIC) [2], are two most promising candidates of the next-generation electron−positron collider. The ILC is capable of operating at centre-of-mass energies from 250 GeV to 500 GeV. CLIC can reach centre-of-mass energies from 350 GeV to 3 TeV. Both colliders will be able to measure Higgs couplings precisely via processes like $e^+e^- \to ZH$ and measure top quark mass and couplings via processes such as $e^+e^- \to t\bar{t}$.

The optimisation of the design of the detectors for the future linear colliders is crucial to improve the ability to reconstruct events. By reconstructing the event to the

per-particle level, an event can be studied in detail. At the same time, physics simulation studies are important to demonstrate the physics read of the future linear collider.

chapter 2 starts with an overview of Standard Model of particle physics, including brief discussions on 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 is dedicated to the discussion on identifying a Higgs boson from vector bosons using the tau pair decay channel.

In chapter 3, the detector designs currently considered for two future electron-positron linear colliders, the ILC and CLIC, are described. After a short introduction of the two colliders, the physics programme for these future colliders is discussed, followed by the impact of physics requirements on the detector design. Afterwards, the International Large Detector (ILD), one detector option for the ILC, is discussed in details, followed by overviews on each sub-detector in the ILD. The chapter finishes with a discussion on the modified ILD detector concept for CLIC.

In chapter 4, the software for event simulation, event reconstruction, and event analysis is discussed. The event reconstruction focus on PandoraPFA, a world-leading pattern-recognition software for particle flow calorimetry. PandoraPFA aims to reconstruct individual particles in the event. Jet algorithms is also used, followed by a discussion on multivariate analysis, where different fitting models, optimisation, and overfitting are discussed. The multivariate analysis is used extensively in the chapter 5, chapter 6, and chapter 7.

Chapter 5 describes several new PandoraPFA algorithms for photon reconstruction. These algorithms reconstruct photons and address the issues of the photon fragments in the event reconstruction. After presenting the algorithms, the performances of these algorithms are provided using photon samples and jet samples.

In chapter 6, a classification of the tau lepton decay modes is presented. The analysis contains the event generation, simulation, reconstruction, and the use of the multivariate classifier for the classification. The performance of the tau decay mode classification is given, followed by an electromagnetic calorimeter optimisation study based on the tau decay mode classification. Lastly, the tau decay mode classification is used in a proof-of-principle analysis to demonstrate the ability to use the tau pair polarisation

correlation as a signature for Higgs boson using the tau pair decay process, where both $\tau^- \to \pi^- \nu_\tau$.

In chapter 7, a full CLIC_ILD detector simulation study is 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 review of the Standard Model of Particle Physics, with an emphasis on the Higgs mechanism and the Higgs boson. A general parametrisation of the Higgs theory is discussed, which supplies the theoretical background for the physics analysis in chapter 7. Lastly a discussion of the usage of the tau pair polarisation correlations as a signature of Higgs boson is presented, which motivates the study in chapter 6.

2.1 Overview of the Standard Model

The Standard Model (SM) [6–9] is a quantum field theory concerning the fundamental particles and three fundamental interactions of nature: the electromagnetic; the weak; and the strong interactions. The fundamental particles in the SM consist of bosons and fermions. The bosons mediate the fundamental forces between particles: 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. The properties of the force-exchange bosons and Higgs boson are listed in table 2.1.

The other fundamental particles are spin-$\frac{1}{2}$ fermions. For each fermion in the SM, there is an anti-fermion with the same mass and spin, but opposite charge. These

Force	Boson	Mass	Spin	Charge / e
Electromagnetic	photon	0	1	0
Weak	W^+	80.385(15) GeV	1	1
	W^-	80.385(15) GeV	1	−1
	Z	91.1876(21) GeV	1	0
Strong	gluon	0	1	0
-	Higgs	125.1(3) GeV	0	0

Table 2.1: Masses, spins, and charges of fundamental bosons in the SM. Values are taken from [6].

fermions have three generations. Each generation of fermions has the same interaction property, but different masses. Experimental evidences of three generations include the measurements of the Z boson decay-width, which strongly suggested three generations of neutrinos [10].

These fermions come in two distinct categories: leptons and quarks. The neutral leptons (the neutrinos) only experience the weak forces. The charged leptons ($e^{\pm}$, $\mu^{\pm}$, $\tau^{\pm}$) experience the weak forces and the electromagnetic forces. Quarks experience all three fundamental forces described by the SM. Properties of these fermions are listed in table 2.2.

Many SM predications have been experimentally verified. Some recent highlights include the discovery of the top quark in 1995 [11], the tau neutrino in 2000 [12], and the Higgs boson in 2012 [3, 4]. However, there are observations which are not explained by the SM. One issue is that the SM does not incorporate the gravitational force. Another issue is that the SM does not natively allow neutrino masses and mixings. The SM also does not explain the existence of the dark matter. There are many theories beyond the Standard Model (BSM) trying to provide an explanation for these issues. One example is the generalisation of the Higgs theory to allow non-SM coupling strengths [13, 14].

2.2 Quantum electrodynamics

QED is a quantum gauge field theory explaining electromagnetic interactions. Quantum field theory (QFT) is the theoretical framework for constructing quantum mechanical

Type	Generation	Fermion	Mass	Charge / e
Lepton	1	e^-	$0.548579909070(16)\,\text{MeV}$	-1
		ν_e	-	0
	2	μ^-	$105.6583745(24)\,\text{MeV}$	-1
		ν_μ	-	0
	3	τ^-	$1776.86(12)\,\text{MeV}$	-1
		ν_τ	-	0
Quark	1	u	$2.2^{+0.6}_{-0.4}\,\text{MeV}$	$+\frac{2}{3}$
		d	$4.7^{+0.5}_{-0.4}\,\text{MeV}$	$-\frac{1}{2}$
	2	c	$1270 \pm 30\,\text{MeV}$	$+\frac{2}{3}$
		s	$98^{+8}_{-4}\,\text{MeV}$	$-\frac{1}{3}$
	3	t	$173210 \pm 510 \pm 710\,\text{MeV}$	$+\frac{2}{3}$
		b	$4180^{+40}_{-30}\,\text{MeV}$	$-\frac{1}{3}$

Table 2.2: Masses and charges of the fundamental fermions in the SM. All fermions are spin-$\frac{1}{2}$ particles. For each fermion in the SM, there is an anti-fermion with the same mass and spin, but opposite charge. Neutrinos are known to have non-zero mass from the observation of neutrino flavour oscillations. The upper bound on the neutrino mass is 2 eV. For the top quark mass, the statistical uncertainties is listed first, followed by systematic uncertainties. Values are taken from [6].

models of fundamental particles. Particles are treated as excited states of the underlying physical field in the QFT. A gauge theory is a type of field theory in which the Lagrangian is invariant under a continuous group of local transformations. Gauge invariance or gauge symmetry refers to when a field is transformed, but the Lagrangian is not.

QED is an abelian gauge theory with the U(1) symmetry group. The gauge field, which mediates the interaction between the charged spin-$\frac{1}{2}$ fields, is the electromagnetic field, denoted A^{μ}. The QED Lagrangian [15] for a spin-$\frac{1}{2}$ field interacting with the electromagnetic field is given by:

$$\mathcal{L}_{QED} = \overline{\psi}\left(i\gamma^{\mu}D_{\mu} - m\right)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}, \tag{2.1}$$

where ψ is the spin-$\frac{1}{2}$ Dirac field satisfying the Dirac equation, given by the Lagrangian density:

$$\mathcal{L} = \overline{\psi}\left(i\gamma^{\mu}D_{\mu} - m\right)\psi, \tag{2.2}$$

where the γ^{μ} are the Dirac gamma matrices with $\mu \in \{0, 1, 2, 3\}$; $\overline{\psi}$ is defined as $\psi^{\dagger}\gamma^{0}$; $F_{\mu\nu} = \partial_{\mu}A_{\nu} - \partial_{\nu}A_{\mu}$ is the electromagnetic field tensor; e is the coupling constant, which is equal to the electric charge; m is the mass of the electron; and the gauge covariant derivative is given by:

$$D_{\mu} \equiv \partial_{\mu} + ieA_{\mu}, \tag{2.3}$$

where A_{μ} is the covariant four-vector potential of the electromagnetic field generated by the electron itself.

2.3 Quantum chromodynamics

Quantum chromodynamics (QCD) is the quantum field theory of strong interactions. QCD theory is invariant under local non-Abelian SU(3) transformations. There are eight gauge bosons, the gluons, corresponding to the eight ($8 = 3^2 - 1$) generators of the SU(3) symmetry group. Gluons carry colour charges. There are three types of colour charge, sometimes labelled as red, green, and blue. Anti-particles carry anticolour. A Quarks are associated with a single colour. Gluons are made up of a colour and an anticolour

(or superposition of colour−anticolour pair). The QCD Lagrangian is given by:

$$\mathcal{L}_{QCD} = \sum_{f\in u,d,s,c,b,t} \overline{\psi_i} \left(\left(i\gamma^\mu \partial_\mu - g_s \gamma^\mu G^a_\mu \frac{\lambda^a}{2} \right)_{ij} - m_f \delta_{ij} \right) \psi_j - \frac{1}{4} G^a_{\mu\nu} G^{a\mu\nu}, \tag{2.4}$$

where ψ represents a quark with a colour charge, indicated by i or j; m controls the mass of the quark; g_s is the strong coupling constant; a is the colour charge; λ^a represents one of the eight Gell-Mann matrices; and $G^a_{\mu\nu}$ represents the gauge invariant gluon field strength tensor, given by:

$$G^a_{\mu\nu} = \partial_\mu \gamma^a_\nu - \partial_\nu \gamma^a_\mu - g_s f_{abc} G^b_\mu G^c_\nu, \tag{2.5}$$

where G^b_μ is the gluon field with colour charge b.

2.4 The electroweak interaction

The electroweak interaction can be thought as an extension to QED to incorporate the weak force, the force describing nuclear radioactive decay. The unification of the electromagnetic and the weak force is accomplished under an $\mathrm{SU(2)_L} \times \mathrm{U(1)}$ gauge symmetry group. The corresponding gauge bosons are the three W bosons (W^1, W^2, and W^3) from $\mathrm{SU(2)_L}$ gauge symmetry , and the B boson from U(1) gauge symmetry. All gauge bosons are initially massless. Fermion mass terms are forbidden under $\mathrm{SU(2)_L}$ gauge symmetry.

The electroweak Lagrangian can be written as

$$\mathcal{L}_{Electroweak} = \mathcal{L}_{Boson} + \mathcal{L}_{Fermion} + \mathcal{L}_{Higgs} + \mathcal{L}_{Yukawa}\,. \tag{2.6}$$

First consider $\mathcal{L}_{Boson}$, given by:

$$\mathcal{L}_{Boson} = -\frac{1}{4} W^i_{\mu\nu} W^{i\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu}, \tag{2.7}$$

$$W^i_{\mu\nu} = \partial_\nu W^i_\mu - \partial_\mu W^i_\nu - g\varepsilon^{ijk} W^j_\mu W^k_\nu, \tag{2.8}$$

$$B_{\mu\nu} = \partial_\nu B_\mu - \partial_\mu B_\nu, \tag{2.9}$$

where the B field is invariant under U(1) transformations; the W field is invariant under non-Abelian SU(2) transformations; and the indices, i, j, and k, indicate three W fields.

The term $\mathcal{L}_{Fermion}$ describes the massless fermion fields coupling to the fermions, and the propagation of the fermion fields. The left-handed (ψ_L) and the right-handed fermions (ψ_R) are treated differently. The right-handed fermions are singlets. The left-handed fermions are in doublets with the corresponding fermions of the same generation. The term $\mathcal{L}_{Fermion}$ is given by:

$$\mathcal{L}_{Fermion} = \sum_{\psi \in fermions} \overline{\psi}_L \gamma^\mu D_\mu^L \psi_L + \overline{\psi}_R \gamma^\mu D_\mu^R \psi_R, \tag{2.10}$$

where covariant derivatives D_μ^L and D_μ^R are defined as

$$D_\mu^L = \partial_\mu + ig\frac{\tau_i}{2}W_\mu^i + ig' Y_\psi B_\mu, \tag{2.11}$$

$$D_\mu^R = \partial_\mu + ig' Y_\psi B_\mu. \tag{2.12}$$

The structure of 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) and Y_ψ is the hypercharge associated with the fermion field ψ. The W field couples with strength g to the fermion field. The B field couples with strength g' to the particles carrying weak hypercharge Y.

The term $\mathcal{L}_{Higgs}$ describes the Higgs field. After electroweak symmetry breaking of the Higgs field, the mass terms of the gauge bosons are introduced. The term $\mathcal{L}_{Yukawa}$ produces the mass terms of the quarks and charged leptons. Firstly a general spontaneous symmetry breaking mechanism is provided, followed by a description of the electroweak symmetry breaking.

2.4.1 Spontaneous symmetry breaking

Consider a complex scalar field, with the Klein-Gordon Lagrangian:

$$\mathcal{L} = \partial^\mu \psi^* \partial_\mu \psi - m^2 |\psi|^2 = \partial^\mu \psi^* \partial_\mu \psi - V(\psi), \tag{2.13}$$

where m is the mass term and $V(\psi)$ is the potential of the field ψ. This Lagrangian has a global symmetry $\psi \to e^{i\phi}\psi$. The potential can be modified to add an interaction term

without breaking the invariance of the global symmetry:

$$V(\psi) = m^2|\psi|^2 + \lambda|\psi|^4, \tag{2.14}$$

where λ controls the interaction strength. This modified potential has a minimum at $|\psi| = 0$ for $m^2 > 0$. However, if $m^2 < 0$, the minimum of the potential occurs when:

$$\frac{\partial V(\psi)}{\partial|\psi|} = 2m^2|\psi| + 4\lambda|\psi|^3 = 0, \tag{2.15}$$

leading to a non-negative expectation value for the field:

$$|\psi| = \sqrt{\frac{-m^2}{2\lambda}} \equiv \frac{\nu}{\sqrt{2}}, \tag{2.16}$$

where $\nu = \sqrt{-m^2/\lambda}$. The solution that minimises the potential is not unique; it corresponds to a circle of points in the complex ψ plane. By choosing any one of these points, which are degenerate in energy, the symmetry of $\psi \to e^{i\phi}\psi$ is broken. This phenomenon is known as the spontaneous symmetry breaking.

A consequence of spontaneous symmetry breaking is that the perturbation of the field along the degenerate energy direction, which is the circle in complex ψ plane, have no associated potential energy. This is formalised as Goldstone's theorem [16, 17]. The theorem states that spontaneous symmetry breaking always implies the existence of a massless particle.

To demonstrate Goldstone's theorem, using the Lagrangian in equation 2.13 as an example, after the spontaneous symmetry breaking of the field, the perturbation of the field ψ near the field minimum point can be written as

$$\psi = \frac{1}{\sqrt{2}}(\nu + \psi_1 + i\psi_2), \tag{2.17}$$

where $\nu = \sqrt{-m^2/\lambda}$ refers to the minimum point in the potential, and ψ_1 and ψ_2 are real scalar fields. Substituting ψ in the Lagrangian in equation 2.13 gives:

$$\mathcal{L} = \frac{1}{2}\partial^\mu\psi_1\partial_\mu\psi_1 + \frac{1}{2}\partial^\mu\psi_2\partial_\mu\psi_2 - m^2\psi_1^2 + \dots. \tag{2.18}$$

The mass term for the ψ_1 field is $\sqrt{-m^2}$ whereas there is no mass term for the ψ_2 field, as stated by the Goldstone's theorem.

In the previous example, the Lagrangian in equation 2.13 possesses the global symmetry of $\psi \to e^{i\phi}\psi$. Instead, if there is a local U(1) gauge symmetry of $\psi \to e^{i\phi(x)}\psi$, this implies a corresponding field A_μ, which transforms as $A_\mu \to A_\mu - \partial_\mu\phi(x)$. For the gauge invariance, the covariant derivative becomes $D_\mu = \partial_\mu + ieA_\mu$. Hence the Lagrangian in equation 2.13 becomes:

$$\mathcal{L} = (D^\mu\psi)^*(D_\mu\psi) - m^2|\psi|^2 - \lambda|\psi|^4. \tag{2.19}$$

When the field is expanded around the minimum of the potential, $\nu = \sqrt{-m^2/\lambda}$, with $m^2 < 0$, the gauge boson mass term:

$$+\frac{e^2\nu^2}{2}A^\mu A_\mu, \tag{2.20}$$

is obtained from the $(D^\mu\psi)^*(D_\mu\psi)$ term in the Lagrangian. Therefore the spontaneous symmetry breaking of a gauge field gives rise to a gauge boson mass.

2.5 Higgs Mechanism

The Higgs mechanism is an extension of the example of the spontaneous symmetry breaking introduced in the previous section. It can provide mass terms for bosons and fermions that are compatible with the gauge invariance of the SM. A complex scalar Higgs field, Φ_H, transforms as a doublet of SU(2) with hypercharge $Y = \frac{1}{2}$. The Higgs Lagrangian is given by:

$$\mathcal{L}_{Higgs} = \left(D_\mu\Phi_\mathrm{H}\right)^\dagger \left(D^\mu\Phi_\mathrm{H}\right) - \mu^2\Phi_\mathrm{H}^\dagger\Phi_\mathrm{H} + \lambda\left(\Phi_\mathrm{H}^\dagger\Phi_\mathrm{H}\right)^2, \tag{2.21}$$

where λ and μ are constants. The $\mathrm{SU(2)_L} \times \mathrm{U(1)}$ symmetry of the electroweak Lagrangian demands that the covariant derivative of the Higgs field takes the form

$$D_\mu = \left(\partial_\mu + ig\frac{\tau_i}{2}W_\mu^i + ig'\frac{1}{2}B_\mu\right), \tag{2.22}$$

where g is the coupling constant of the $\mathrm{SU(2)_L}$ gauge symmetry; g' is the coupling constant of the U(1) gauge symmetry; and the τ_i are Pauli matrices. The Higgs potential is given by:

$$V(\mathrm{H}) = \mu^2\Phi_\mathrm{H}^\dagger\Phi_\mathrm{H} - \lambda\left(\Phi_\mathrm{H}^\dagger\Phi_\mathrm{H}\right)^2. \tag{2.23}$$

The Higgs potential is minimised when

$$\sqrt{\Phi_{\mathrm{H}}^{\dagger}\Phi_{\mathrm{H}}} = \frac{\nu}{\sqrt{2}} = \sqrt{\frac{\mu^2}{2\lambda}}. \tag{2.24}$$

By expanding the Higgs field about the minimum point of the potential, the non-zero vacuum expectation value (VEV) can be written as:

$$\left\langle \Phi_{\mathrm{H}} \right\rangle = \begin{pmatrix} 0 \\ \frac{\nu}{\sqrt{2}} \end{pmatrix}, \tag{2.25}$$

with a real ν. Substituting the Higgs VEV into the $\mathcal{L}_{Higgs}$ in equation 2.21, the $\left(D_\mu \Phi_{\mathrm{H}}\right)^{\dagger}\left(D^\mu \Phi_{\mathrm{H}}\right)$ term becomes

$$-\frac{1}{8}\begin{pmatrix} 0 & \nu \end{pmatrix}\begin{pmatrix} gW_\mu^3 + g'B_\mu & g\left(W_\mu^1 - iW_\mu^2\right) \\ g\left(W_\mu^1 + iW_\mu^2\right) & -gW_\mu^3 + g'B_\mu \end{pmatrix}\begin{pmatrix} gW_\mu^3 + g'B_\mu & g\left(W_\mu^1 - iW_\mu^2\right) \\ g\left(W_\mu^1 + iW_\mu^2\right) & -gW_\mu^3 + g'B_\mu \end{pmatrix}\begin{pmatrix} 0 \\ \nu \end{pmatrix}. \tag{2.26}$$

Ignoring the negative sign, the expression simplifies to

$$\frac{\nu^2 g^2}{8}\left(W_\mu^1 - iW_\mu^2\right)\left(W_\mu^1 + iW_\mu^2\right) + \frac{\nu^2}{8}\left(gW_\mu^3 - g'B_\mu\right)^2. \tag{2.27}$$

The physical fields W_μ^+ and W_μ^- can be identified with the first part of the equation 2.27, as

$$W_\mu^+ = \frac{1}{\sqrt{2}}\left(W_\mu^1 - iW^2\mu\right), \tag{2.28}$$

$$W_\mu^- = \frac{1}{\sqrt{2}}\left(W_\mu^1 + iW^2\mu\right). \tag{2.29}$$

The physical fields Z_μ and A_μ are associated with W_μ^3 and B_μ. Since the Z boson is massive and the photon is massless, the second part of the equation 2.27 should give rise to Z boson mass term only, with no mass term for photon. This can be achieved by rearranging the second part of the equation 2.27:

$$\frac{\nu^2}{8}\left(gW_\mu^3 - g'B_\mu\right)^2 = \frac{\nu^2\left(g^2 + g'^2\right)}{8}\left(\frac{g}{\sqrt{g^2 + g'^2}}W_\mu^3 - \frac{g'}{\sqrt{g^2 + g'^2}}B_\mu\right)^2. \tag{2.30}$$

A convenient way to connect g and g' is to use the Weinberg mixing angle [18], denoted as θ_{W}. The Weinberg mixing angle is defined as

$$\cos\left(\theta_{\mathrm{W}}\right) = \frac{g}{\sqrt{g^2 + g'^2}}, \tag{2.31}$$

$$\sin\left(\theta_{\mathrm{W}}\right) = \frac{g'}{\sqrt{g^2 + g'^2}}. \tag{2.32}$$

The equation 2.30 can be rewritten using the Weinberg mixing angle:

$$\frac{\nu^2\left(g^2 + g'^2\right)}{8}\left(\cos\left(\theta_{\mathrm{W}}\right)W_\mu^3 - \sin\left(\theta_{\mathrm{W}}\right)B_\mu\right)^2. \tag{2.33}$$

The physical field Z_μ can be immediately identified as:

$$Z_\mu = \cos\left(\theta_{\mathrm{W}}\right)W_\mu^3 - \sin\left(\theta_{\mathrm{W}}\right)B_\mu. \tag{2.34}$$

Consequently, the physical field A_μ with associated massless photon can be written as:

$$A_\mu = \sin\left(\theta_{\mathrm{W}}\right)W_\mu^3 + \cos\left(\theta_{\mathrm{W}}\right)B_\mu. \tag{2.35}$$

The equation 2.27 can be written in terms of the physical fields W_μ^+, W_μ^-, Z_μ and A_μ:

$$\frac{(g\nu)^2}{4}W_\mu^+W^{-\mu} + \frac{\left(g^2 + g'^2\right)\nu^2}{8}Z_\mu Z^\mu. \tag{2.36}$$

The first term gives mass of the W^+ and W^- vector bosons. The second term gives mass of the Z vector boson. There is no mass term for the photon. The spontaneous symmetry breaking of the Higgs field breaks the electroweak $\mathrm{SU(2)_L} \times \mathrm{U(1)}$ gauge symmetry to the U(1) gauge symmetry of the electromagnetism. The masses of the W^+, W^- and Z bosons are given by:

$$m_{\mathrm{W}^+} = m_{\mathrm{W}^-} = \frac{g\nu}{2}, \quad m_{\mathrm{Z}} = \frac{\nu\sqrt{g^2 + g'^2}}{2} = \frac{m_{\mathrm{W}}}{\cos\left(\theta_{\mathrm{W}}\right)}. \tag{2.37}$$

2.6 Higgs boson

For the Higgs doublet complex field in the SM, there are four real scalar degrees of freedom. Three degrees of freedom are "eaten" to form the longitudinal polarisations of

the $W_{\mu}^{\pm}$ and Z_{μ} fields. The remaining one real scalar degree of freedom forms the Higgs boson. The properties of the Higgs bosons can be shown in the unitary gauge, where three degrees of freedom are manifestly eaten. The Higgs field is given by:

$$H(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \nu + h(x) \end{pmatrix}, \tag{2.38}$$

where $h(x)$ is the real scalar field of the Higgs boson and ν the Higgs vacuum expectation value. The Higgs boson is not charged under electromagnetism as the field is real. The coupling of the Higgs boson to other fields can be calculated out by replacing ν with $\nu + h(x)$ in equation 2.36:

$$m_W^2 \left(\frac{2h}{\nu} + \frac{h^2}{\nu^2} \right) W_{\mu}^{+} W^{-\mu} + \frac{m_Z^2}{2} \left(\frac{2h}{\nu} + \frac{h^2}{\nu^2} \right) Z_{\mu} Z^{\mu}. \tag{2.39}$$

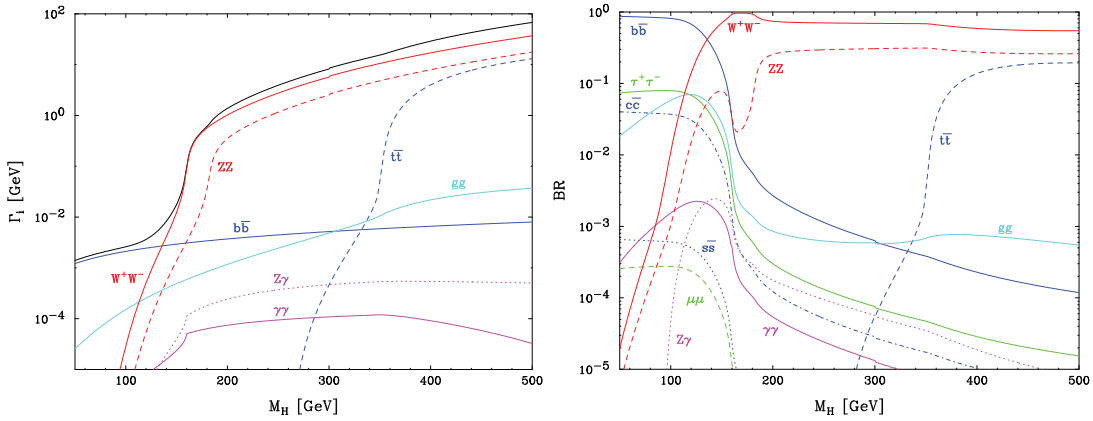
The Higgs boson self-interaction terms are obtained by replacing ν with $\nu + h(x)$ in the Higgs field potential in equation 2.23:

$$\frac{\mu^2}{2} (\nu + h)^2 - \frac{\lambda}{4} (\nu + h)^4 \supset -\lambda \nu^2 h^2 - \lambda \nu h^3 - \frac{\lambda}{4} h^4 \tag{2.40}$$

The quadratic term, $-\lambda \nu^2 h^2$, is the Higgs boson mass term, $m_{\mathrm{H}} = \sqrt{2\lambda}\nu$. The terms in h^3 and h^4 give trilinear and quadlinear Higgs self-interaction terms.

Once the Higgs boson mass is known, λ can be determined and the Higgs boson decay widths and branching fractions can be calculated. Figure 2.1 shows the Higgs boson partial decay widths and the branching ratios as a function of the Higgs boson mass for different Higgs decay modes.

The Higgs boson mass is measured to be $125.09 \pm 0.24\,\mathrm{GeV}$ [6]. Because the Higgs boson is lighter than a pair of heavier particles such as $\mathrm{W^+W^-}$ or ZZ, the process $\mathrm{H} \to \mathrm{W^+W^-}$ and $\mathrm{H} \to \mathrm{ZZ}$ are forbidden kinematically. However, the quantum field theory allows such processes to happen, if one of the decay products 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: a) the Higgs boson partial decay widths, and b) Higgs boson branching ratios, plotted as a function of the Higgs boson mass, m_{H}. In a), the black curve shows the total decay width. Both figures are taken from [19].

2.7 Yukawa couplings

The Yukawa sector of the electroweak Lagrangian provides mass terms for quarks and charged leptons after the spontaneous symmetry breaking of the Higgs field. The corresponding term in the Lagrangian is:

$$\mathcal{L}_{Yukawa} = -\lambda^u \overline{q_L} \Phi^c_{\mathrm{H}} u_R - \lambda^d \overline{q_L} \Phi_{\mathrm{H}} d_R - \lambda^e \overline{l_L} \Phi_{\mathrm{H}} e_R + h.c., \tag{2.41}$$

where q_L is the left-handed quark doublet field; u_R is the up-type right-handed quark singlet field; d_R is the down-type right-handed quark singlet field; l_L is the left-handed lepton doublet field; e_R is the right-handed charged lepton singlet field; λ is a constant; $\Phi^c_{\mathrm{H}} \equiv i\sigma^2 \mathrm{H}^*$ is an SU(2) doublet field with hypercharge $Y = -\frac{1}{2}$; $h.c.$ indicates the Hermitian conjugate terms; and the Lagrangian is summed over all possible quarks and leptons. When the Higgs vacuum expectation value is substituted into $\mathcal{L}_{Yukawa}$, the Yukawa interaction terms give the fermion mass terms:

$$m_u = \frac{\lambda^u \nu}{\sqrt{2}}, \; m_d = \frac{\lambda^d \nu}{\sqrt{2}}, \; m_e = \frac{\lambda^e \nu}{\sqrt{2}}. \tag{2.42}$$

2.8 Beyond the Standard Model Higgs Models

A number of BSM Higgs theories have been proposed. For example, the light Higgs could be a composite bound state of new strongly-interacting sector at the TeV scale. If the composite Higgs is pseudo Nambu-Goldstone boson from spontaneous global symmetry breaking, the Higgs can be naturally light [13]. In this model, the couplings of the Higgs would deviate from those in the SM for Higgs interactions at the TeV scale.

An important physics process for testing the Higgs theory is the double Higgs production via vector boson fusion at the TeV scale [20–22]. For the composite Higgs scenario, the scattering amplitude for this process increases with energy. It is difficult to measure the double Higgs production at the LHC due to the large SM background rate [21]. However, a multi-TeV electron−position linear collider, such as the Compact Linear Collider, would be able to measure the cross section for this process [23].

The study of double Higgs production via W^+W^- fusion can probe the Higgs trilinear self coupling, g_{HHH}, and quartic coupling, g_{WWHH}. The coupling g_{HHH} is associated with the terms in h^3 in Higgs potential in equation 2.40. The coupling g_{WWHH} is associated with the terms in h^2 in Higgs interaction terms with other fields in equation 2.39. Leading-order Feynman diagrams for double Higgs production via W^+W^- fusion are shown in figure 2.2. The diagram shown in figure 2.2a contains the triple Higgs vertex, which is sensitive to the Higgs trilinear self coupling g_{HHH}. The diagram in figure 2.2b is sensitive to the quartic coupling g_{WWHH}. Figures 2.2c and 2.2d show Feynman diagrams for irreducible background processes containing two HW^+W^- vertices.

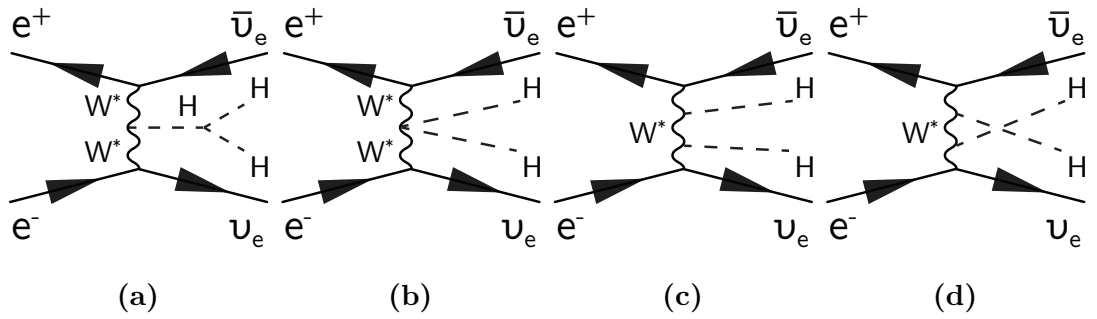


Figure 2.2: The main Feynman diagrams for the leading-order $e^+e^- \to HH\nu_e\overline{\nu}_e$ processes.

Following the assumption made in [21, 22], 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 [22]. In this description, after the electroweak symmetry breaking , the

Lagrangian is given by:

$$\mathcal{L} = \frac{1}{2}(\partial_\mu h)^2 - V(h) + \left(m_W^2 W_\mu^+ W^{-\mu} + \frac{m_Z^2}{2} Z_\mu Z^\mu\right)\left[1 + 2a\frac{h}{\nu} + b\frac{h^2}{\nu^2} + \dots\right], \tag{2.43}$$

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

$$V(h) = \frac{1}{2}m_h^2 h^2 + d_3\left(\frac{m_h^2}{2\nu}\right)h^3 + d_4\left(\frac{m_h^2}{8\nu^2}\right)h^4 + \dots, \tag{2.44}$$

and a, b, d_3 and d_4 are dimensionless parameters. Higher-order terms in h are omitted. The parameters a and b are proportional to the coupling strengths of the VVh and $VVhh$ vertices, where V represents a vector boson, and the parameters d_3 and d_4 are proportional to the trilinear and quadlinear h self-coupling strength, respectively. Comparing with the $\mathcal{L}_{Higgs}$ in the SM (see equation 2.39 and equation 2.40), it can be seen that $a = b = d_3 = d_4 = 1$ in the SM, and all higher order terms vanish. However, BSM Higgs model allow a, b, d_3, d_4 to take different values.

Consider a pair of the longitudinal polarised vector bosons (V_L) coupling to two h fields. The scattering amplitude for $V_L V_L \to hh$ can be written as:

$$A = a^2(A_{SM} + A_1\delta_b + A_2\delta_{d_3}), \tag{2.45}$$

where A_{SM} is the SM amplitude and:

$$\delta_b \equiv 1 - \frac{b}{a^2}, \tag{2.46}$$

$$\delta_{d_3} \equiv 1 - \frac{d_3}{a}. \tag{2.47}$$

The term A_1 grows like the square of energy at a large center-of-mass energy, $E \gg m_V$. The terms A_{SM} and A_2 have no energy dependence. Therefore, the parameter δ_b controls the magnitude of the increasing of the scattering amplitude as a function of energy. In an electron−positron collider, this scattering process can be studied via the double Higgs production $e^+e^- \to \nu\bar{\nu}hh$ channel, where the cross section can be written as

$$\sigma = a^4\sigma_{SM}\left(1 + A\delta_b + B\delta_{d_3} + C\delta_b\delta_{d_3} + D\delta_b^2 + E\delta_{d_3}^2\right), \tag{2.48}$$

where σ_{SM} is the SM cross section. Variables that increase with the increasing of the centre-of-mass energies are suitable for studying the cross section dependence on parameters δ_b and δ_{d_3}. 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 2.3 shows that the m_{hh} and H_T distributions are sensitive to the values of δ_b and δ_{d_3} [22]. The changes in the m_{hh} and H_T distributions can be related to the change in δ_b and δ_{d_3}. Therefore deviations of δ_b and δ_{d_3} from those SM values, 1, could be established using the m_{hh} and H_T distributions. It should be noted that figure 2.3 shows a generator-level study; detector effect will affect the distributions because of, for example, the loss of the reconstruction efficiency in the barrel/endcap gap region.

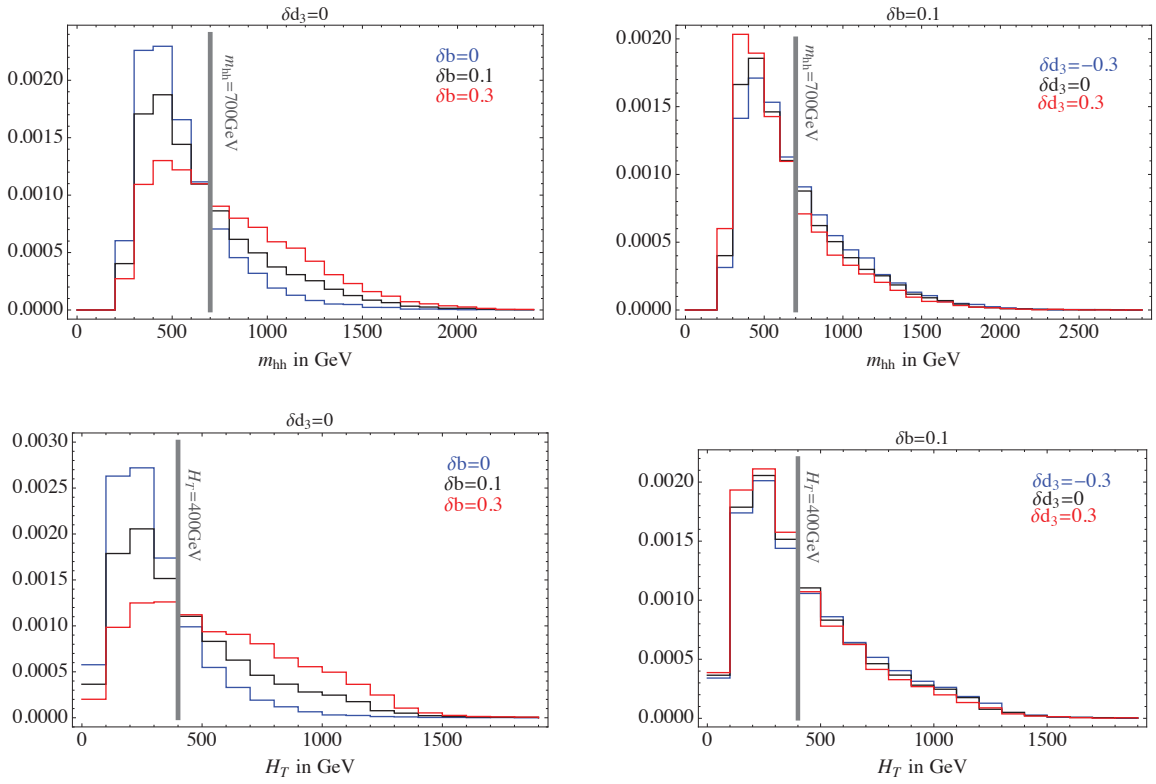


Figure 2.3: Normalised differential cross sections $d\sigma/dm_{hh}$ and $d\sigma/dH_T$ for $e^+e^- \to \nu\bar{\nu}hh$ for CLIC at $\sqrt{s} = 3\,\text{TeV}$ after applying generator-level identification cuts, for several values of δ_b and δ_{d_3}. The plot is taken from [22].

In the expression of the cross section of the double Higgs production via $e^+e^- \to \nu\bar{\nu}hh$ in equation 2.48, the parameter a, which is proportional to g_{VVH}, enters as an overall factor. Figure 2.4 shows the comparison of cross sections as a function of the centre-of-mass energy, for different the Higgs production modes. Up to a centre-of-mass energy of $\sqrt{s} = 3\,\text{TeV}$, the cross sections of the single Higgs production are two orders of magnitude larger than the cross sections of the double higgs production. The cross section of

$e^+e^- \to \nu\overline{\nu}h$ channel is given by:

$$\sigma = \sigma_{SM}\Big(1 + A\Delta a + B\Delta a^2\Big), \tag{2.49}$$

where $\Delta a \equiv 1 - a$ is the change in a, and A and B are two dimensionless coefficients. The measurement of the parameter a, using $e^+e^- \to \nu\overline{\nu}h$ channel, would be performed before the measurement of the δ_b and δ_{d_3} for the double Higgs production.

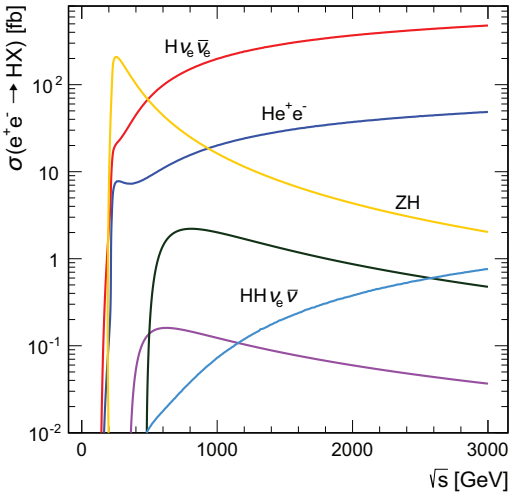


Figure 2.4: Cross sections as a function of centre-of-mass energy for Higgs production processes at an electron-positron collider for a Higgs mass of 126 GeV. The cross section values correspond to unpolarised beams and do not include the effect of beamstrahlung. The plot is taken from [24].

Therefore, for the purpose of measuring g_{VVHH} and g_{HHH} via double Higgs production, it is sufficient to treat the parameter a as a known constant. Hence only a two-dimensional fit of the parameters δ_b and δ_{d_3} would be performed to extract values of δ_b and δ_{d_3}.

2.9 Tau pair polarisation correlations as a signature of Higgs boson

The tau lepton has been studied extensively in the past at the Large Electron Positron Collider (LEP) [25]. The tau lepton is a fundamental particle, with a negative electric charge and a spin of $\frac{1}{2}$. It has the same fundamental interaction property as an electron, but a much larger mass. Unlike the stable electron, because the tau lepton is massive, it decays via the weak interaction with a mean decay lifetime of $(290.3 \pm 0.5) \times 10^{-15}$ s [26].

The tau lepton has many decay modes. The decay modes with branching ratio above 2% are listed in table 2.3.

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

Table 2.3: Decay modes, final state particles and branching ratios of the seven major τ^- decays, taken from [6].

A scalar Higgs boson with spin-0 can decay to $\tau_L^+\tau_L^-$ or $\tau_R^+\tau_R^-$, whereas a vector boson Z with spin-1 can decay to $\tau_L^+\tau_R^-$ or $\tau_R^+\tau_L^-$, where L, R denotes the tau lepton helicity, due to the helicity conservation. Therefore, by studying the correlation between the polarisations of the tau pair from a boson decay, one can determine statistically if the parent boson is a scalar or a vector.

The tau pair polarisation correlation can be studied using various tay decay modes. Here reference [27] is followed and the $\tau^- \to \pi^-\nu_\tau$ decay mode is used as the example. The Higgs and Z boson decay to a tau pair, where both tau leptons subsequently decay via $\tau^- \to \pi^-\nu_\tau$, can be represented as:

$$X \to \tau_\alpha^+\tau_\beta^- \to \pi^+\pi^- + \nu s, \tag{2.50}$$

where X is either H or Z, and α, β are the tau lepton 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:

$$z = \frac{E_{\pi^-}}{E_{\tau^-}}, \tag{2.51}$$

$$\overline{z} = \frac{E_{\pi^+}}{E_{\tau^+}}. \tag{2.52}$$

For a single tau decay, the differential cross section distribution can be written as:

$$\frac{1}{\Gamma_\tau}\frac{d\Gamma}{dz} = Br\Big(\tau^- \to \pi^-\nu_\tau\Big) f\Big(\tau_\alpha^- \to \pi^-; z\Big), \tag{2.53}$$

where $Br(\tau^- \to \pi^-\nu_\tau)$ is the branching fraction of $\tau^- \to \pi^-\nu_\tau$ decay mode. The form f can be obtained by working outing the matrix element from the Feynman diagram and integrating the square of the matrix element over the phase space [28]:

$$f\Big(\tau_\alpha^- \to \pi^-; z\Big) = 1 + P_\alpha(2z-1), \tag{2.54}$$

where $P_L = -1$ and $P_R = +1$. Hence for the tau pair decay, the differential cross section distribution is of the form:

$$\frac{d^2N(X \to \tau^+\tau^- \to \pi^+\pi^- + \nu's)}{dz\,d\overline{z}} = \Big(Br\Big(\tau^- \to \pi^-\nu_\tau\Big)\Big)^2 \sum_{\alpha,\,\beta} C^X_{\alpha\beta} f\Big(\tau_\alpha^- \to \pi^-; z\Big) f\Big(\tau_\beta^+ \to \pi^+; \overline{z}\Big), \tag{2.55}$$

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

$$C^{\mathrm{H}}_{LL} = C^{\mathrm{H}}_{RR} = \frac{1}{2}. \tag{2.56}$$

In contrast, the non-zero correlation coefficients for the $\mathrm{Z} \to \tau^+\tau^-$ are:

$$C^{\mathrm{Z}}_{LR} = \frac{1}{2}\Big(1 - P_\tau\Big),\ C^{\mathrm{Z}}_{RL} = \frac{1}{2}\Big(1 + P_\tau\Big), \tag{2.57}$$

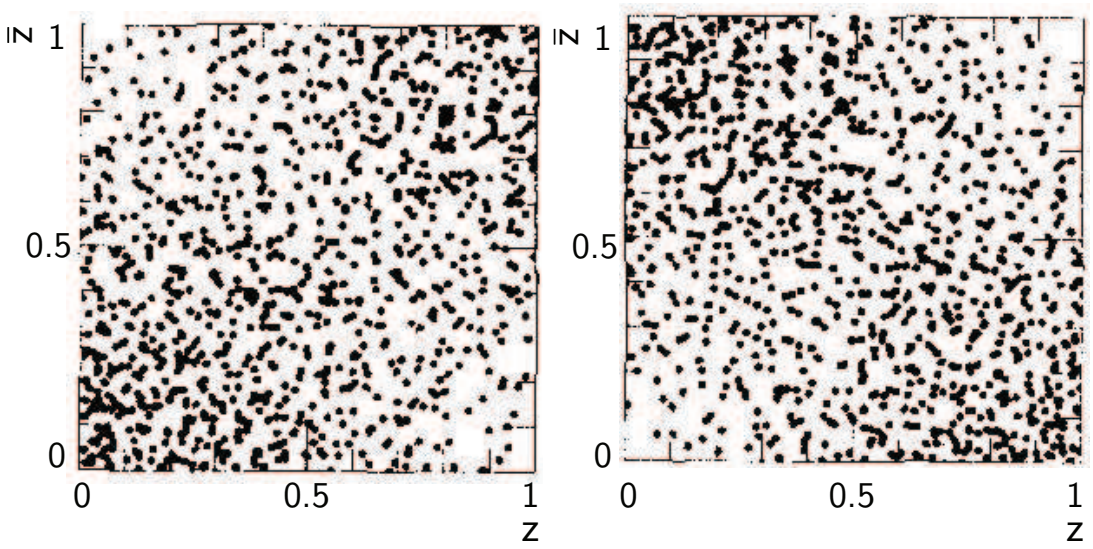
where P_τ is the mean non-zero tau polarisation of Z decays. The tau polarisation is not zero because the process $\mathrm{Z} \to \tau^+\tau^-$ is not parity-conserving. In the SM:

$$P_\tau = \frac{-2va}{v^2 + a^2}, \tag{2.58}$$

where the parameter $v = -\frac{1}{2} + \sin^2\theta_{\mathrm{W}}$ and $a = -\frac{1}{2}$ are the respective vector and axial-vector $\mathrm{Z}\tau^+\tau^-$ couplings.

Figure 2.5 shows the resulting two-dimensional distributions of $\overline{z} = \frac{E_{\pi^+}}{E_{\tau^+}}$ versus $z = \frac{E_{\pi^-}}{E_{\tau^-}}$ for $\mathrm{Z} \to \tau^+\tau^-$ and $\mathrm{H} \to \tau^+\tau^-$ channels, where both tau leptons decay via $\tau^- \to \pi^-\nu_\tau$. The difference of the tau pair polarisation correlation between Z and H is clear. The energy distribution of the charged pion from $\mathrm{Z} \to \tau^+\tau^-$ has the form of

$\overline{z} \sim z$, whilst the distribution from $H \to \tau^+\tau^-$ has the form of $\overline{z} \sim (1 - z)$. Therefore, in $Z \to \tau^+\tau^-$ process, a high-energy $\pi^\pm$ is likely to be associated with a high-energy $\pi^\mp$. In $H \to \tau^+\tau^-$ process, the opposite is favoured. If the tau pair decay from Higgs boson is observed, the decay can be recognised in the $\tau^- \to \pi^-\nu_\tau$ mode as a high-energy $\pi^\pm$ with a low-energy $\pi^\mp$. Hence, the tau decay product energy distribution can be a clean signature for $H \to \tau^+\tau^-$.



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

Figure 2.5: Two-dimensional distribution of $\overline{z} = E_{\pi^+}/E_{\tau^+}$ plotted against $z = E_{\pi^-}/E_{\tau^-}$ for a) $Z \to \tau^+\tau^-$, and b) $H \to \tau^+\tau^-$ processes, where both tau leptons decay via $\tau^- \to \pi^-\nu_\tau$, adapted from reference [28].

Chapter 3

Detectors for Future Electron-Positron Linear Colliders

'The person attempting to travel two roads at once will get nowhere.'
— Xun Kuang, 313 BC – 238 BC

Two leading candidates for next-generation electron-positron linear particle colliders are the International Linear Collider (ILC) [1], and the Compact Linear Collider (CLIC) [2]. This chapter provides an overview of the two colliders, followed by the physics programme at these colliders, the detectors requirements, and the description of detectors for the ILC and CLIC.

3.1 International Linear Collider

The ILC is a high-luminosity future electron-positron linear particle collider. The machine will be built in two 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 1 TeV. The layout of the collider complex is shown in figure 3.1. Two detector concepts have been developed for the ILC: the International Large Detector (ILD) [29] and the Silicon Detector (SiD) [30]. Both detectors are shown in figure 3.2.

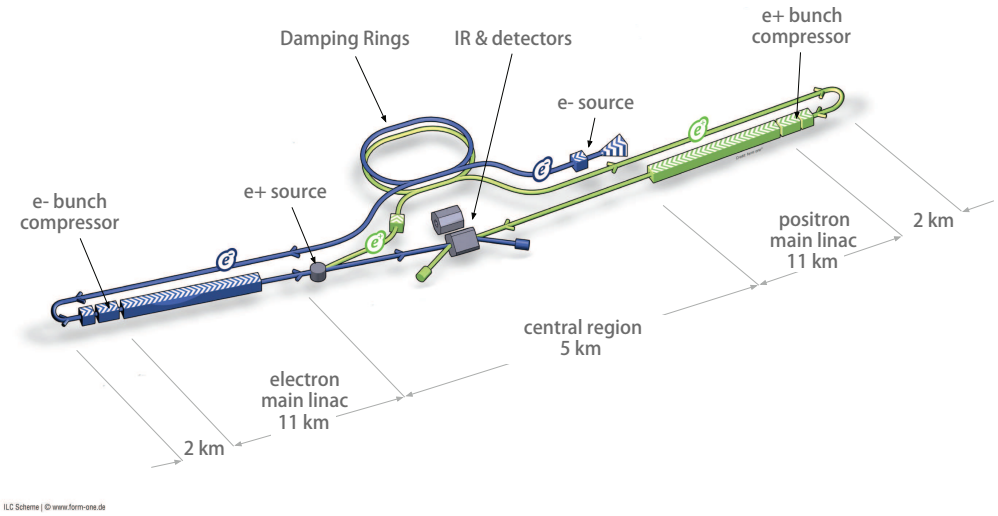
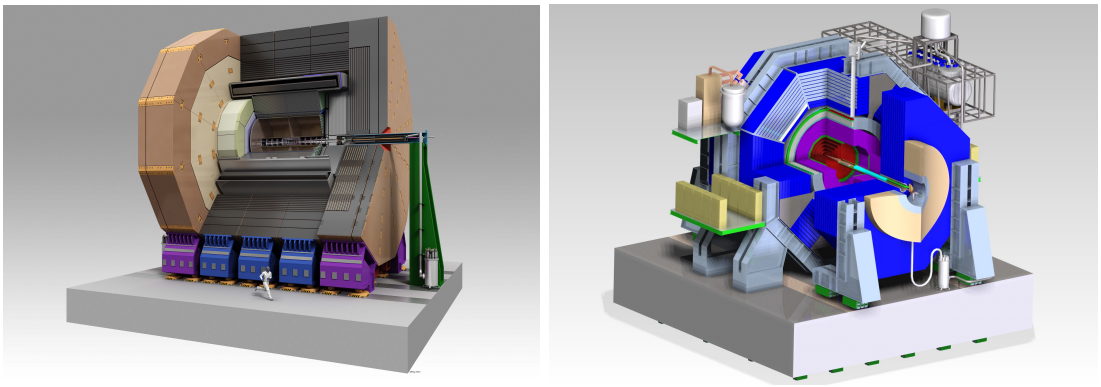


Figure 3.1: Schematic layout of the International Linear Collider, indicating all the major subsystems (not to scale), taken from [31].



(a) **(b)**

Figure 3.2: a) the International Large Detector, and b) the Silicon Detector. Both detector concepts are developed for the International Linear Collider. Both figures are taken from [31].

3.2 Compact Linear Collider

CLIC is a potential next-generation electron-positron linear particle collider at CERN [2]. CLIC is designed to be built in three stages: a first stage of a centre-of-mass energy of 380 GeV; a second stage of a centre-of-mass energy of 1.4 TeV; and the final stage of a centre-of-mass energy of 3 TeV. The layout of the CLIC complex at the final stage is shown in figure 3.3.

The ILC and CLIC share some common features. Both colliders will be linear colliders as oppose to circular colliders, like the Large Hadron Collider (LHC). Detectors for both colliders will use the high granularity particle flow calorimetry [2,31]. One major difference between the two colliders is the operating energy. Due to a higher centre-of-mass energy at CLIC, there are significant beam related backgrounds. The e^+e^- incoherent pair background has a major influence on the design of the inner region and the forward region of the detectors [2]. The pile-up of 3.2 $\gamma\gamma \to$ hadrons background events per bunch, also integrating over 60 bunch crossings, need to be mitigated for physics analyses [2]. Another difference between the ILC and CLIC is that the timing separation between bunches is much shorter at CLIC. The CLIC beam contains 312 bunch trains with a train repetition rate of 50 Hz, separated by 0.5 ns between each bunch train. This short timing separation suggests that the detector will integrate over a number of bunch crossings.

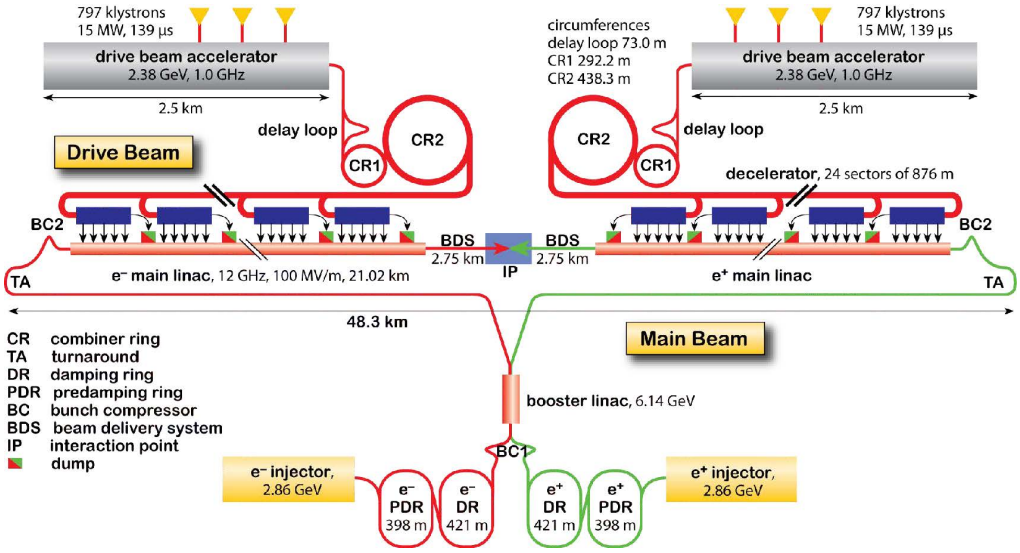


Figure 3.3: The layout of the Compact Linear Collider at the final stage of a centre-of-mass of energy of 3 TeV, taken from [32].

3.3 Physics at future linear colliders

An e^+e^- linear collider has advantages over the current hadron collider, the LHC. Those advantages include:

i) Events in the e^+e^- collider will be cleaner than those in the hadron collider. In the LHC, many proton−proton collisions per bunch crossing are expected [33], generating hundreds of particles from parton collisions. In the e^+e^- collider, the main source of background comes from photon−photon collisions [1, 2]. Depending on the operating energy and scheme, there will be only a few of these photon−photon collisions per bunch crossing. Particles produced from these collisions are mainly in the forward direction, which can be identified relatively easily.

ii) Electroweak interactions in the e^+e^- collider will be democratic as the photon couples to all particles in and beyond the Standard Model equally [6–9]. The production of pairs of all particles will be at a similar rate. In the LHC, the non-perturbative strong interaction is the main channel for the particle production. Heavy particles has a lower production rate than light ones, as the parton distributions fall sharply for a composite object like a proton [34].

iii) In the LHC, calculation of the cross section depends on quantum chromodynamics and the proton structure function, which have large systematic errors. In an e^+e^- collider, the initial particles, e^+ and e^-, are point-like particles, interacting through electroweak forces only. Consequently, theoretical uncertainties are smaller.

iv) The physics at an e^+e^- collider can be studied in detail. Without complicated underlying events, complete events can be reconstructed. Polarised beams of electrons and positrons with known initial and final polarisation states also could be used to enhance the production of certain interactions, for example, electron−positron annihilation with opposite helicities.

The physics programmes for the ILC and CLIC, which is a driving force behind the detector design, share some common features. At a centre-of-mass energy of 250 GeV, the collider could operate as a Higgs factory, allowing measurements of precision Higgs couplings via channels like $e^+e^- \to ZH$. At a centre-of-mass energy of 350 GeV, the collider can continue to measure Higgs couplings, as well as to measure top quark mass and couplings via channels such as $e^+e^- \to t\bar{t}$. At a centre-of-mass energy of 1 TeV and beyond, the collider would be able to produce rare Higgs decays, allowing measurements of Higgs

self-couplings and probing composite Higgs sector, and to search for supersymmetric particles [5].

3.4 Detector requirements

Many physics processes at a future linear collider can be characterised by multi-jet final states, often with charged leptons or missing momentum associated with neutrinos. The reconstruction of the invariant masses of two or more jets is crucial for event reconstruction and event selection. At the Large Electron−Positron Collider (LEP), kinematic fitting [35] allowed precise invariant mass reconstruction. At a future linear collider, reconstruction invariant mass of multiple jets for final states with missing momentum will rely heavily on the intrinsic jet energy resolution of the detector.

One goal of jet energy resolution at a future linear collider is to be able to separate W and Z bosons, by reconstructing their invariant massese via quark-jets using the hadronic decay channel. The idealised reconstructed W and Z boson masses distributions for different jet mass resolutions are shown in figure 3.4. As the invariant mass resolution is comparable to the gauge boson widths, i.e. $\sigma_m/m \approx \Gamma_\mathrm{W}/m_\mathrm{W} \approx \Gamma_\mathrm{Z}/m_\mathrm{Z}$, a separation of 2.5σ in the masses distributions implies a jet energy resolution of 3.5% [36] for a range of jet energies from 50 GeV to 1 TeV.

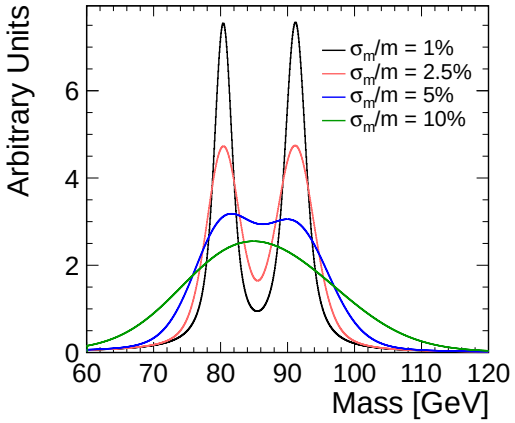


Figure 3.4: Ideal W/Z boson mass separation for different jet mass resolutions obtained using a Gaussian smearing of Breit−Wigner distribution, taken from [2].

3.5 Particle Flow Calorimetry

A jet energy resolution of 3.5% is unlikely to be achieved with a traditional calorimeter design. Traditionally, jet energies are measured as a sum of energies deposited in the electromagnetic (ECAL) and hadronic calorimeter (HCAL), giving a jet energy resolution of the form

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

The stochastic term α is typically greater than 60% [29, 31], and the constant term β is a few percent [29, 31]. To achieve a jet energy resolution of 3.5% or better, the stochastic term should be less than 30% with a small constant term, which is unlikely to be achieved by a traditional calorimeter.

In a typical jet, about 62% of the jet energy is from charged particles, 27% from photons, 10% from long-lived neutral hadrons, and 1.5% from neutrinos [37, 38]. In a traditional approach to calorimetry, the jet energy resolution is limited by the relatively poor energy resolution of the hadronic calorimeters.

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 energies of individual particles, where the energies of the charged particles are measured in the tracking detectors, and the energies of neutral particles are measured in calorimeters. Hence the hadronic calorimeter only measures about 10% of the jet energy.

As shown in table 3.1, assuming 30% of the jet energy (photon energy) is measured with $\sigma_E/E = 15\%/\sqrt{E(\mathrm{GeV})}$, and 10% of the jet energy (hadron energy) is measured with $\sigma_E/E = 55\%/\sqrt{E(\mathrm{GeV})}$ [31], a jet energy resolution of $\sigma_E/E = 19\%/\sqrt{E(\mathrm{GeV})}$ can be obtained. This satisfies the jet energy resolution requirement for separating W and Z bosons 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, results from imperfect reconstruction rather than the intrinsic detector performance limit the particle flow performance [36].

In the particle flow approach to calorimetry, the sum of calorimeter energies is replaced by a complex pattern-recognition problem, which is solved by the Particle Flow reconstruction Algorithm (PFA). Detailed simulations of the ILC and the CLIC detector

Component	Detector	Energy fraction	Energy resolution	Jet energy resolution
Charged particles (C)	Tracker	$\sim 0.6 E_j$	$10^{-4} E_C^2$	$< 3.6 \times 10^{-5} E_j^2$
Photons (γ)	ECAL	$\sim 0.3 E_j$	$0.15\sqrt{E_\gamma}$	$0.08\sqrt{E_j}$
Neutral hadrons(N)	HCAL	$\sim 0.1 E_j$	$0.55\sqrt{E_N}$	$0.17\sqrt{E_j}$

Table 3.1: Contributions from different particle components to the jet energy resolutions (all energies in GeV). The table lists the approximate fractions of charged particles, photons, and neutral hadrons in a jet of energy E_j, and the assumed single particle energy resolutions. The table is adapted from [36].

concepts using the PandoraPFA [36, 39] particle flow reconstruction algorithms have demonstrated that a jet energy resolution of approximately 3% can be achieved for jet energies in the range of 100 GeV to 1 TeV.

Particle flow calorimetry works by fully reconstructing particles and associating calorimeter hits to tracks in tracking detectors. This places stringent requirements on the calorimeter designs. The ECAL and the HCAL need to be highly granular for an excellent spatial resolution to correctly associate calorimeter hits to the inner detector tracks. The tracking system needs to have an excellent momentum resolution for the momentum measurements of the charged particles.

3.6 International Large Detector

Two detector concepts have been developed for the ILC. The motivation for having two detectors is to have multiple independent measurements within one collider for cross-checking, complementarity, and competition between collaborations. The two detectors are both designed to be general purpose detectors. The Silicon Detector, SiD [30], is a compact detector with silicon tracking modules and a magnetic field of 5 T. The other detector, the International Large Detector, ILD [29], is a larger detector with a time projection chamber as the main tracking detector.

The ILD detector concept has been optimised for particle flow techniques. Figure 3.5 shows the longitudinal cross section of top quadrant of the ILD detector concept. 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); forward calorimeters (FCAL); a superconducting solenoid; and muon chambers embedded

within the iron return yokes. The key parameters of the ILD are listed in table 3.2. The section below describes the sub-detectors of the ILD detector concept referred to as the ILD_o1_v05 option in the MOKKA detector simulation [40], used for the ILD technical design report [31].

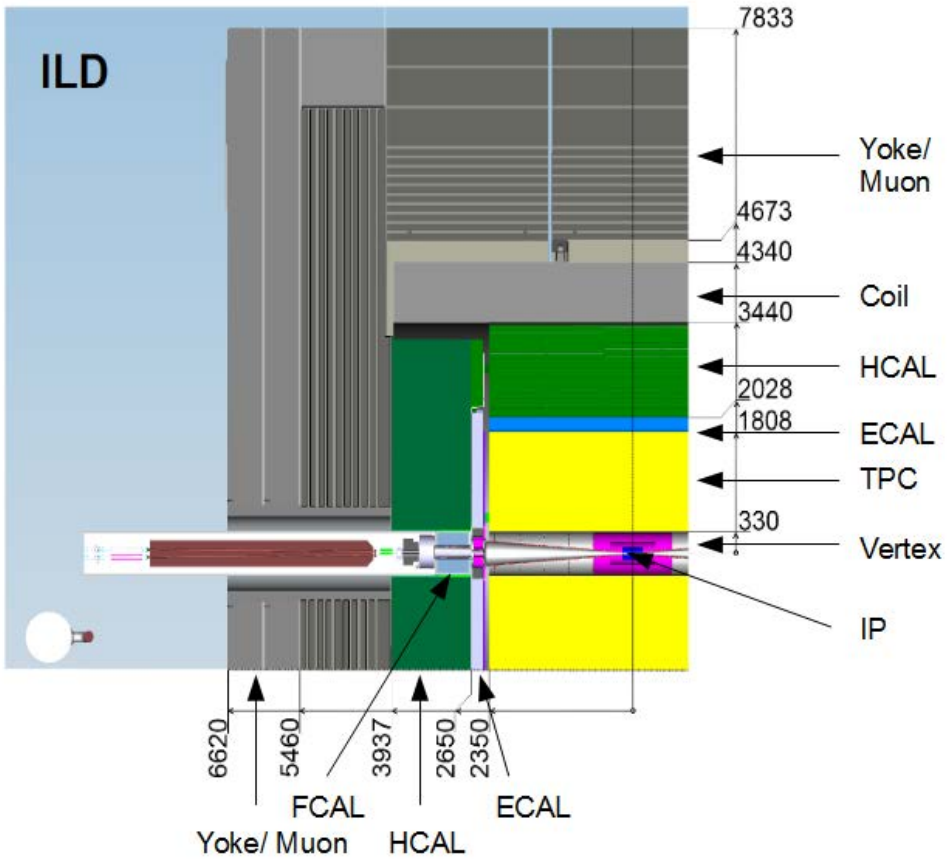


Figure 3.5: The longitudinal cross section of top quadrant of the ILD, taken from [31]. From interaction point (IP) outwards, there is: a tracking system comprising a large time projection chamber (TPC) augmented with silicon tungsten layers; highly granular electromagnetic calorimeters (ECAL) and hadronic calorimeters (HCAL); forward calorimeters (FCAL); a superconducting solenoid; and muon chambers embedded within the iron return yokes. The dimensions are in units of mm.

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. Figure 3.6 shows the structure of the vertex detector.

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

Table 3.2: 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. The table is adapted from [2].

The structure is of three, nearly cylindrical, concentric layers of double-sided ladders. Each ladder contains pixel sensors on both sides at 2 mm separation between two layers. This results in six measured positions for each charged particle traversing the detector. The first double layer is half the 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 3.3. The radii covered by the detector range from 16 mm to 60 mm.

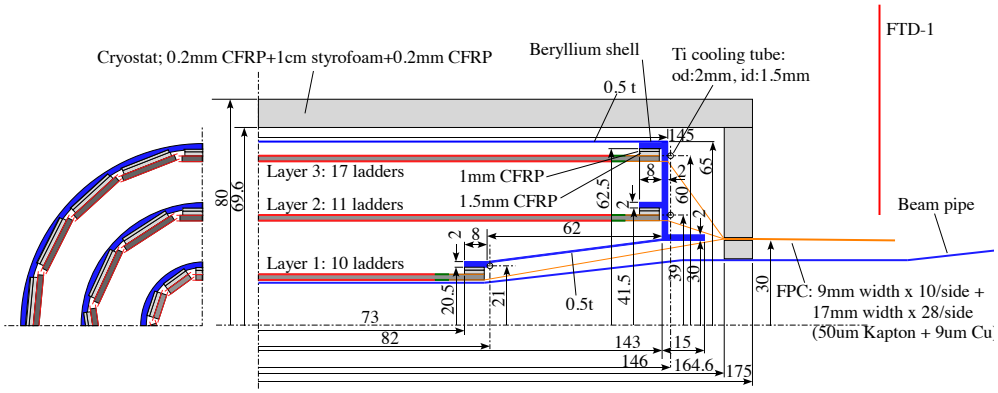


Figure 3.6: Structure of ILD vertex detector, taken from [31].

	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

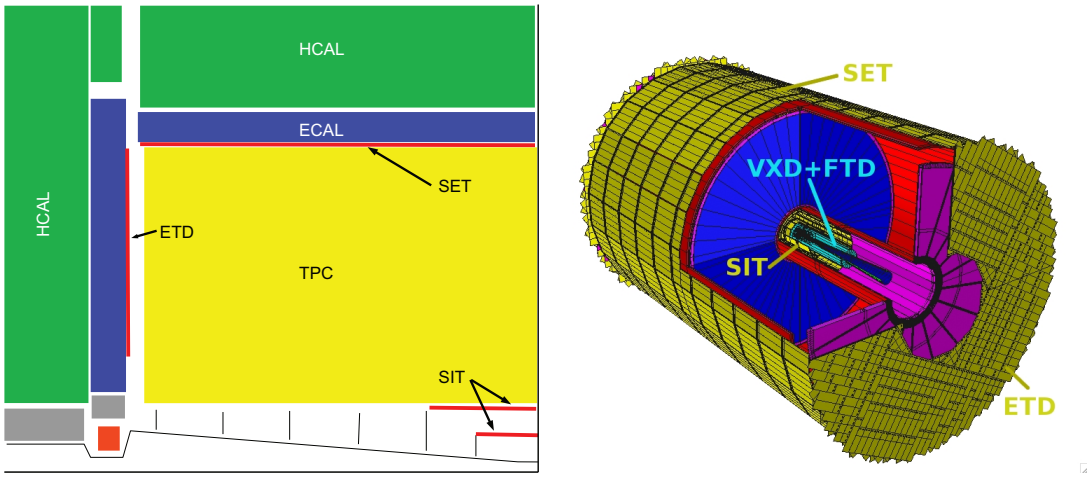
Table 3.3: Key parameters for vertex detector in the ILC. The spatial resolution (σ) and readout times are for the CMOS option. The table is adapted from [31].

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 silicon end cap tracking component (ETD) behind the endplate of the TPC, and a silicon forward tracker (FTD) in the forward region. A top quadrant view of the ILD silicon envelope system with the TPC is shown in figure 3.7. The SIT, SET, and ETD are made up of two single-sided strip layers tilted by a small angle. The FTD is a system of two silicon-pixel disks and five silicon-strip disks. The main parameters of the silicon system and the TPC can be found in table 3.4.

A TPC tracking detector has several advantages: a) tracks can be measured with a large number of three-dimensional (r, ϕ, z) spatial points; b) the continuous tracking allows precise reconstruction of tracks; and c) the TPC uses a minimum amount of material, which minimises the photon to electron pair conversion.

The silicon intermediate tracker (SIT) and the silicon envelope tracker (SET) provide spatial point measurements before and after the TPC in the barrel region. This helps to improve the overall momentum resolution by providing points to link the vertex detector with the TPC, and to extrapolate tracks from the TPC to the calorimeters. The FTD improves the low angle coverage of the tracking system, where the low angle is not covered by the TPC.



(a) (b)

Figure 3.7: a) A top quadrant view of the ILD silicon envelope system, SIT, SET, FTD, and ETD, with TPC, ECAL, and HCAL, and b) a 3D detailed GEANT4 simulation description of the silicon system as sketched in the quadrant view in a). Both plots are adapted from figures in [31].

	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	$\pm$ 2350 mm	up to 0.98

Table 3.4: Main parameters of the central silicon tracking systems (SIT, SET, and ETD) and the TPC. The table is adapted from [31].

3.6.3 Electromagnetic Calorimeter

The silicon−tungsten sampling electromagnetic calorimeters in the ILD consist of an octagonal barrel and two end cap systems. The fine granularly ECAL is located inside the HCAL. Both ECAL and HCAL are inside the superconducting solenoid. Figure 3.8a shows the position of the electromagnetic calorimeter in the ILD detector and the trapezoidal form of the modules.

The particle flow paradigm has a large impact on the ECAL design. In addition to measuring and separating photons, the ECAL needs to allow the reconstruction of detailed shower profiles to separate electromagnetic showers from hadronic showers, as approximately 50% of hadronic showers start in the ECAL.

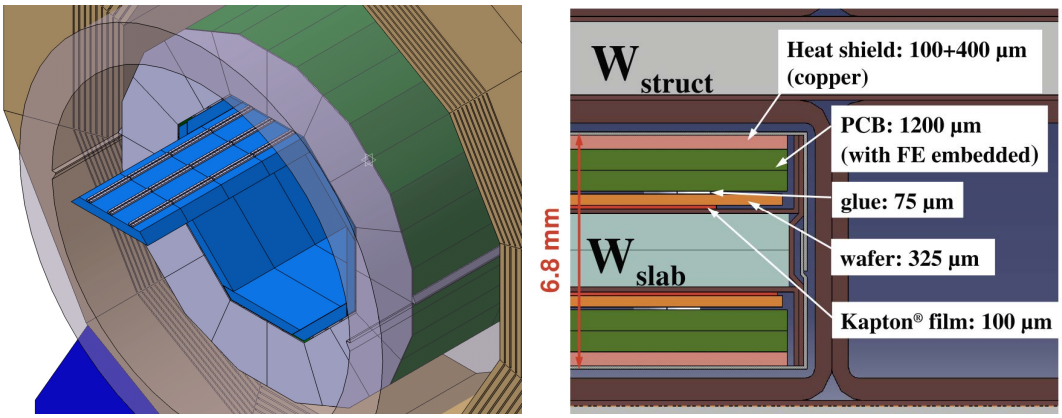
From test beam data and simulation studies [41–43], a sampling calorimeter with a longitudinal segmentation below one radiation length and the transverse segmentation respectively below one Molière radius is required. A compact design is realised with tungsten as the absorber material and silicon pad diodes as the active material. A cross section of an ECAL layer is shown in figure 3.8b. Tungsten is a dense material with a large ratio of interaction length to radiation length. A small radiation length will promote the start of the electromagnetic shower earlier in the calorimeter, whilst a large interaction length will reduce the fraction of hadronic showers starting in the ECAL.

The ECAL, which is about 20 cm thick, has 30 longitudinal layers providing about 24 radiation lengths. The first 20 layers use 2.1 mm thick absorber plates and the last 10 layers have 4.2 mm thick absorber plates.

The choice of thin silicon layers offers a great spatial resolution. The chosen size of $5.1 \times 5.1\,\mathrm{mm}^2$ silicon pads provides enough segmentation to meet the requirements of the particle flow paradigm.

3.6.4 Hadronic Calorimeter

The principal role of the HCAL is to separate neutral hadron showers from other particles, and to measure (neutral) hadron energies. The ILD HCAL is a sampling calorimeter with steel absorber and scintillator tiles as the active medium. The layout of the HCAL is 48 longitudinal layers with $3 \times 3\,\mathrm{cm}^2$ scintillator tiles, using an analogue read out system. The layout of a technological prototype, the "EUDET prototype" [44] is shown in figure 3.9.



(a) (b)

Figure 3.8: a) The electromagnetic calorimeter (in blue) within the ILD detector. b) A cross section through the electromagnetic calorimeter layers. Both plots are taken from [31].

For the absorber material, stainless steel is chosen for mechanical and calorimetric reasons. Steel allows a self-supporting structure without auxiliary supports. At the same time, iron has a moderate ratio of hadronic interaction length ($\lambda_I = 17\,\mathrm{cm}$) to electromagnetic radiation length ($X_0 = 1.8\,\mathrm{cm}$), which allows a fine longitudinal sampling in X_0 with a reasonable number of layers in a given total hadronic absorption length. The longitudinal system, including the ECAL, provides about 6 interaction lengths, which is sufficient to contain the hadronic showers.

The scintillator tiles provide both energy and position measurement. The transverse segmentation, chosen by optimisation studies [36] is about $3 \times 3\,\mathrm{cm}^2$. This level of segmentation is sufficient to meet the requirement of the particle flow [31].

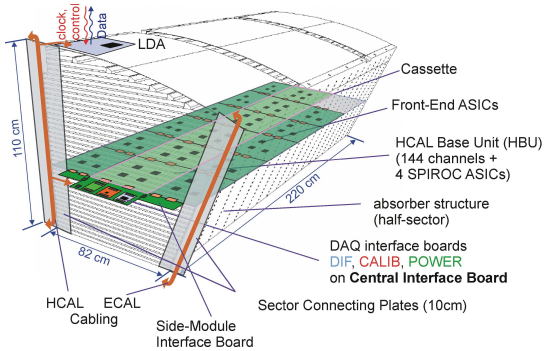


Figure 3.9: The schematic view of a CALICE analogue HCAL technological prototype module, taken from [31].

3.6.5 Solenoid, Yoke and Muon system

A large superconducting solenoid, outside the calorimeters, produces a nominal 3.5 T magnetic field. Figure 3.10 shows the cross section of the ILD magnet.

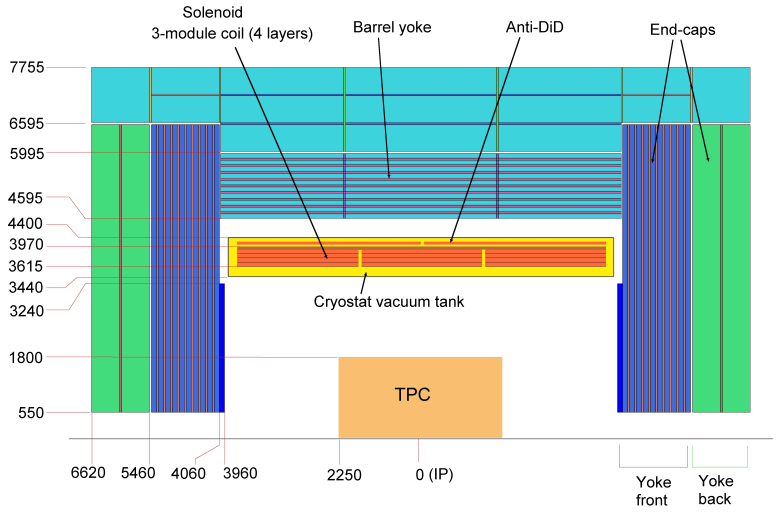


Figure 3.10: The ILD magnet cross section. Dimensions are in mm. Figure is taken from [31].

The iron yoke returns the magnetic flux. The yoke is designed to ensure that the maximum magnetic field at 15 m radial distance from the detector is 50 Gauss to ensure safety [45].

The iron yoke is also instrumented with scintillator strips as active layers to act as a muon detector. A highly efficient muon detector is provided by the $3 \times 3\,\mathrm{cm}^2$ scintillator tiles. The layout of the muon detector is shown in figure 3.11.

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 [46].

3.6.6 Very Forward Calorimeters

The detectors in the forward region provide luminosity measurements and forward coverage of calorimeters. A system of precision and radiation resistent calorimeters is required. Figure 3.12 shows the forward calorimeters of the ILD.

The luminosity calorimeter (LumiCAL) counts Bhabha scattering to measure the luminosity to precision of 10^{-3} at a centre-of-mass energy of 500 GeV [47]. The beam

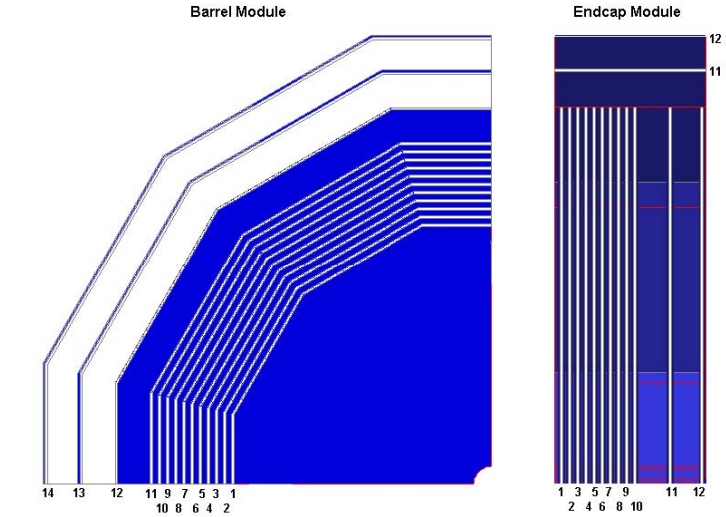


Figure 3.11: Sensitive layers of the ILD muon system, taken from [31].

calorimeter (BeamCAL), which is hit by many beamstrahlung pairs after each bunch crossing, extends the forward coverage. The BeamCAL also provides a measurement of the bunch-by-bunch luminosity. An additional hadron calorimeter in the forward region, LHCAL, extends the angular coverage of the HCAL to that of the LumiCAL.

The calorimeters in the forward region also provides enough information for high-energy electron tagging [48], which aids event reconstruction at a high centre-of-mass energy. Table 3.5 lists the key parameters of the LumiCAL and the BeamCAL in the ILD.

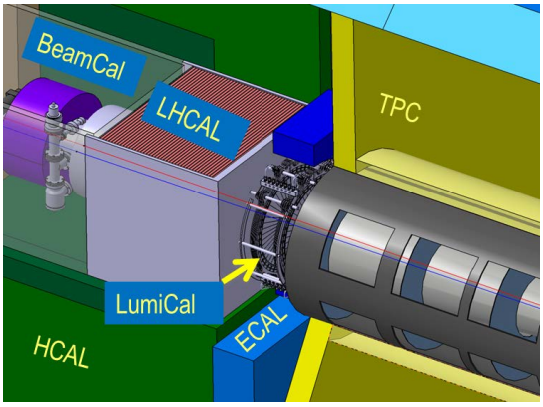


Figure 3.12: The calorimeter in the forward region of the ILD, taken from [31]. The LumiCAL, the BeamCAL, and the LHCAL are the luminosity calorimeter, the beam calorimeter, and the forward hadronic calorimeter, respectively.

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

Table 3.5: Comparison of the key parameters for the LumiCAL and the BeamCAL at the ILD and the CLIC_ILD detector concepts. The table is adapted from [2].

3.7 Detector optimisation

Detector optimisation studies were performed to choose the optimal parameters for the ILD sub-detectors [31]. Here the optimisation studies for the ECAL and the HCAL are presented.

3.7.1 Electromagnetic Calorimeter optimisation

The metric for optimisation is the jet energy resolution, which defined as the root-mean-square divided by the mean for the smallest width of distribution that contains 90% of entries, using $e^+e^- \to Z'Z'$ samples, where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$, at barrel region. The angular cut is to avoid the barrel/endcap overlap region. The light quark decay of the Z' is used to avoid the complication of missing momentum from semi-leptonic decay of heavy quarks. Using 90% of the entries is robust and focuses on the Gaussian part of the distribution.

The optimisation of the ILD ECAL design was performed as a function of the number of longitudinal layers, whilst keeping other geometry constant. Figure 3.13 shows the jet energy resolution for a single jet as a function of the number of longitudinal layer, for four different jet energies. 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. Therefore, 30 longitudinal layers is chosen for the ECAL.

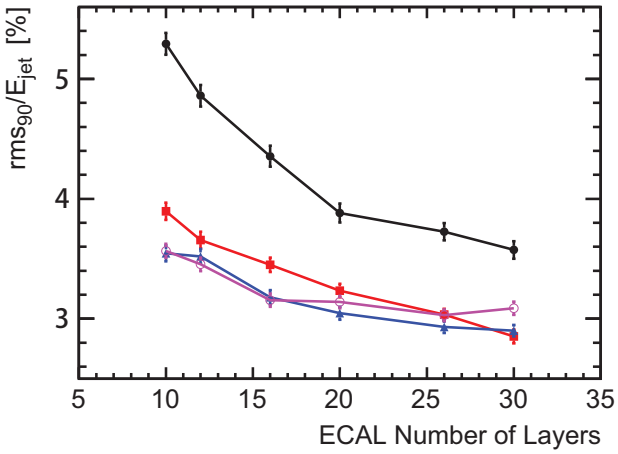


Figure 3.13: The single jet energy resolution as a function of the number of longitudinal ECAL layers, with different total jet energy using $e^+e^- \to Z'Z'$ samples, where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$, at barrel region, adapted from [31].

The ILD ECAL design was optimised as a function of the transverse cell sizes, whilst keeping other geometry constant. Figure 3.14 shows the jet energy resolution for a single jet, plotted as a function of transverse scintillator cell sizes for four different jet energies. The $10 \times 10\,\mathrm{mm}^2$ cell size is needed to meet the jet energy requirement of σ_E/E <3.8% for the jet energies relevant at $\sqrt{s} = 1\,\mathrm{TeV}$, with $5 \times 5\,\mathrm{mm}^2$ cell size being preferable.

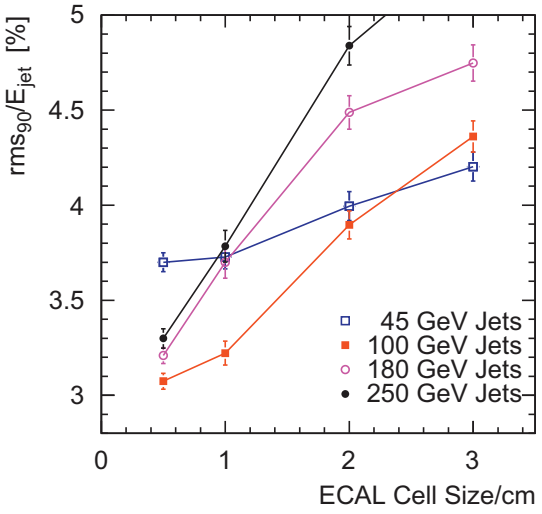


Figure 3.14: The single jet energy resolution as a function of the ECAL transverse cell sizes, with different total jet energy using $e^+e^- \to Z'Z'$ samples, where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$, at barrel region, adapted from [36].

3.7.2 Hadronic calorimeter optimisation

For the ILD HCAL design, the transverse cell sizes has been optimised using the jet energy resolution as the metric. The jet energy resolutions as a function of HCAL scintillator square cell sizes for foure different jet energies are shown in figure 3.15. 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 $3 \times 3\,\mathrm{cm}^2$ scintillator cell size is chosen for the HCAL design.

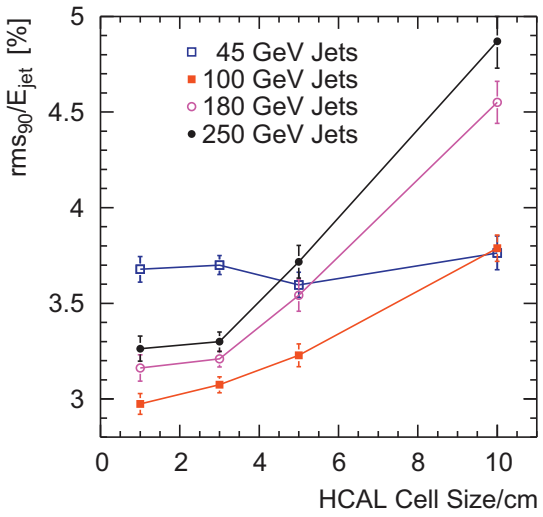


Figure 3.15: The single jet energy resolution as a function of the hadronic calorimeter scintillator cell sizes, with different total jet energy using $e^+e^- \to Z'Z'$ samples, where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$, at barrel region, adapted from [31].

3.8 CLIC_ILD detector concept

There are two detector concepts studied in the CLIC conceptual design report [2], the CLIC_ILD and the CLIC_SiD concepts. The CLIC_ILD detector concept is based on the ILD detector concept. Figure 3.16 shows the longitudinal cross section of the CLIC_ILD detector. A comparison of key parameters for the ILD and the CLIC_ILD detector concepts is shown in table 3.2.

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 and a smaller beam jet. The detector is also required to provided time stamping at the nanosecond level.

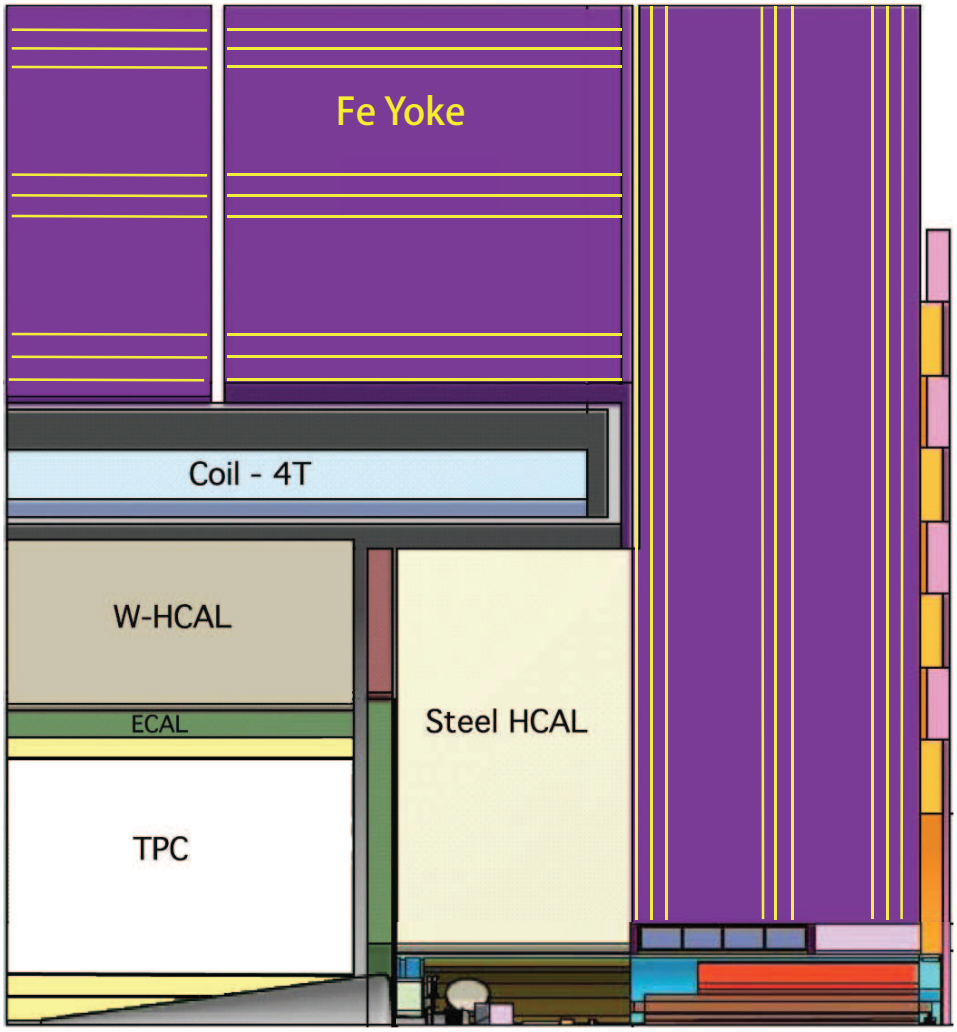


Figure 3.16: The longitudinal cross section of top quadrant the CLIC_ILD detector concept, taken from and [2]. From interaction point (IP) outwards, there is: a tracking system comprising a large time projection chamber (TPC) augmented with silicon tungsten layers; highly granular electromagnetic calorimeters (ECAL) and hadronic calorimeters (HCAL); forward calorimeters (FCAL); a superconducting solenoid; and muon chambers embedded within the iron return yokes.

For the CLIC_ILD tracking detector, the same silicon−TPC hybrid structure is used. At 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. 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 $12^\circ \lesssim \theta \lesssim 168^\circ$

For the CLIC_ILD design, the same ECAL as the ILD is assumed, as the requirements of a CLIC detector are satisfied by the ECAL design at the ILD.

For the CLIC_ILD HCAL, extra layers are added to contain the hadronic shower for the higher centre-of-mass energies of CLIC. The increased thickness is justified by the simulation studies [2], where the jet energy resolution degrades quickly for a thinner HCAL. To sustain the same inner bore radius as the ILC detector solenoid, a more dense material, tungsten, is chosen as the absorber material in the HCAL barrel.

The magnetic field is increased to 4 T for a better jet energy resolution [36] at a higher centre-of-mass energy. Due to the different magnetic field strength, the iron yoke thickness is increased to 230 cm.

The CLIC_ILD adopts a similar very forward calorimetry system as that of the ILD. The dimensions of the elements are changed due to a difference in the beam crossing angles (20 mrad for CLIC and 14 mrad for the ILC). A comparison of the key parameters for the LumiCAL and the BeamCAL at the ILD and the CLIC_ILD is shown in table 3.5.

Chapter 4

Event generation, simulation, reconstruction, and analysis software

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

— Confucius, 551 BC – 479 BC

In this chapter, event generation. simulation, reconstruction, and analysis software for the future linear colliders are discussed. The event reconstruction focuses on the PandoraPFA event reconstruction, which is the framework for the photon reconstruction algorithms in chapter 5. The multivariate analysis (MVA) is discussed in details due to its complexity.

4.1 Event generation

Monte Carlo (MC) samples were generated for physics analyses in this thesis. Most events used in this thesis were generated with the WHIZARD software [49, 50]. Some simple events, such as single-photon-per-event samples, were generated by writing the event manually in the HEPEVT format [51]. The PYTHIA software [52] was used to describe parton showering, hadronisation and fragmentation. The fragmentation parameters for the PYTHIA were tuned to OPAL data [53] from the Large Electron-Positron Collider

(LEP). The TAUOLA software [54] was used to describe the tau lepton decay with correct spin correlations of the tau 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 with default parameters.

Particle masses and widths used to generate SM samples for studies with CLIC detectors, used in chapter 7, are listed in table 4.1.

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

Table 4.1: Particle masses and widths used for the generation of SM samples for CLIC detectors, taken from [2]. The Higgs boson mass is specified for individual samples.

4.1.1 CLIC 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 period, the total integrated luminosities of the electron−photon and photon−photon interactions are different to that of the electron−positron interaction.

As all events were generated assuming a total integrated luminosity of the electron−positron interaction, a correction in the total integrated luminosity is needed for electron-photon and photon-photon interactions where the photon is produced from initial state radiation via Beamstrahlung.

A simulated study [55] was performed to identify the ratios of the integrated luminosities of the positron−photon, electron−photon, and photon−photon interactions where initial-state photons are from Beamstrahlung, to the electron−positron interaction at CLIC. Results of the study is summarised in table 4.2. For the physics analysis in chapter

7, integrated luminosities for processes with initial-state photons from Beamstrahlung are corrected with the values in table 4.2.

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

Table 4.2: Luminosity ratios of total integrated luminosities of the positron−photon, electron−photon, and photon−photon interactions where initial-state photons are from Beamstrahlung, to the electron−positron interaction, for CLIC, at $\sqrt{s} = 1.4\,\text{TeV}$ and 3 TeV. Values are taken from [55].

4.2 Event Simulation

The simulation software used to simulate the interaction of particles through the detector material is the GEANT4 software [56]. The ILD and CLIC_ILD detector geometry descriptions are provided by the Mokka software [40]. The QGSP_BERT physics list from GEANT4 the is used to simulate the detailed development of hadronic showers in the detector. Since events were generated with head-on collisions, the event simulation introduced the crossing angles (20 mrad for CLIC and 14 mrad for ILC) by applying a corresponding Lorentz boost to all particles in the events.

4.2.1 CLIC beam induced backgrounds

It is necessary to include beam induced background to study CLIC detectors under realistic conditions. Two most significant types of backgrounds [2] in the CLIC colliding environment are $\gamma\gamma \to$ hadrons and incoherent e^+e^- pairs. The $\gamma\gamma \to$ hadrons background is produced when the interaction of real and virtual beamstrahlung photons from the colliding beams leads to two-photon interactions, resulting in hadronic final states [57, 58]. The incoherent e^+e^- pairs are produced with interactions of both real or virtual beamstrahlung photons with individual particles of the other beam, producing e^+e^- pairs in the strong electromagnetic fields [59].

Table 4.3 shows the amount of energies deposited from $\gamma\gamma \to$ hadrons and the incoherent pairs in different parts of the CLIC_ILD detector. The energies in the calorimeter are integrated over 300 ns from the start of the bunch train. The $\gamma\gamma \to$ hadrons is the dominant background in all calorimeters except the HCAL endcap. For the study presented in chapter 7, only the $\gamma\gamma \to$ hadrons background is included in the simulation.

Subdetector	Incoherent Pairs (TeV)	$\gamma\gamma \to$ hadrons (TeV)
ECAL Endcaps	2	11
ECAL Barrel	-	1.5
HCAL Endcaps	16	6
HCAL Barrel	-	0.3
Total Calorimeter	18	19
Central Tracker	-	7

Table 4.3: Amount of energies deposited from $\gamma\gamma \to$ hadrons and the incoherent pairs in the different parts of the CLIC_ILD subdetectors. Numbers correspond to the background for an entire CLIC bunch train and were obtained for nominal background rates. The reconstructed calorimeter energies are integrated over 300 ns from the start of the bunch train. The table is adapted from [2].

The simulation of $\gamma\gamma \to$ hadrons uses the photon spectrum from GUINEAPIG [60] and a parametrisation of the total cross section of the $\gamma\gamma \to$ hadrons process based on [61]. The average number of $\gamma\gamma \to$ hadrons events per bunch crossing within the detector acceptance at $\sqrt{s} = 3$ TeV is 3.2, for a $\gamma\gamma$ centre-of-mass energy greater than 2 GeV [62]. The PYTHIA software is used to simulate the hard interaction and the hadronisation of these $\gamma\gamma \to$ hadrons events.

The hits from simulated $\gamma\gamma \to$ hadrons events were superimposed to simulated e^+e^-, $e^\pm\gamma$, and $\gamma\gamma$ collisions before the event reconstruction. The $\gamma\gamma \to$ hadrons backgrounds included are resulted from 60 bunch crossings, corresponding to a time window of -5 ns to $+25$ ns around the generated physics event, with a 0.5 ns timing separation between bunch crossings to mimic the CLIC train structure [2]. For each bunch crossing, the number of $\gamma\gamma \to$ hadrons events superimposed were chosen from a Poisson distribution with a mean of 3.2.

4.3 Event Reconstruction

Reconstruction software runs in the MARLIN framework [63]. The event reconstruction contains following steps: digitisation of simulated calorimeter hits, reconstruction of tracks in the tracking system (using pattern recognition algorithms) [64], and particle flow objects (PFOs) reconstruction with PandoraPFA [36, 39].

Different MARLIN processors are used to reconstruct tracks: ClupatraProcessor [64] for tracks in the TPC, ForwardTrackingProcessor [64] for tracks in the FTD, and SiliconTrackingProcessor [64] for tracks in other silicon tracking detectors. A final MARLIN tracking processor, FullLDCTrackingProcessor [64], is used to combine tracks segments produced from individual processors.

4.3.1 PandoraPFA event reconstruction

The PandoraPFA event reconstruction is used in the studies for future e^+e^- linear colliders. Originally developed with the ILD detector concept [36], PandoraPFA has been adapted to the CLIC condition and shows its ability to deliver required energy resolutions [2,39]. There are over 60 e^+e^- linear collider specific reconstruction algorithms. Each algorithm aims to address a particular topological issue in the reconstruction. In the recent development, the core base codes for basic objects and memory managements were factorised into the Pandora C++ Software Development Kit [65].

In the subsequent sections, the main steps in the PandoraPFA reconstructions are summarised. The details of the PandoraPFA event reconstruction can be found in [36, 39, 65]. The inputs of the PandoraPFA event reconstruction are digitised calorimeter hits and reconstructed tracks, with some detector information, such as magnetic field strength, to aid the reconstruction. The output are reconstructed Particle Flow Objects (PFOs).

4.3.1.1 Track processing

Tracks from the inner tracking detectors are important inputs for the PandoraPFA reconstruction. A helical track fit using last 50 reconstructed hits in the tracking detector is performed to project the track onto the front of the ECAL. Afterwards, special topologies of tracks are identified based on the likely origin of the track. The topologies

of tracks include 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 topology of the "prongs" is also identified when a charged particle decays to multiple charged particles. This information about special topologies, along side with helical track fit, the track projection onto the front of the ECAL, and the original track parameters, is stored and passed onto the subsequent reconstruction.

4.3.1.2 Calorimeter hit processing

The other important inputs of the PandoraPFA reconstruction are the calorimeter hits from calorimeters. The properties of a calorimeter hit and the extra calculated information about calorimeter hits are stored and used in later steps. The properties of a calorimeter hit include its position, its layer in the calorimeter, and its energy response from the calorimeter digitiser. The extra calculated information about calorimeter hits includes likelihood of the hit originated from a minimum ionising particle (MIP).

Isolated hits, often originated from low energy neutrons in a hadronic shower, can be at a significant distance from the point of the production. Therefore, they are of little use to the PandoraPFA reconstruction as it is impossible to associate isolated hits to the correct hadronic shower. These hits are identified and not used during the clustering stage. However, these isolated hits participate in the reconstruction in the particle flow object (PFO) creation step to contribute to the energy estimation of the PFO.

4.3.1.3 Particle Identification

Dedicated particle identification algorithms find calorimeter hits associated with neutral particles, such as muons and photons. These calorimeter hits associated with muons and photons are removed from the subsequent reconstruction for charged particles. Identified muons and photons re-enter the reconstruction at the fragment removal stage (see section 4.3.1.8). Chapter 5 describes the photon reconstruction algorithms in details.

4.3.1.4 Clustering

Cone-based clustering algorithms are used to group calorimeter hits into clusters. The output clusters are further processed, merged, or split based on their topological properties.

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 direction of the cone. Afterwards the cone with the associated hits forms the cluster.

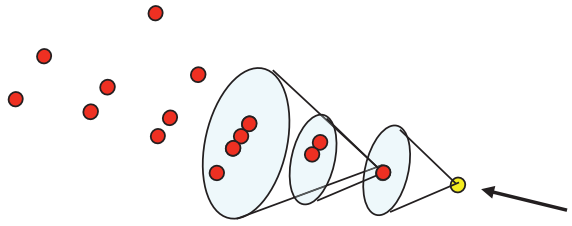


Figure 4.1: Illustration of the cone based clustering algorithm, taken from [66]

The cone-based clustering algorithm is used in PandoraPFA reconstruction 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 illustrates the cone-based clustering algorithm used in the PandoraPFA reconstruction. The seed for the cone clustering is typically the projection of a track onto the front of the ECAL. The initial cone direction is taken as the direction of the seed. Afterwards, a cone with a specified opening angle 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.

4.3.1.5 Topological cluster association

After the initial clustering, clusters are further refined using topological information of calorimeter hits in the calorimeters. The initial clustering scheme tends to form small

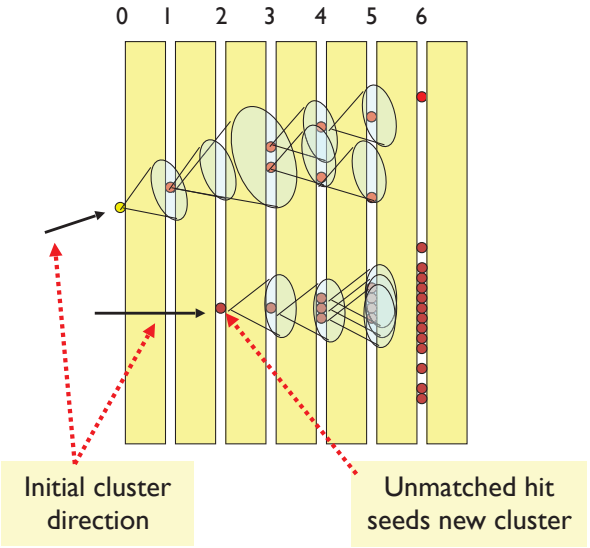


Figure 4.2: Illustration of the clustering algorithm used in the PandoraPFA, taken from [66]

clusters. These small clusters are then merged based on clear topological signatures. Main rules for topological association algorithms are schematically shown in figure 4.3. Some merging signatures include combining track segments, connecting track segments with gaps, connecting track segments to hadronic showers, and merging clusters when they are within close proximity.

4.3.1.6 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 detectors. The associations are made according to the proximity of the first layer of the cluster and the track projection onto the front of the ECAL, whilst demanding a match between the track direction and the initial cluster direction, and a match between the track momentum and the cluster energy.

4.3.1.7 Re-clustering

The track−cluster association scheme described in the previous section works well for events with jets of less than 50 GeV energies. In a dense jet environment with higher energy jets, electromagnetic and hadronic showers are boosted and are likely to overlap

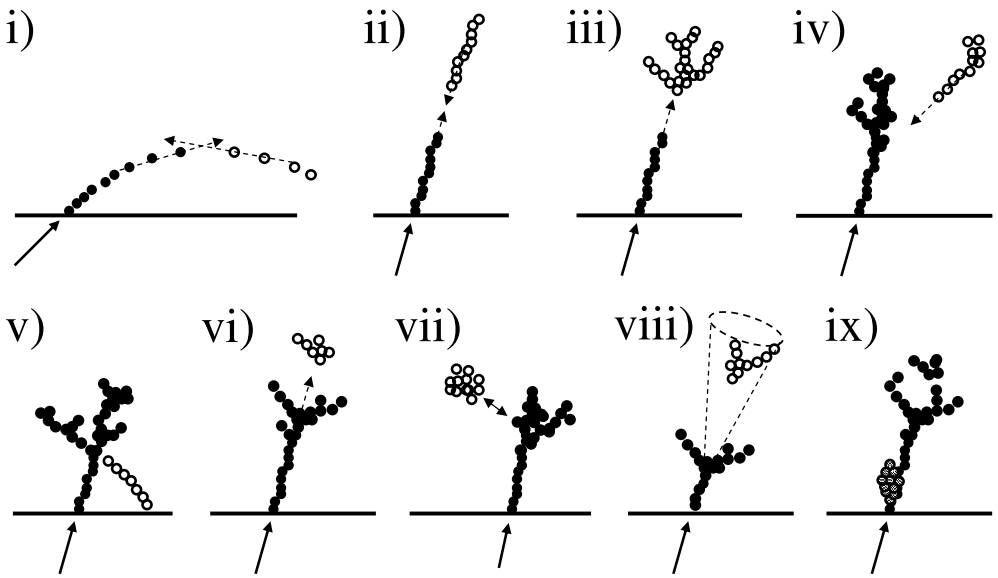


Figure 4.3: The main topological rules for cluster merging: i) looping track segments; ii) track segments with gaps; iii) track segments pointing to hadronic showers; iv) track-like neutral clusters pointing back to a hadronic shower; v) back-scattered tracks from hadronic showers; vi) neutral clusters which are close to a charged cluster; vii) a neutral cluster near to a charged cluster; viii) cone association; and ix) recovery of photons which overlap with a track segment. In each case the arrow indicates the track, the filled points represent the hits in the associated cluster and the open points represent the hits in the neutral cluster. Figures are taken from [36].

each other. Therefore, it is important to refine the track-cluster association based on the information from the momentum of the track and the energy of the cluster.

The re-clustering stage improves the compatibility of the cluster energy and the associated track momentum. It is performed on a statistical basis. If mismatch between the cluster energy and the associated track momentum is identified, 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. Out of many temporary clusters, the temporary cluster with the best track−momentum cluster−energy match is chosen, and that cluster is associated with the track.

A schematic diagram of the re-clustering stage is shown in figure 4.4. 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 of the cluster to track.

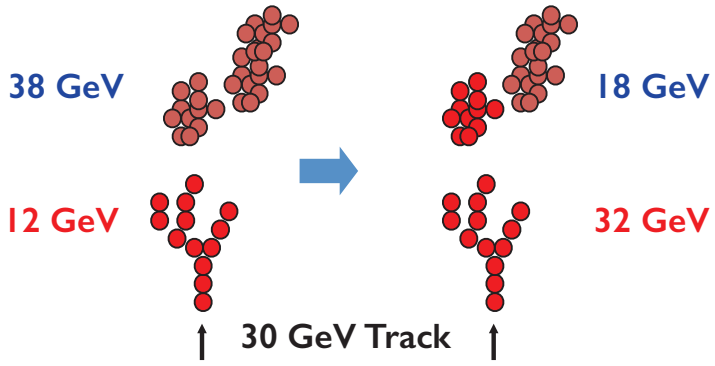


Figure 4.4: Illustration of the re-clustering algorithm in PandoraPFA, taken from [66]. Arrows indicate the tracks. Dark red dots represent the calorimeter hits in the associated cluster. Slightly fainter red dots represent the calorimeter hits in the neutral cluster. The initial cluster energy is less 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.

4.3.1.8 Fragment removal

This stage of the PandoraPFA reconstruction will focus on merging clusters that are likely to be fragments of other particles. Typically a fragment is merged if it is close to the parent cluster and the fragment has a low energy. Algorithms for photon fragment merging are described in details in chapter 5.

4.3.1.9 Particle Flow Object Creation

The last stage of the reconstruction is to create the output objects, Particle Flow Objects (PFOs). The PFOs contain clusters and associated tracks and calorimeter hits. Particle identification for electrons, muons, and photons are applied to PFOs.

4.3.2 CLIC beam induced backgrounds suppression

Following the discussion on CLIC beam induced background, two software have been developed to suppress $\gamma\gamma \rightarrow$ hadrons backgrounds: a track selector and a PFO selector [39].

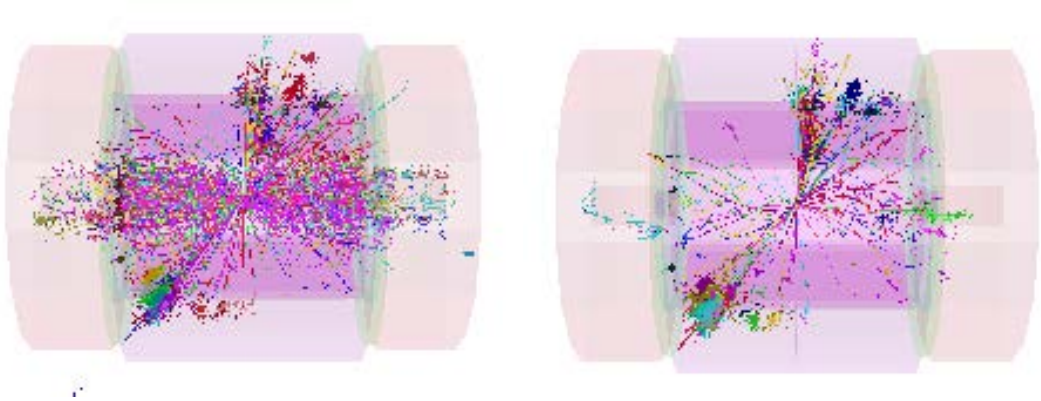
The track selector, CLICTrackSelector processor [39], removes poor quality and fake tracks that are likely from the beam induced background. It examines the number of track hits in individual tracking subdetectors and places a track-quality cut. It also places a cut on the arrival time of the track onto the front of the ECAL. 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 [39] 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 $\gamma\gamma \rightarrow$ hadrons often have low p_{T} and are distributed in time across the reconstruction integration timing window. In contrast, the PFOs from physics processes have a range of p_{T}, and have times close to the time of the brunch crossing that contains the physics event.

For the best performance of the background suppression, the PFO selector uses different p_{T} and timing cuts for the central part of the detector and for the forward part of the detector. The PFO selector also uses different p_{T} and timing cuts for different types of particles: photons, neutral PFOs, and charged PFOs.

Three configurations of these cuts are developed: “loose”, “normal”, and “tight” PFO selections. As the name suggested, “loose” PFO selection corresponds to a looser cut of p_{T} and time, preserving PFOs with a larger range of p_{T} and a larger range of times than the “tight” PFO selection.

The optimal configuration of the PFO selection to suppress the backgrounds 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. Figure 4.5a shows reconstructed particles in a simulated $e^+e^- \to HH \to t\bar{b}b\bar{t}$ event which 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. The effect of applying tight PFO selection 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.



(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 [39].

4.4 Analysis software

4.4.1 Monte Carlo truth linker

It is extremely useful to be able to associate reconstructed objects to the Monte Carlo 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 MC particle contributing most to a reconstructed PFO or a group of PFOs (a jet) can be determined.

4.4.2 Jet algorithms

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 [67], where descriptions on later developments can be found in reviews [68–70].

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 4.3.1.4 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 with the seed and into the jet. 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 used in this thesis is the FastJet C++ software package [71,72]. The notations in the subsequent discussion follow the convention in [71].

4.4.2.1 Longitudinally invariant k_t algorithm

Longitudinally invariant k_t algorithm [73,74] 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 $y_i = \frac{1}{2}\ln\frac{E_i + p_{zi}}{E_i - p_{zi}}$ and ϕ_i are particle i's rapidity and azimuthal angle. The free parameter R controls the jet radius.

The d_{ij} and d_{iB} are calculated for all pairs of particles. If the minimum value of all d_{ij} and d_{iB} is a d_{ij}, particle i and j are merged and the four momentum of particle i is updated as the sum of the two particles. If the minimum value is a d_{iB}, particle i is set to be a final jet and removed from the list of particles. The above procedure is repeated until no particles are left.

The exclusive variant is similar to the inclusive variant. First difference is that when the minimum value is a d_{iB}, particle i forms part of the beam jet. The beam jet contains particles that are considered to be from the beam induced background and discarded at the end of the jet clustering. The second difference is that when all d_{ij} and d_{iB} are above a threshold, d_{cut}, the jet clustering will stop.

The exclusive mode allows a specified number of jets to be found, where the d_{cut} is automatically determined. In contrast, the inclusive mode would find as many jets as the algorithm allows.

4.4.2.2 Durham algorithm

The Durham algorithm [75], also known as e^+e^- k_t algorithm, is commonly used in the e^+e^- collider experiments. It only has one pair-wise distance metric:

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

where E_i is the energy of particle i and θ_{ij} is the angle 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 a threshold, d_{cut}.

Compared to the longitudinally invariant k_t algorithm, the Durham algorithm 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. And the total energy of the event is known, which is the same as the collision energy (for events with missing momentum).

4.4.2.3 Jet algorithm for CLIC

Although CLIC is a e^+e^- collider, the beam-induced background is significant and deposits a large amount of energy in the detector. Therefore, traditional e^+e^- jet algorithms, like the Durham algorithm, are not suitable for the CLIC colliding environment. Studies [2,76] have shown that jet algorithms for the pp colliders give better performances for the CLIC colliding environment. Therefore, longitudinally invariant k_t algorithm is often used in analyses with the CLIC environment.

4.4.2.4 The y parameter

The y parameter is a measure of the number of jets in an event. It describes the transition of going from N clustered jets to $N+1$ clustered jets using an exclusive jet algorithm. For example, y_{23} would be the d_{cut} value for an 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 a small number to a machine acceptable range is to take negative logarithm of the number.

4.5 Multivariate Analysis

Multivariate analysis (MVA) has become increasingly important in high energy physics. MVA is typically used in physics analysis to classify signal events from background events. Compared to the traditional cut-based method, modern machine learning techniques offer much improvement to data analysis. The implementation of the machine learning based MVA used in this thesis is provided by TMVA software [77].

MVA can be used for classification or regression. Classification classifies a testing event into one of several classes. Regression of a testing event gives an output in a

continuous numerical range. The focus in this section is on the classification, as the MVA is often used to select one type of events from other type of events in a physics analysis.

A typical MVA classification involves two classes, sometimes referring to as the signal class and the background class. Before using the MVA classification, a machine learning model (classifier) needs to be trained with training data. The model uses a set of discriminative variables as inputs, which have different distributions for the signal and the background. To use the MVA classification, the trained model will be applied onto the testing data. The classification response of the model on a testing sample is a two-class outcome: the signal or background.

This two-class classification scheme can be easily 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 6.5 and in the flavour tagging classifier in section 7.5.

4.5.1 Optimisation and overfitting

One important concept with the MVA is the optimisation and the overfitting of the model. 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 well, but the model would not be optimal for a testing sample. A simple model is less prone to statistical fluctuation of samples, however, the model might be too simple to achieve the optimal modelling. The former case is known as overfitting, or overtraining. The latter case is called underfitting, or undertraining.

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

One definition of overfitting is when the efficiency of the signal selection in the training samples increases, but the efficiency of the signal selection in the testing sample decreases, with the increase of the model complexity. Figure 4.6 shows the signal selection efficiency as a function of the model complexity, using an example 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 fraction of the signal selected when the background fraction is 1%, reported by the TMVA training process. The depth of 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 a tree depth of six onwards, overfitting occurs. In this particular example, one should choose a tree depth fewer than seven to avoid overfitting.

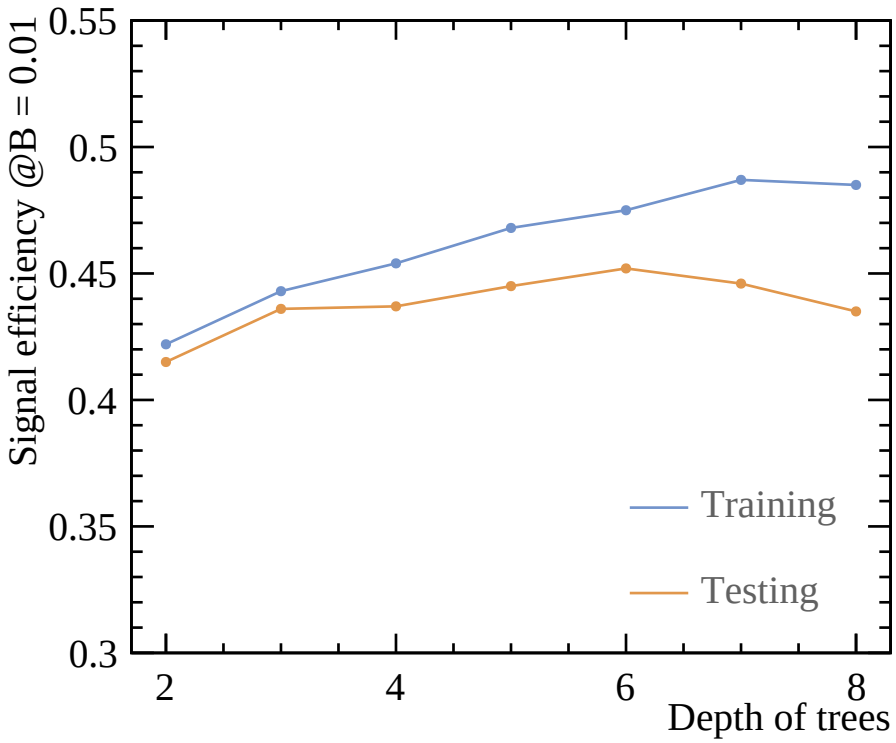


Figure 4.6: Example of the signal selection efficiency as function of the model complexity. The example is chosen from the double Higgs analysis at $\sqrt{s} = 3\,\text{TeV}$, using the Boosted Decision Tree model. The efficiency of the signal selection is defined as the fraction of the signal selected when the background fraction is 1%, reported by the TMVA training process. The depth of the tree reflects the complexity of the model. From a tree depth of six onwards, overfitting occurs.

4.5.2 Choice of models

The model to fit the data 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 to the input variables.

To rigourously identify the best model, individual optimisation of models are required, which is computationally very expensive. However, as researchers in the machine learning suggested, the boosted decision tree is probably the best out-of-the-box machine learning method [78]. A neutral network model could potentially perform better than the boosted decision tree model, but it requires more tuning, and it is less intuitive to interpret. For these reasons, the boost decision tree model (BDT) is chosen for to be used in physics analyses in this thesis. Before describing the BDT model in details, some simpler models will be described.

4.5.3 Rectangular Cut model

The rectangular cut method, probably the most intuitive model, optimises cuts to maximise some pre-defined metrics. The metric could be the signal efficiency for a given background efficiency. Alternatively, the metric can be the significance, $\frac{S}{\sqrt{S+B}}$, where S and B are respective numbers of signal and background passing the rectangular cuts.

Discriminative variables give better separation power when they are gaussian-like and statistically independent [78]. Therefore it is preferable to decorrelate the variables and gaussian transform them before using the rectangular cut MVA.

4.5.4 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 of the projective likelihood model is discussed in section 5.5.

The likelihood classifier calculates the probability density for each discriminative variable, for the signal class and the background class (hence PDE approach). The overall signal and background likelihood are defined as products of the individual probability density of each variable, for the respective signal class and background class. The likelihood ratio, R, can be then defined as the signal likelihood divided by the sum of the e signal likelihood and the background likelihood.

Similar to the rectangular cut method, the likelihood model works better with decorrelated, gaussian-like variables.

4.5.5 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 splitting node, the splitting point, uses a cut on 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"), depending on whether the event passes the cut. At each splitting node, samples are divided into two categories: signal-like and background-like sub-samples. For each sub-sample, this splitting process, tree growing, starts at the root node, and stops after certain criterion are met. The root node is the first splitting node. The stopping criterion could be the minimum number of events in a node, the maximum number of layers of the tree, or a minimum/maximum signal purity of the end nodes. The end nodes, where the tree stops growing and the sub-samples are not split, contains signal and/or background events. If there are more with more signal than background events in an end node, it is called a signal-like end node. The opposite is called a background-like end node.

The training of the decision tree refers to finding the optimal cut at the splitting node by minimising a given 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 trained decision tree. The signal class is the PhD student and the background class is the undergraduate student. The attributes of samples are listed table 4.4. The top diamond box, "Party ends before 1am", is the root node. All diamond boxes are splitting nodes. Rectangle boxes are end nodes. The signal-like end node is represented by the red rectangle and the background-like end nodes are represented by blue rectangles. The depth of this decision tree is two. The metric to find the optimal cut at the splinting node is the Gini index.

To demonstrate first step of the training of the model, details of finding the optimal cut at the root node are outlined. There are two possible cuts for the root node, "Party ends before (after) 1am" and "(Not) Know where free pizza is". If the cut at the root node is "Leave party 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 leave parties before 1 am. Gini index is given by

$$2p(1-p) \backsimeq 0.36. \tag{4.4}$$

If the cut at the root node 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 is located. Gini index is given by

$$2p(1-p) \backsimeq 0.44. \tag{4.5}$$

Therefore, by choosing the cut that minimise the Gini Index, the optimal cut for the root node is "Leave party 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 a student who leaves parties before 1 am and knows where a free pizza is located, then the student is classified as a PhD student.

PhD students	Leave party before 1 am	Leave party after 1 am
Know where free pizza is	10	0
Not know where free pizza is	0	0
Undergraduates	**Leave party before 1 am**	**Leave party after 1 am**
Know where free pizza is	0	5
Not know where free pizza is	3	2

Table 4.4: The attributes of the PhD students and undergraduate students for the decision tree example shown in figure 4.7.

4.5.5.1 To improve decision tree

It is very easy to construct a decision tree that fits the training data very well, but the tree would not be optimal for the testing sample. To overcome the instability of the

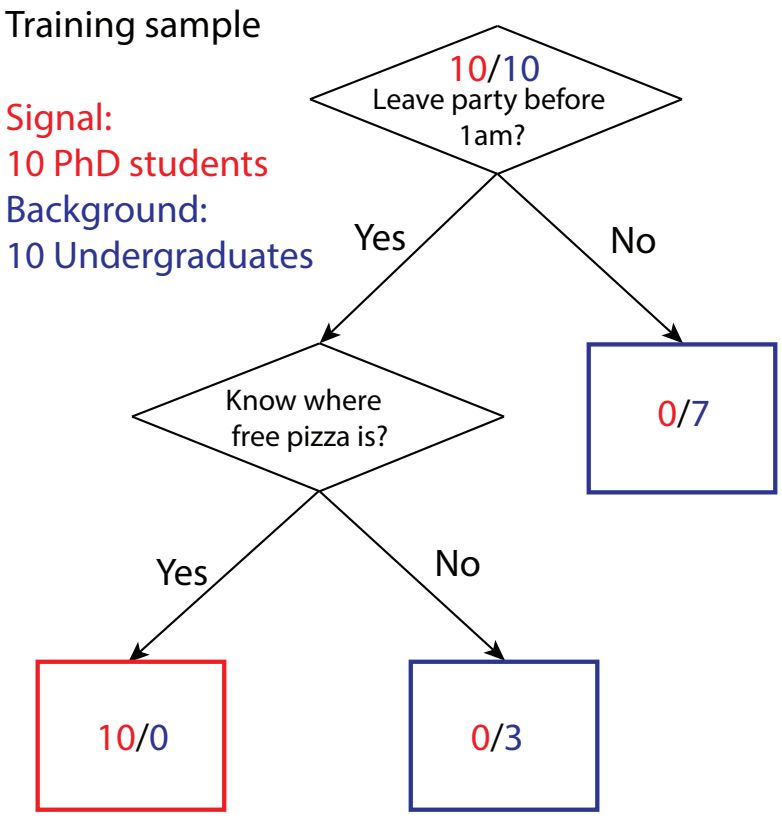


Figure 4.7: Example of a decision tree. Numbers in each node represent number of PhD students (red) and number of undergraduate students (blue). Diamond boxes represent splitting nodes. Rectangular boxes represent end nodes. Blue boxes are background-like end nodes. Red boxes are signal-like end-nodes.

decision tree model, many methods have been developed. Some of the most successful ones are boosting, bagging, and random forest.

Boosting: the basic idea of boosting is that the tree growing 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.

Bagging: also known as boot-strap, it is a method that select a random subset of the training sample, and use the subset at training.

Random Forest: when a tree is grown, a randomly selected subset of discriminative variables are used to grow the tree.

4.5.6 Boosted Decision Tree model

Boosted decision tree (BDT) contains a forest of decision trees, where the boosting is used to grow trees. There are two common boosting methods: adaptive boosting and gradient boosting. The adaptive boosting, first introduced in [79], is discussed in further details, as it is simpler to understand than the gradient boosting. The adaptive boosting algorithm, adapted from [78], 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, where M is the total number of trees. For iteration m:
  - Create/grow a m^{th} tree with weighted samples obtained from $(m-1)^{th}$ iteration.
  - 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 in m^{th} tree, $w_{i,m} = w_{i,m-1} e^{\alpha_m B_{i,m}}$.
- The output, $G(x)$, for a testing event x, is a weighted vote from all M trees:

$$G(x) = \begin{cases} -1, & \text{if } \sum_{m=1}^{M} \alpha_m G_m(x) < 0, \\ 1, & \text{otherwise.} \end{cases} \tag{4.6}$$

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). The parameter B represents if an event is misclassified. For the i^{th} event in the m^{th} tree, $B_{i,m} = 1$ if the event is misclassified, or 0 if the event is correctly classified. The parameter $w_{i,m}$ represent the event weight for i^{th} event in m^{th} tree.

In each iteration, if the i^{th} event is misclassified, the event 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 decision tree with a few layer would be a weak classifier. By sequentially applying many weak classifiers with weighted samples, the final "forest" is very robust with a very good performance.

4.5.6.1 Optimisation of Boosted Decision Tree

The most important parameter is the depth of a tree, which determines how many end nodes the tree has. It also affects the complexity of the BDT model. If the depth of a tree is set to a large value, it could leads to the overfitting of the model.

The number of trees is another important parameter. Past studies show that using many small trees yields the best result [78]. However, intuitively a large number of trees leads to overfitting. But it has been shown that a large number does not lead to overfitting [78]. 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 paramter.

The boosting algorithm has two variants in TMVA implementation: adaptive boost and gradient boost.

The learning rate of the adaptive boost controls how fast the event weight changes in each boosting iteration. Studies show that a small learning rate ($\sim$0.1) with many trees works better than a large learning rate with fewer trees [78].

The shrinkage rate in the gradient boost is similar to the learning rate parameter 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 [78].

The usual choice of the metric to optimise cuts for tree growing is either the Gini index or the cross-entropy. The two different metrics make little differences to model performances.

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, has little impact on the model performance. But because variables are binned, variables should be pre-processed before feeding into the training model. For example, the variable should be limited to a range to avoid the extreme values that distort the variable distribution. If the original distribution of the variable is highly skewed, the variable should be transformed to obtain a more uniform distribution.

For the end node, the output can be either signal-like or background-like, based on the majority of the training events in the end node. Numerically, it can correspond 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 model performance improves.

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

4.5.7 Multiple classes

The above discussion assumes two classes, the signal class and the background class. The classification 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. The second way is called “one versus all”, when each class is trained against all other classes combined.

Using a three-class example, A, B, and C class, “one versus one" scheme trains A class against B class; B class against C class; and C class against A class. “One versus

all" scheme would train A class against non-A classes; B class against non-B classes; and C class against non-C classes.

TMVA multiple class implementation, MULTICLASS, uses the “one versus all" scheme. For each class, the MULTICLASS classifier will train the class against all other classes. This process is repeated for each class, resulting in multiple classifiers. The overall classifier output for a single event is a normalised response using all trained classifiers, where the sum of the classifier outputs for a single event is one. Individual response of a trained classifier for an event can be treated as the likelihood. In the applying stage, the event is classified into a class if the classifier for that class gives the highest output response amongst all classifiers for that event.

The advantage of using the MULTICLASS classifier instead of a two-class classifier for samples with multiple classes is the classifier outputs are correctly adjusted for multiple classes. Hence one event can be unambiguously classified into only one class. The issue with the MULTICLASS classifier is that powerful discriminative variables for each individual class need to enter the training stage simultaneously, resulting in a large number of discriminative variables in the MULTICLASS classifier.

Chapter 5

Photon Reconstruction in PandoraPFA

'I dreamed I was a butterfly, flitting around in the sky; then I awoke. Now I wonder: Am I a man who dreamt of being a butterfly, or am I a butterfly dreaming that I am a man?'

— Zhuang Zi, 369 BC – 286 BC

Many aspects of the photon reconstruction are important. A good single photon energy resolution and the ability to reconstruct two spatially close photons are necessary to reconstruct particles using decay channels involving photons, such as $\pi^0 \to \gamma\gamma$.

The ability to correctly reconstruct photons in a dense jet environments improves the charged particle reconstruction by simplifying the pattern recognition problem for the charged particle reconstruction.

The photon reconstruction algorithms presented in this chapter have benefited many physics analyses. The most recent example of such a physics study is the $H \to \gamma\gamma$ simulation study at CLIC [80].

This chapter starts with an overview of the electromagnetic shower produced by photons passing through a thick absorber. It then discusses the photon reconstruction algorithms within the PandoraPFA framework, followed by a description of the performances of these algorithms. Part of this chapter has been published in the proceedings of 2015 International Workshop on Future Linear Colliders [81].

5.1 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 pair production and bremsstrahlung generate many low-energy photons and electrons, producing shower-like structures in the detector. Two suitable length scales to describe the EM shower are the radiation length and the Molière radius [82, 83].

The radiation length of a material describes the EM longitudinal shower profile, defined as the mean distance travelled by an electron where an electron loses its energy by a factor of $1/e$ via bremsstrahlung, also the $7/9$ of the mean free path for pair production by a high energy photon [84].

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 [85] :

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

where t is the number of radiation lengths; the parameter E_0 is the shower energy; the parameter b varies slightly with material but it is sufficient to use $b = 0.5$ for the purpose of photon reconstruction [6]; and the parameter a is given by [36]:

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

where E_c is the critical energy. The critical energy is defined as the energy of the electron at which the rate of losing energy by bremsstrahlung is the same as the rate of losing energy by ionisation [86]. The alternative definition of the critical energy is the energy at which the energy loss by ionisation per radiation length is the same of the particle energy [87]. This parametrisation of the EM longitudinal shower profile should only be used to describe an average behaviour of the EM shower, as the fluctuation of the individual EM shower profile is significant.

The EM transverse shower profile can be described as a narrow cone widening as the shower develops. 90% of the shower energy is contained in a fiducial cylinder with a radius of one Molière radius, along the direction of the shower.

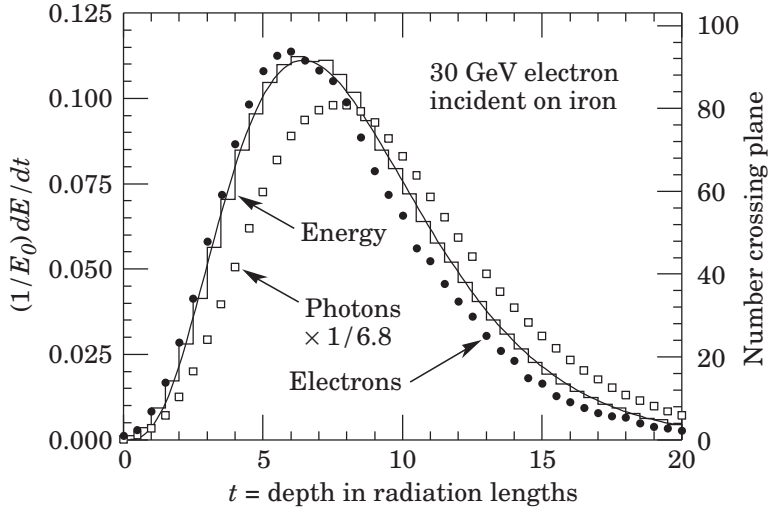


Figure 5.1: An EGS4 simulation of a 30 GeV electron-induced electromagnetic shower in iron. The histogram shows fractional energy deposition as a function of radiation lengths, and the curve is a gamma-function fit to the distribution. Circles and squares are the number of electrons and photons respectively with total energy greater than 1.5 MeV crossing planes with scale on right. Plot is taken from [6].

5.2 Overview of photon reconstruction in PandoraPFA

Five algorithms are developed to tackle different issues in the photon reconstruction. The most important photon algorithm is the Photon Reconstruction algorithm. It 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.

Three algorithms 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. The last photon algorithm is a photon splitting algorithm. The algorithm separates accidently merged photons.

5.3 Photon Reconstruction algorithm

The Photon Reconstruction algorithm is a photon reconstruction algorithm at the early stage of the reconstruction. It corresponds to "Particle ID" stage of the

PandoraPFA reconstruction chain, described in section 4.3.1.3. Main steps of the Photon Reconstruction algorithm, shown in figure 5.2, are: forming photon clusters; finding photon candidates; photon ID test; and optional fragments removal.

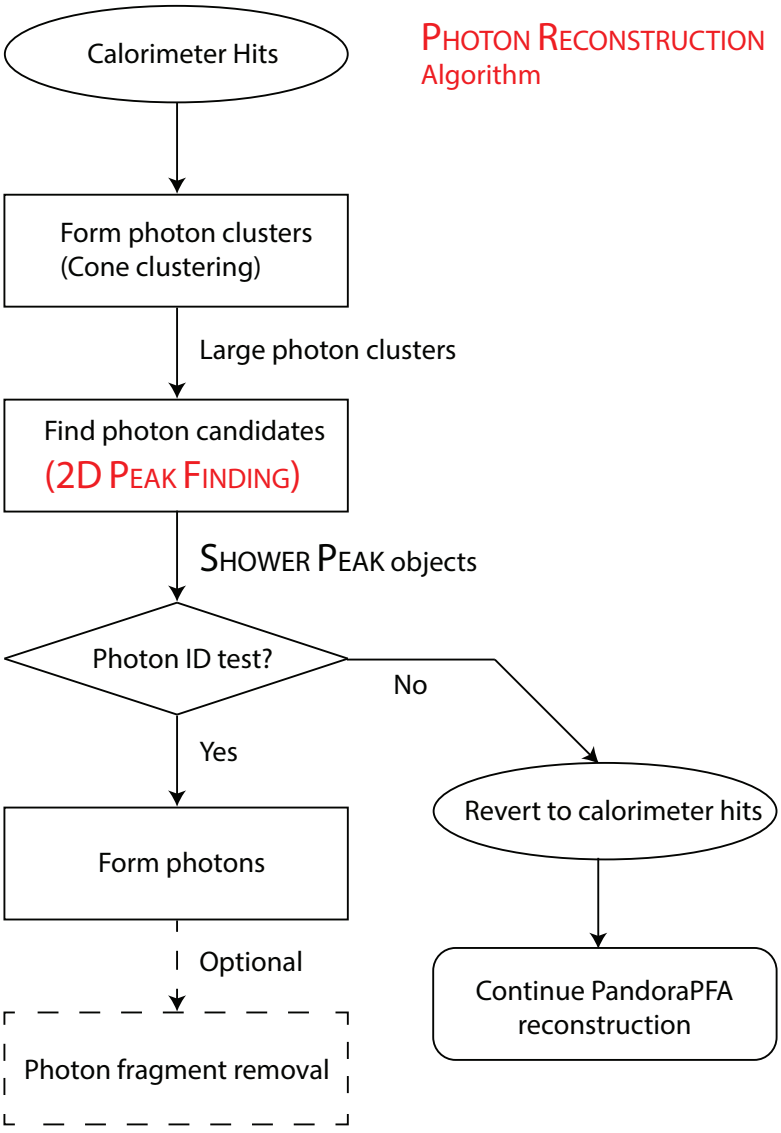


Figure 5.2: Main steps of the Photon Reconstruction algorithm: forming photon clusters; finding photon candidates; photon ID test; and optional fragments removal.

5.3.1 Forming 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 algorithms form muons and remove calorimeter hits associated with muons from the

reconstruction. The calorimeter hits associated with reconstructed muons are not used to form photons.

This step forms photon clusters from calorimeter hits in the ECAL. The clusters are formed in a way such that calorimeter hits from one photon would not be split into two clusters, but one cluster may contain calorimeter hits from multiple photons. The algorithm uses the cone clustering algorithm provided inside PandoraPFA to form clusters. Since the target for reconstruction is the neutral photon, the cone clustering algorithm uses high-energy calorimeter hits in the ECAL as initial seeds, instead of using track projections as initial seeds.

5.3.2 Finding photon candidates

This step refines photon clusters into smaller photon candidates. Each photon candidate should contain calorimeter hits from one photon only.

A three-dimensional cluster is split into several smaller photon candidates, if the cluster contains several photons. The three-dimensional splitting problem is harder than a two-dimensional one. Therefore, a translation is needed to map the three-dimensional problem to a more manageable two-dimensional problem. This translation relies on the characteristic EM transverse shower profile. Along the direction of the shower, an EM shower can be modelled as a dense shower core with peripheral calorimeter hits around the core. When the energies of the calorimeter hits of the cluster are projected onto a two-dimensional plane, an EM shower core would appear as a mountain-like structure in the plane. Figure 5.3 shows an example of a photon cluster projected onto a two-dimensional plane, where two EM showers are identified. Hence, by identifying a peak in the two-dimensional plane, the EM shower core is identified.

A high-performance two-dimensional peak-finding algorithm is the key to identify multiple photon candidates within a cluster. Due the complexity of the peak finding procedure, a peak-finding algorithm is developed and discussed in section 5.4. The output of the two-dimensional peak-finding algorithm is a collection of ShowerPeak objects. Each ShowerPeak object corresponds to one photon candidate and associated calorimeter hits.

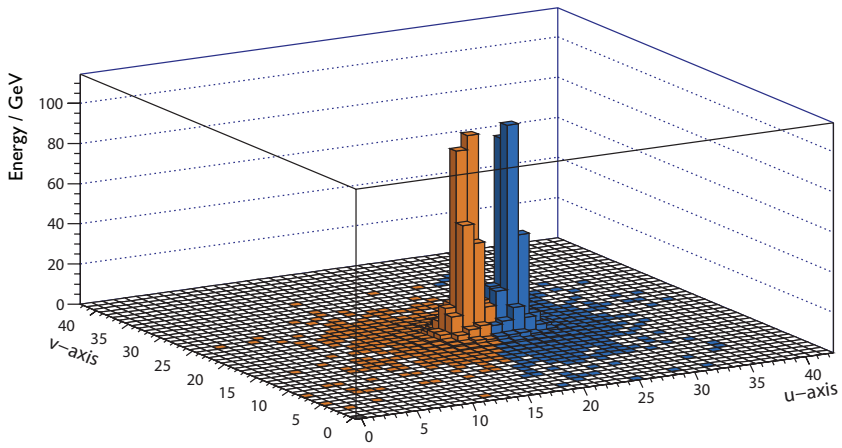


Figure 5.3: Two 500 GeV photons (yellow and blue) belonged to a photon cluster, just resolved in a transverse plane orthogonal to the direction of the flight. The axes U and V are orthogonal axes in units of the ECAL cell sizes. The height of a bin in the histogram is the sum of the calorimeter hit energy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

5.3.3 Photon ID test

This step applies the photon ID test on the Shower Peak object. The photon ID test uses a multidimensional likelihood classifier. A set of variables, which exploit features of electromagnetic showers, are used. The response from the classifier determines if a Shower Peak object is a photon. If it is a photon, the Shower Peak object would be tagged as a photon and the Shower Peak object does not participate in the subsequent event reconstruction. The identified photon re-enters the event reconstruction at the fragment removal stage after the charged particle reconstruction. If a Shower Peak object is not a photon, the Shower Peak object will be discarded. Calorimeter hits associated with discarded the Shower Peak object will be passed onto the next stage of the reconstruction. The likelihood classifier is further discussed in section 5.5

5.3.4 Photon Fragment removal

The photon fragment removal algorithm merges small photon fragments to identified photons. The algorithm is optional as it is not used by the default setting of the event reconstruction. Since this algorithm shares the same logic as another fragment removal algorithm, two algorithms are discussed togethers in section 5.6.

This step marks the end of the PHOTON RECONSTRUCTION algorithm. The outputs are a collection of reconstructed photons, separated from non-photon calorimeter hits.

5.4 Two-dimensional peak-finding algorithm

As discussed in section 5.3.2, identifying photon candidates inside a cluster is translated to identifying peaks in a two-dimensional plane, using a two-dimensional peak-finding algorithm (2D PEAK FINDING algorithm). The 2D PEAK FINDING algorithm aims to correctly identify peak positions in a two-dimensional histogram and to associate calorimeter hits to identified peaks.

There are two variants of the 2D PEAK FINDING algorithm: the neutral cluster variant and the charged cluster variant. The base algorithm is the neutral cluster variant. The charged cluster variant is used when the cluster is close to the projection of a track onto the front of the ECAL. Main steps of the neutral cluster variant is shown in figure 5.4: initialising a two-dimensional histogram; projecting calorimeter hits to the histogram; identifying local peaks; associating non-peak bins to peaks; filtering peaks; and forming SHOWER PEAK objects.

5.4.1 Initialising two-dimensional histogram

This step initialises a two-dimensional (2D) histogram to host the projection of the calorimeter hits of the 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 the axes and the direction of the cluster form an orthogonal basis in the three-dimensional space.

5.4.2 Projecting calorimeter hits to histogram

This step projects the calorimeter hits associated with the cluster onto the 2D histogram initialised in the previous step. The projection is chosen such that the cluster centroid position is projected onto the centre of the histogram. The relative distance between the calorimeter hit position and the cluster centroid position is converted into a distance

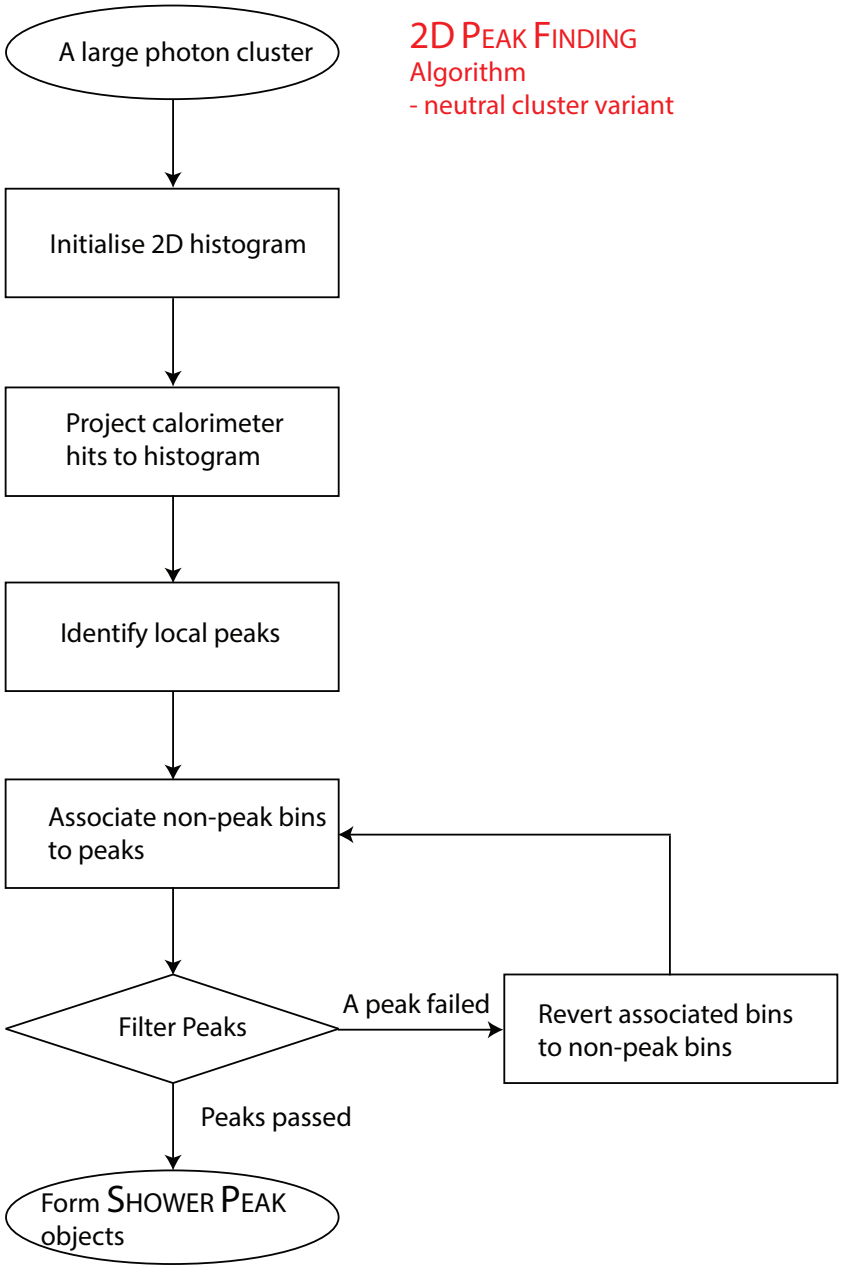


Figure 5.4: Main steps of the neutral cluster variant of 2D PEAK FINDING algorithm: initialising a two-dimensional histogram; projecting calorimeter hits to the histogram; identifying local peaks; associating non-peak bins to peaks; filtering peaks; and forming SHOWER PEAK objects.

vector. The distance vector, $\vec{s_i}$, of a calorimeter 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 calorimeter hit i; $\langle\vec{a}\rangle$ is the centroid position of cluster a; and d_{cell} is the ECAL square cell length. The coordinate of the calorimeter hit projection onto the histogram is calculated from the scalar products of the distance vector with the axes vectors.

The height of a bin in the 2D histogram is the sum of the energies associated with the calorimeter hits that fall in that particular bin. Each bin contains calorimeter hits that projected onto the bin. One bin size along either axes on the 2D histogram corresponds to one ECAL square cell length.

5.4.3 Identifying local peaks

This step identifies all local peaks in the 2D histogram. A local peak is defined as a bin where its height is above all eight neighbouring bins. The 2D histogram is scanned to identify all local peaks.

5.4.4 Associating non-peak bins to peaks

Having identified all local peaks, this step associates non-peak bins to a particular peak based on the energy of the peak and the distance of the non-peak bin to the peak bin. A non-peak bin should be associated to a high-energy peak bin that is close to the non-peak bin.

The peak bin to associate a non-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 2D histogram, and E_i is the height (energy) of the peak bin i. For each non-peak bin, the metric is iterated over all peak bins.

5.4.5 Filtering peaks

The performance of the 2D Peak Finding algorithm is improved by peak filtering. In a 2D histogram, such as the one in figure 5.3, major peaks with many associated non-peak bins most likely correspond to physical photons, while the minor peaks with few associated non-peak bins likely come from fluctuations in the energy deposition of the EM shower. To discard minor peaks, every time after all non-peak bins are associated with peak bins, minor peaks with fewer than three bins associated (including the peak bin) are discarded. These discarded bins are re-associated with other peak bins. This process iterates until all peak bins have at least three bins associated.

The Shower Peak object is created after filtering peaks. One Shower Peak object contains one peak bin and associated non-peak bins. The associated calorimeter hits within 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.

5.4.6 2D Peak Finding algorithm charged cluster variant

If a photon candidate is close to the projection of the track onto the front of the ECAL, it is more likely that the candidate is a charged hadron. Misidentifying a charged hadron as a photon leads to a significant degradation in the reconstruction performance, because there will be double counting of energies from the track and from the charged hadron misidentified as a photon. However, if a photon next to a charged hadron is carefully reconstructed, the charged particle reconstruction is improved.

This step aims to carefully identify photon candidates next to charged hadrons, by using track information and features of EM showers. The important information is that the EM shower typically starts 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 of 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 onto the front of the ECAL, charged cluster variant of the 2D Peak Finding algorithm is applied.

The "Create and iterate sub-clusters" step performs the following. The ECAL is sliced longitudinally to create fiducial spaces. For example, the default three slices will result

in three ECAL fiducial volumes. Each fiducial volume covers the space from the front of the ECAL to a third, two thirds, and the back of the ECAL. Three sub-clusters are created from calorimeter hits that are contained in each fiducial volume.

After creating sub-clusters, the neutral cluster variant of the 2D Peak Finding algorithm is applied to each sub-cluster. The sub-cluster in the first third of the ECAL is processed first. The sub-cluster in the whole of the ECAL is processed lasted. For each sub-cluster, a collection of Shower Peak objects are created from the 2D Peak Finding algorithm.

The Shower Peak objects created from each sub-cluster undergo the "Shower Peak filter" step. All peaks from the first sub-cluster are preserved. For the next sub-cluster, a peak is only preserved if the peak bin position can be linked to a peak bin position in the previous sub-cluster, allowing a shift in the peak bin position by no more than one neighbouring bin. Furthermore, if the peak bin is within one neighbouring bin of a track projection bin, the peak is discarded. Only the peaks that are preserved in every sub-cluster through the iteration of "Shower Peak filter" step will be used to form the final Shower Peak objects.

5.4.7 Inclusive mode

The 2D histogram is iterated many times during the algorithm. The time complexity of iterating the histogram is $O(n^2)$ for a n bins by n bins sized histogram (default $n = 41$). Therefore, for the purpose of speed, it is undesirable to have a large number of bins. Having a small finite-sized histogram speeds up the computation. However, because of the finite size of the histogram, only calorimeter hits projected onto the histogram would be considered for the peak finding algorithm. Calorimeter hits projected outside the histogram would not be used 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. However, for photon splitting, all calorimeter hits from the parent photon should be used to form daughter photons. Hence the inclusive mode of the 2D Peak Finding algorithm is developed, and allows calorimeter hits projected outside the histogram to be associated with identified peaks.

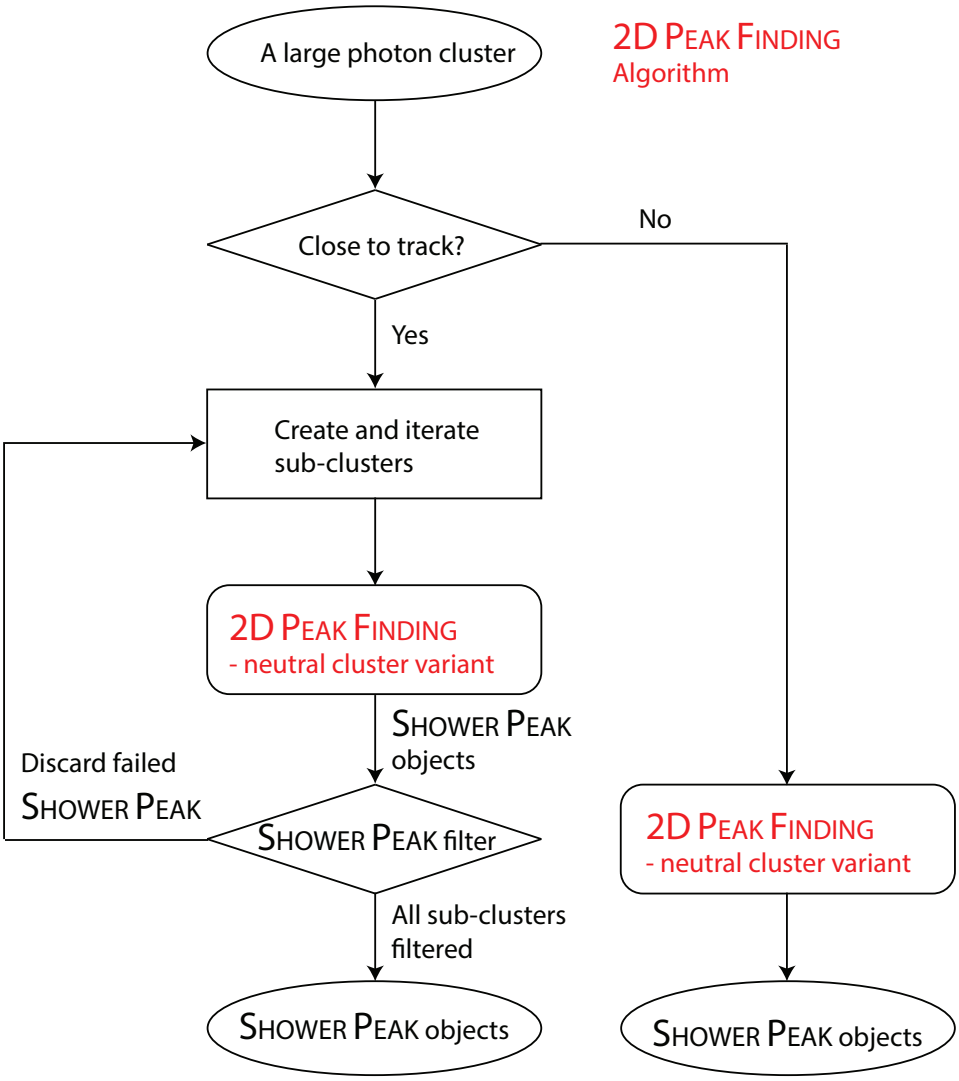


Figure 5.5: Main steps of the 2D Peak Finding algorithm, including the charged cluster variant: identifying whether the cluster is close to a track; create and iterate sub-clusters; apply 2D Peak Finding algorithm neutral cluster variant to sub-clusters; filter Shower Peak objects in sub-clusters; create final Shower Peak objects.

5.5 Likelihood classifier for photon ID

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

5.5.1 Likelihood classifier variables

Variables used in the likelihood classifier exploit the differences between a characteristic electromagnetic shower and a hadronic shower, and the fact that a photon is less likely to be close to track projections onto the front of the ECAL, than a cluster of a charged paritcle. Variables used in the classifier are listed in table 5.1.

Two variables are obtained from the EM longitudinal shower profile: the variable t_0 is the start layer from the longitudinal shower profile, shown in figure 5.6a; and δl is fractional difference of the observed shower profile to the expected EM shower profile described in equation 5.1:

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

where the quantity 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 true photon, t_0 and δl are expected to be small, as an EM shower should start in the first few layers of the ECAL and the shower profile should be similar to an expected EM shower profile.

Three variables are obtained from the transverse EM shower profile: the variable $\langle w \rangle$ is the energy weighted root-mean-square distance of all bins in a Shower Peak to its peak bin, a measure of the transverse shower size, shown in figure 5.6c; the variable $\delta\langle w_{UV} \rangle$ is the smallest ratio of the two energy weighted root-mean-square distances of all bins in a Shower Peak to its peak bin in each of the U, V axis direction, a measure of the circularity of the transverse shower; the last variable, $\delta E_{cluster}$, is the ratio between the energy of the Shower Peak object to the cluster energy, a measure of the dominance of a Shower Peak in a cluster.

The last variable used in the classifier, d, is the distance between the candidate and the closest track projection onto 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 EM shower profile	δl,t_0
Transverse EM shower profile	$\langle w \rangle$,$\delta\langle w_{UV} \rangle$, $\delta E_{cluster}$
Distance to track	d

Table 5.1: Variables used in the likelihood classifier for photon ID test.

5.5.2 Projective Likelihood classifier

Projective likelihood classifier with probability density estimators is used for the photon ID due to its low requirement on computing resources, comparing to a boost decision tree classifier or a neutral network classifier.

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 energy ranges. There are 8 photon energy ranges, obtained by binning the distribution of photon energies at 0.2, 0.5, 1, 1.5, 2.5, 5, 10, 20 GeV. The variable distributions for non-photon are binned in the same energy ranges, according to the energy of the non-photon.

The training stage of the classifier uses simulated $e^+e^- \to Z'Z'$ samples, where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$, at a centre-of-mass energy of 500 GeV. The events at centre-of-mass energy of 500 GeV allow the training of photon with energies greater than 20 GeV.

In the applying stage of the classifier, for a given candidate with the candidate energy in the energy bin α, the classifier output is given by

$$\mathrm{PID}_\alpha = \frac{N \prod_i^6 P_i}{N \prod_i^6 P_i + N' \prod_i^6 P'_i} \tag{5.6}$$

where P_i and P'_i are the probability of the candidate fallen in the respective photon and non-photon i^{th} variable probability distributions in the energy bin α; the variables N

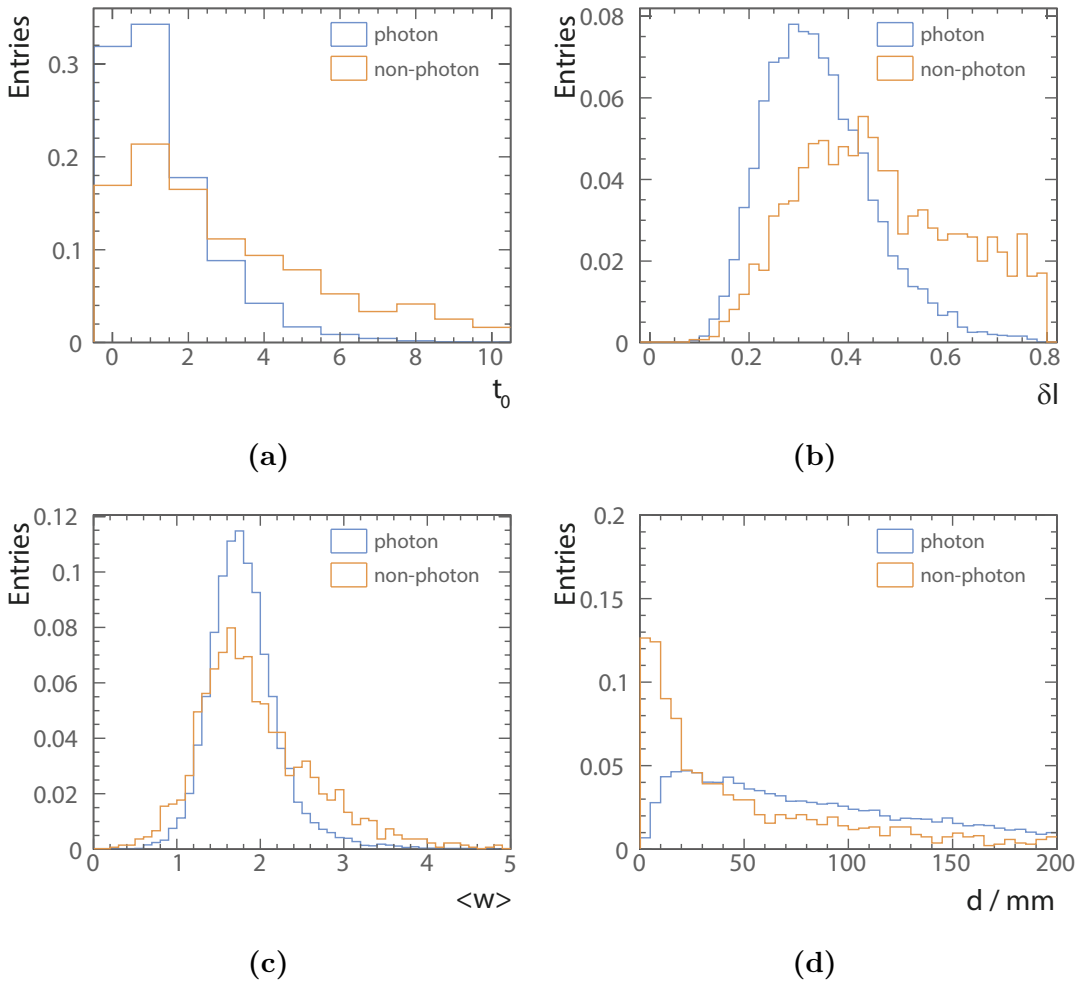


Figure 5.6: Distributions for a) the start layer from the longitudinal shower profile (t_0), b) the fractional difference of the observed shower profile to the expected EM shower profile (δl), c) the energy weighted root-mean-square distance of all bins in a ShowerPeak to its peak bin ($\langle w \rangle$), and d) the distance between the photon candidate and the closest track projection onto the front of the ECAL (d) are shown. All plots are normalised that the area under curve is 1. The particle ID is determined using the truth information. All plots are generated with simulated $e^+e^- \to Z'Z'$ samples, where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$, at $\sqrt{s} = 500\,\text{GeV}$.

and N' are the number of respective photons and non-photons in the energy bin α in the training sample.

During applying stage of the classifier, a candidate passes the photon ID test if

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

where E is the candidate energy. Two values of the cuts on PID is because it is more likely to misidentify a low-energy particle as a photon. A low-energy EM shower does not have a dense shower core, and is more difficult to identify. Hence for candidates with energy between 0.2 and 0.5 GeV, PID > 0.6 is required instead of PID > 0.4.

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 (the fragment), and identified as a photon or a neural hadron. Figure 5.7 shows a typical creation of such a photon fragment. A fragment typically does not have the electromagnetic shower structure, and has much lower energy than a main photon.

(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 two algorithms share same logics for merging, the algorithm used after the charged particle reconstruction will be discussed here.

A photon and a fragment form a 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.

Depending on whether the fragment is reconstructed as a photon or a neutral hadron, the photon−fragment pairs is further classified into photon−photon-fragment pairs and photon−neutral-hadron-fragment pairs, because they have different kinematic and topological distributions. The pairs are subsequently classified into low energy and high energy pairs, depending on whether the fragment energy (E_p) is above 1 GeV.

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. The description of each variable used in the cuts will be provided first, followed by the description of the logics of the cuts.

5.6.1 Variables

There are three distance variables: the variable d_c gives the distance between centroids of the PFO in the photon−fragment pair, which is a computationally quick measurement; the variable d_h is the minimum distance between calorimeter hits of each PFO in the pair; the variable d is the mean energy weighted intra-layer distance between each PFO in the pair, illustrated schematically in 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} layer of the ECAL; the parameter d_l^i is the minimum distance between calorimeter hits of the pair in the i^{th} layer; and E_f^i is the energy of the fragment in the i^{th} layer of the ECAL.

Other quantities used in the merging metric include: E_m, the energy of the main photon; E_f, the energy of the fragment; E_{p1} and E_{p2}, the energies of the two most energetic EM showers, identified by the 2D Peak Finding algorithm, ordered by descending energy, using the photon−fragment pair as input; N_{calo}, the number of the

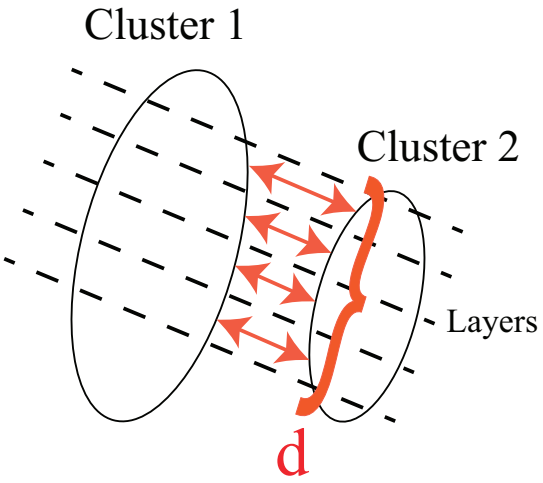


Figure 5.8: Illustration of mean energy weighted intra-layer distance between each PFO in the pair, d.

calorimeter hits in ECAL in the fragment; and $|\cos(\theta_Z)|$, the absolute value of the cosine of the polar angle of the main photon with respect to the beam direction.

Here each set of logics for merging fragments are discussed, using the cuts for photon−photon-fragment with fragment energy $< 1\,\text{GeV}$ as an example. Fragments passing any one set of cuts will be merged.

5.6.2 Transverse shower comparison cuts

One logic for merging is when the photon−fragment pair looks like one EM shower 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 most energetic peak found by the 2D Peak Finding algorithm. It also demands that the second energetic peak should have less than half of the fragment energy, $\frac{E_{p2}}{E_f} < 0.5$. And the most energetic peak should have more energies than the main photon, $E_{p1} > E_m$. Lastly the fragment should be close to the main photon, $d < 30\,\text{mm}$.

5.6.3 Low energy fragment cuts

The logic for merging is when the fragment has a low energy and is spatially close to the main photon: $d < 20\,\text{mm}$ and the energy of the fragment is less than $0.2\,\text{GeV}$.

5.6.4 Small fragment cuts

Fragments that are spatially close to the main photon and have very few number of associated calorimeter hits will be merged. Two sets of cuts are developed. Either the pair satisfies $d < 30\,\text{mm}$; $d_c < 50\,\text{mm}$; and number of calorimeter hits in the fragment less than 40, or the pair satisfies $d < 30\,\text{mm}$ and number of calorimeter hits in the fragment less than 50. The multiple sets of cuts allow the merging of a fragment with fewer number of calorimeter hits with a slightly larger distance separation to the main photon, or the merging of a fragment with a slightly bigger number of calorimeter hits with a smaller distance separation to the main photon.

5.6.5 Small fragment forward region cuts

This logic merges low-energy fragment in the end cap region of the detector. The cut demands: $d_c < 60\,\text{mm}$; $|\cos(\theta_Z)| > 0.7$; the energy of the fragment less than $0.6\,\text{GeV}$; and the number of calorimeter hits in the fragment less than 40.

5.6.6 Relative low energy fragment cuts

The merged fragment should be relatively low energetic. The distance between the pair should satisfies $d < 40\,\text{mm}$ and $d_h < 20\,\text{mm}$. The ratio of the fragment energy to the main photon energy should be less than 0.01.

Cuts for high-energy fragments (E_f >1 GeV) only has logics for transverse shower comparison and relative low energy fragment, as the cut on the absolute low-energy fragment is not applicable for the high-energy fragments.

Neutral hadron fragments originated from charged particles are more likely to have low energies, but high-energy neutral hadron fragments are more likely to be originated from photons. Hence cuts for photon−neutral-hadron-fragment pair for low-energy fragment only merge fragments that are very close to the main photon, with very few calorimeter hits, or has a relative very small energy. The cuts for photon−neutral-hadron-fragment pair for high-energy fragment, on the other hand, are more generous, allow merging fragments that has energy of up to 20% of the main photon energy.

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 the smallest distance metric, d, will be merged.

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.

5.6.7 Photon fragment recovery algorithm after the Photon Reconstruction algorithm

The photon fragment removal algorithm immediately after the PHOTON RECONSTRUCTION algorithm shares the same logics as the stated above. It differs sightly in the values of the cuts, which are stricter as the algorithm is careful not to make mistakes at merging fragments. The cuts for merging fragments are listed in table 5.3.

5.7 Photon fragment recovery algorithm in the HCAL

There is another type of fragments originated from the leakage effect of the ECAL. When a high-energy EM shower is not fully contained in the ECAL, the shower deposits energy in the HCAL, which often forms a neutral hadron in the HCAL. An example of a 500 GeV photon reconstructed into a main photon in the ECAL (yellow) and a neutral hadron fragment in the HCAL (blue) is shown in figure 5.9. This section presents an algorithm to merge fragments in the HCAL to the main photon.

Photon fragments in the HCAL are spatially close to the main photon. A cone obtained from fitting the main photon, if extended to the HCAL, should contain most of the calorimeter hits of the fragment. These features allow a set of cuts developed to merge fragments in the HCAL, which are listed in table 5.4.

This algorithm uses photons in the ECAL and neutral hadrons in the HCAL as inputs. The algorithm then iterates over all pairs of reconstructed photons and neutral

$E_f \leqslant 1\,\text{GeV}$	Photon−photon	Photon−neutral-hadron
Transverse shower comparison, or	$d < 30\,\text{mm}$; $\frac{E_{p1}}{E_m+E_f} > 0.9$; $\frac{E_{p2}}{E_f} < 0.5$; $E_{p1} > E_m$	-
Low energy fragment, or	$d < 20\,\text{mm}$; $E_f < 0.4\,\text{GeV}$	$d < 20\,\text{mm}$; $d_c < 40\,\text{mm}$
Small fragment 1, or	$d < 30\,\text{mm}$; $N_{calo} < 40$; $d_c < 50\,\text{mm}$	$d < 50\,\text{mm}$; $N_{calo} < 10$; $d_h < 50\,\text{mm}$
Small fragment 2, or	$d < 50\,\text{mm}$; $N_{calo} < 20$	-
Small fragment forward region, or	$N_{calo} < 40$; $d_c < 60\,\text{mm}$; $E_f < 0.6\,\text{GeV}$; $\lvert\cos(\theta_Z)\rvert > 0.7$	-
Relative low energy fragment	$d < 40\,\text{mm}$; $d_h < 20\,\text{mm}$; $\frac{E_f}{E_m} < 0.01$	$d < 40\,\text{mm}$; $d_h < 15\,\text{mm}$; $\frac{E_f}{E_m} < 0.01$
$E_f > 1\,\text{GeV}$	Photon−photon	Photon−neutral-hadron
Transverse shower comparison, or	$\frac{E_{p1}}{E_m+E_f} > 0.9$; $E_{p2} = 0$ 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, or	$d < 40\,\text{mm}$; $d_h < 20\,\text{mm}$; $\frac{E_f}{E_m} < 0.02$	$d < 40\,\text{mm}$; $d_h < 20\,\text{mm}$; $\frac{E_f}{E_m} < 0.02$
Relative low energy fragment 2, or	-	$d < 40\,\text{mm}$; $d_h < 20\,\text{mm}$; $\frac{E_f}{E_m} < 0.1$; $E_f > 10\,\text{GeV}$
Relative low energy fragment 3	-	$d < 20\,\text{mm}$; $d_h < 20\,\text{mm}$; $\frac{E_f}{E_m} < 0.2$; $E_f > 10\,\text{GeV}$

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, after charged hadron reconstruction. Variables 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. Variables E_m and E_f are the main photon energy and the fragment energy, respectively. Variables E_{p1} and E_{p2} are the energies the two largest peaks, found by 2D 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.

$E_f \leqslant 1\,\text{GeV}$	Photon−photon	Photon−neutral-hadron
Transverse shower comparison, or	$d < 20\,\text{mm}$; $\frac{E_{p1}}{E_m+E_f} > 0.9$; $E_{p2} = 0$ or ($\frac{E_{p2}}{E_f} < 0.5$, $E_{p1} > E_m$)	$d < 20\,\text{mm}$; $\frac{E_{p1}}{E_m+E_f} > 0.9$; $E_{p2} = 0$ or ($\frac{E_{p2}}{E_f} < 0.5$, $E_{p1} > E_m$)
Low energy fragment, or	$d < 20\,\text{mm}$; $E_f < 0.2\,\text{GeV}$	-
Small fragment 1, or	$d < 30\,\text{mm}$; $N_{calo} < 20$; $d_h < 13\,\text{mm}$	$d < 50\,\text{mm}$; $N_{calo} < 10$; $d_h < 50\,\text{mm}$
Small fragment 2, or	$d_c < 30\,\text{mm}$; $N_{calo} < 10$; $d_h < 13\,\text{mm}$	-
Relative low energy fragment	-	$d < 40\,\text{mm}$; $d_h < 15\,\text{mm}$; $\frac{E_f}{E_m} < 0.01$
$E_f > 1\,\text{GeV}$	Photon−photon	Photon−neutral-hadron
Transverse shower comparison, or	$d < 20\,\text{mm}$; $\frac{E_{p1}}{E_m+E_f} > 0.9$; $E_{p2} = 0$ or ($\frac{E_{p2}}{E_f} < 0.5$, $E_{p1} > E_m$)	$d < 20\,\text{mm}$; $\frac{E_{p1}}{E_m+E_f} > 0.9$; $E_{p2} = 0$ or ($\frac{E_{p2}}{E_f} < 0.5$, $E_{p1} > E_m$)
Relative low energy fragment	-	$d < 40\,\text{mm}$; $d_h < 20\,\text{mm}$; $\frac{E_f}{E_m} < 0.02$

Table 5.3: The cuts for merging photon−photon-fragment pairs and photon−neutral-hadron-fragment pairs for both low energy and high energy fragments, immediately after photon reconstruction. Variables 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. Variables E_m and E_f are the main photon energy and the fragment energy, respectively. Variables E_{p1} and E_{p2} are the energies the two largest peaks, found by 2D Peak Finding algorithm, ordered by descending energy, respectively. N_{calo} is the number of the calorimeter hits in the fragment.

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).

hadrons. For each pair, a set of variables are calculated and compared to a set of cuts. Photon−fragment pairs passing all the cuts will be merged.

5.7.1 Distance comparison cuts

Fragments in the HCAL should be spatially close to the main photon, measured by three metrics: the variable d_c^l is the distance between the centroid position of the calorimeter hits of the main photon in the last outer layer in the ECAL, and the centroid position of the calorimeter hits of the fragment in the first inner layer of the HCAL; the variable d_{fit}^l is the shortest distance between the direction fitted with the calorimeter hits of the main photon in the last outer layer in the ECAL, and the direction fitted with the calorimeter hits of the fragment in the first inner layer of the HCAL; and d_{fit} is the shortest distance between the direction fitted with the main photon, and the direction fitted with the fragment. These three distances should be small for merging. The cut demands $d_c^l \leqslant 173\,\text{mm}$; $d_{fit}^l \leqslant 100\,\text{mm}$; and $d_{fit} \leqslant 100\,\text{mm}$.

5.7.2 Projection comparison cuts

The direction of the fragment should be similar to the direction of the main photon. The variable r_f is the energy weighted root-mean-square distance of a calorimeter hit in the fragment to the direction fitted with the main photon. The cut requires $r_f \leqslant 45\,\text{mm}$.

5.7.3 Shower width comparison cuts

The shower widths of the fragment and the main photon should be similar. Variable w_m^l is the root-mean-square widths of the calorimeter hits of the main photon in last outer layer in the ECAL. Variable w_f^l is the root-mean-square widths of the calorimeter hits of the fragment in the first inner layer in the HCAL, respectively. The ratio $\frac{w_f^l}{w_m^l}$ needs to be in the range from 0.3 to 5 for the merging. The generous upper bound is because the HCAL cell size is much larger than the cell size of the ECAL.

5.7.4 Cone comparison cuts

When a cone obtained by fitting the main photon in the ECAL is extended to the fragment in the HCAL, the cone should contain a significant amount of the fragment. The variable, $\%N$, the fraction of the calorimeter hits in the fragment in the cone comparing to the calorimeter hits in the fragment, has to be greater than 0.5 for merging.

5.7.5 Energy comparison cuts

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

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

5.8 Photon splitting algorithm

Another aspect in photon reconstruction is to split 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 event reconstruction addresses this issue and tries to split merged photons.

Photon fragment recovery	Cuts
Distance comparison	$d_c^l \leqslant 173\,\text{mm}$; $d_{fit}^l \leqslant 100\,\text{mm}$; $d_{fit} \leqslant 100\,\text{mm}$
Projection comparison	$r_f \leqslant 45\,\text{mm}$
Shower width comparison	$0.3 \leqslant \frac{w_f^l}{w_m^l} \leqslant 5$
Cone comparison	$\%N \geqslant 0.5$
Energy comparison	$\frac{E_f}{E_m} \leqslant 0.1$

Table 5.4: The cuts for merging high energy photon fragment in the HCAL to the main photon in the ECAL. The variable d_c^l is the distance between the centroid position of the calorimeter hits of the main photon in the last outer layer in the ECAL and the centroid position of the calorimeter hits of the fragment in the first inner layer of the HCAL. The variable d_{fit}^l is the shortest distance between the direction fitted with the calorimeter hits of the main photon in the last outer layer in the ECAL and the direction fitted with the calorimeter hits of the fragment in the first inner layer of the HCAL. The variable d_{fit} is the shortest distance between the direction fitted with the main photon and the direction fitted with the fragment. The variable r_f is the root-mean-square energy weighted distance of a calorimeter hit in the fragment to the direction fitted with the main photon. Variables w_m^l and w_f^l are the root-mean-square widths of the calorimeter hits of the main photon in last outer layer in the ECAL, and the calorimeter hit of the fragment in the first inner layer in the HCAL, respectively. Variable $\%N$ is the fraction of the calorimeter hits in the fragment in the cone comparing to the calorimeter hits in the fragment. Variables E_m and E_f are the energy of the main photon and the energy of the fragment, respectively.

If a photon has the topologies of a spatially closed photon pair, the photon should be split. Extra care should be taken if the photon is close to a charged track projection onto the front of the ECAL.

Table 5.5 lists the cuts used in the algorithm. If an energetic photon is identified, the 2D Peak Finding algorithm will be used to identify EM showers in the photon using the transverse shower information. If energy of the photon is bigger than a threshold, E_{c1}, and the energy of the 2^{nd} energetic EM shower is bigger than another threshold, E_{c2}, the photon will be split according to the number of EM showers identified by the 2D Peak Finding algorithm.

The values of E_{c1} and E_{c2} depends on whether the photon is close to a charged track projection onto the front of the ECAL. The cut demands higher energises of the photon and the second energetic EM shower, if the photon is close to the track projection. The number of nearby charged tracks is counted as number of tracks with the track projection onto the front of the ECAL fewer than 100 mm to the photon centroid position. If there is no nearby tracks, the E_{c1} is set to 10 GeV and E_{c2} is set to 1 GeV. If there is one nearby track, the E_{c1} is set to 10 GeV and E_{c2} is set to 5 GeV. If there is more than one nearby track, the E_{c1} is set to 20 GeV and E_{c2} is set to 10 GeV.

The constraint on N_p, the number of EM showers identified in the photon, should be less than five, as one reconstructed photon is unlikely to be accidentally 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 tracks nearby	$E_{c1} = 10\,\text{GeV}$, $E_{c2} = 1\,\text{GeV}$
1 charged tracks nearby	$E_{c1} = 10\,\text{GeV}$, $E_{c2} = 5\,\text{GeV}$
> 1 charged tracks nearby	$E_{c1} = 20\,\text{GeV}$, $E_{c2} = 10\,\text{GeV}$

Table 5.5: Cuts used in the photon splitting algorithm. The parameter E is the photon energy. The parameter E_{p2} is energy of the second energetic peak obtained from 2D Peak Finding algorithm. The parameter N_p is the number of peaks identified by 2D Peak Finding algorithm. The parameters E_{c1} and E_{c2} are the energy threshold values, determined by the number of nearby charged PFOs to the photon.

5.9 Characterising the performance

Three different versions of the PandoraPFA are used to characterise the performance of the photon algorithms:

- with 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,

Without stand-alone photon reconstruction algorithms, PandoraPFA applies a simple photon ID at the end of the event reconstruction. In PandoraPFA version 1, there is a rudimentary photon reconstruction algorithm. In PandoraPFA version 3 contains all the photon algorithms presented in this chapter, as the algorithms were developed during PandoraPFA version 2. In PandoraPFA version 3, the photon algorithms have replaced the photon reconstruction algorithm in PandoraPFA version 1.

Firstly the performance with the full algorithms implemented in PandoraPFA version 3 is compared with the performance with no photon algorithms. Afterwards, the performance with the full algorithms is compared with the performance obtained from PandoraPFA version 1. The performances of individual photon algorithms are then characterised, followed by the characterisation of the performance of the photon algorithms in PandoraPFA version 3.

5.10 Comparing with no photon reconstruction

This section compares the performance with and without photon related algorithms using two-photon-per-event and jet samples. The two-photon-per-event samples 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 photon pair. Events are selected such that there is no early photon conversion in the tracking detector and the photon does not escape the detector. The events are further restricted to photon decaying in barrel and endcap region only, to avoid the barrel/endcap overlap region. The nominal ILD detector model is used to simulate the events.

Figure 5.10 shows the average number of reconstructed photons as a function of MC distance separation between two photons, using two-photon-per-event samples with photon energies of 500 GeV and 50 GeV, reconstructed with and without photon algorithms. Without the photon related algorithms, fragments are produced. Without the photon related algorithms, the number of photon fluctuates between 1 and 1.5 for a distance separation of 0 to 30 mm. With the photon related algorithms, two photons start to be resolved at 10 mm and fully resolved at 20 mm separation. The average number of reconstructed photon is 2 at 20 mm distance separation.

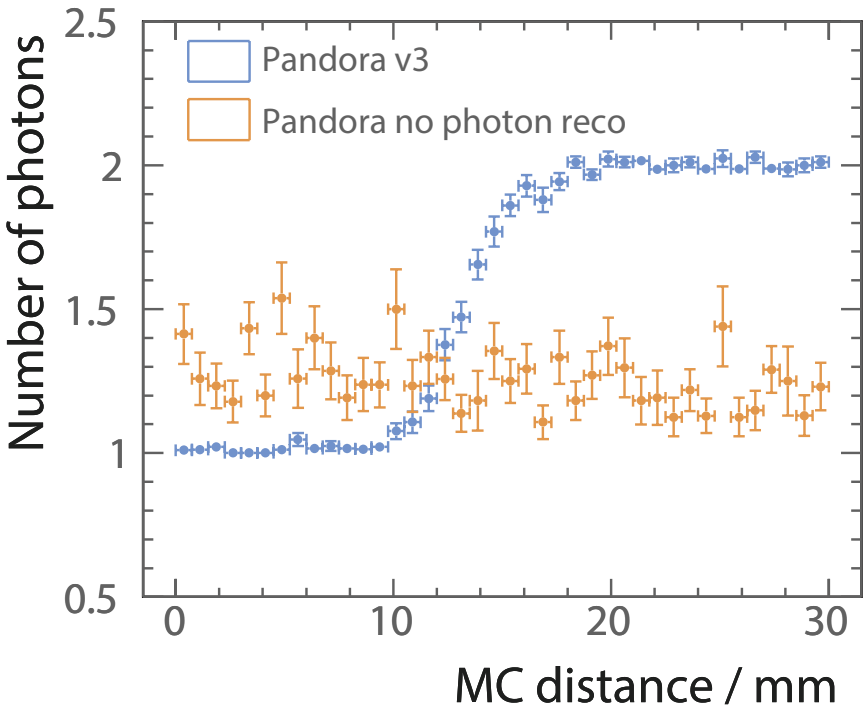


Figure 5.10: Average number of reconstructed photons using two-photon-per-event samples with photon energies of 500 GeV and 50 GeV, without (orange) and with (blue) photon algorithms, as a function of the Monte Carlo distance separation between the photon pair.

The improvement in photon reconstruction leads to a considerable improvement in the jet energy resolution. Jet energy resolution is defined as the root-mean-square divided by the mean for the smallest width of distribution that contains 90% of entries, using $e^+e^- \to Z'Z'$ samples, where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$, at barrel region. The angular cut is to avoid the barrel/endcap overlap region. The light quark decay of the Z' is used to avoid the complication of missing momentum from semi-leptonic decay of heavy quarks. Using 90%

of the entries is robust and focus on the Gaussian part of the jet energy distribution. The total jet energy is sampled at the centre-of-mass energies of 91, 200, 360 and 500 GeV.

As shown in figure 5.11, the jet energy resolutions are much better at $\sqrt{s} = 360\,\text{GeV}$ and 500 GeV with photon algorithms. By identifying photons before reconstructing charged particles in a dense jet environment, there are fewer calorimeter hits left for the charged particle reconstruction. However, at $\sqrt{s} = 91\,\text{GeV}$ and 200 GeV, the jet energy resolution is worse with photon algorithms, because photon algorithms are developed with jet environments at a centre-of-mass energy of 500 GeV.

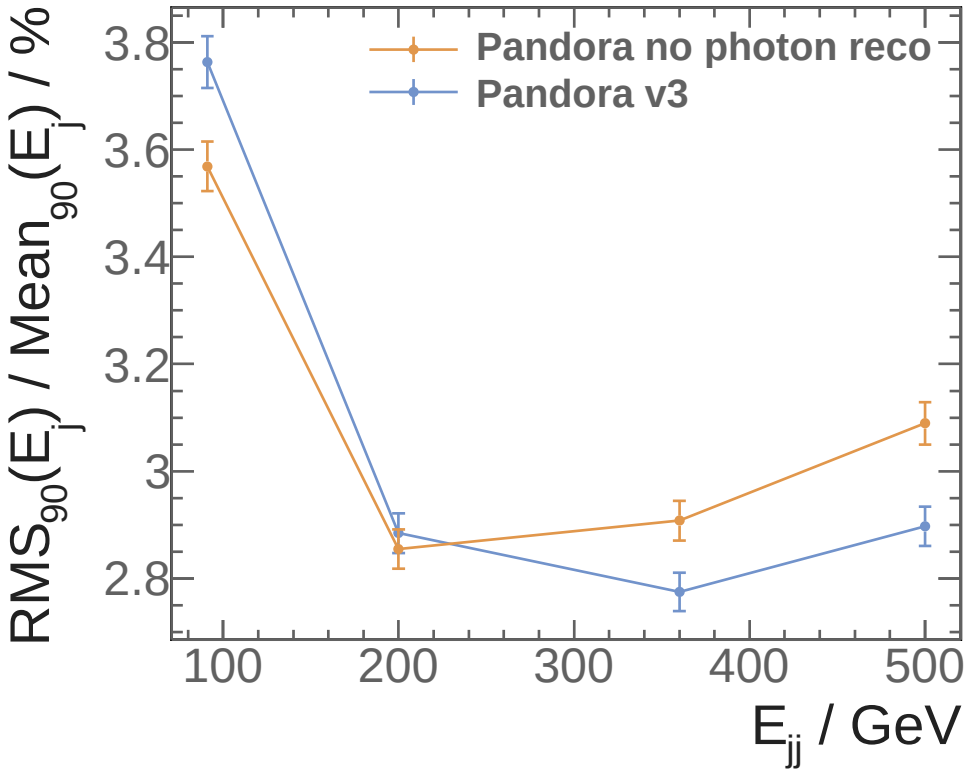


Figure 5.11: Jet energy resolution as a function of the total jet energy using $e^+e^- \to Z'Z'$ samples, where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$, at barrel region. The orange and bottom lines represent the reconstruction without and with photon algorithms, respectively.

To quantify the impact of photon algorithms on jet energy resolution, the same jet samples were reconstructed with the perfect photon reconstruction, which identifies photons by associating calorimeter hits using the MC truth information. The photon confusion terms, which are defined as the quadrature differences of the jet energy resolution between a non-cheated reconstruction and a perfect photon reconstruction,

are listed in table 5.6. The photon confusion terms, except for $\sqrt{s} = 91\,\text{GeV}$, have 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 5.6: Photon confusions as a function of total jet energies in the $e^+e^- \to Z'Z'$ samples, where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$, for reconstruction with and without photon algorithms.

5.11 Comparing with photon reconstruction in PandoraPFA version 1

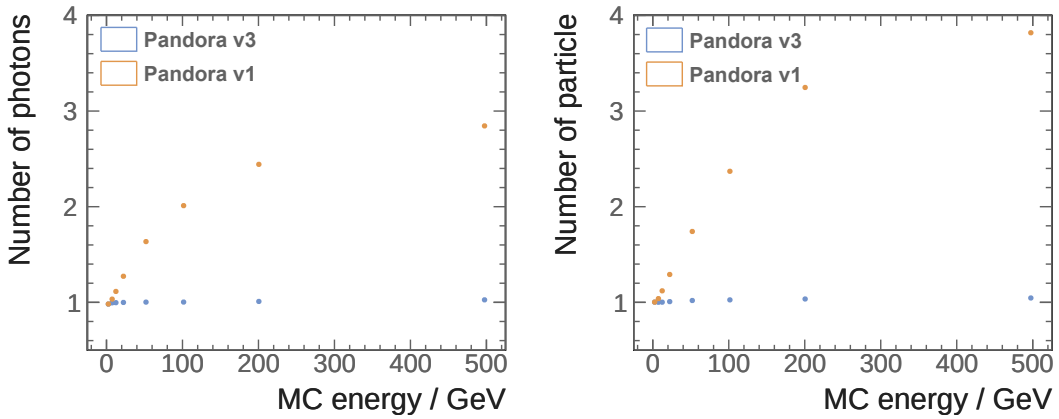
This section reviews the performance improvement with the photon algorithms from PandoraPFA version 1 to version 3, using single-photon-per-event, two-photon-per-event, and jet samples.

The single-photon-per-event samples were generated with an uniform distribution in the solid angle. Other samples are generated and smulated in the same way as in the previous section. The same selection on the single-photon-per-event and two-photon-per-event samples as in the previous section is applied.

Figure 5.12a shows the reduction in fragments reconstructed as photons, using a single-photon-per-event sample. With the reconstruction in PandoraPFA version 3, for a 100 GeV photon sample, the average number of reconstructed photons is reduced to 1 from 2; for a 500 GeV photon sample, the number is reduced to 1.05 from 2.8.

An improvement in the number of reconstructed particles is shown in figure 5.12b. The number of reconstructed particles counts the fragments reconstructed as neutral hadrons and photons. Comparing PandoraPFA version 3 with version 1, for a 100 GeV photon sample, the average number of reconstructed particles is reduced to 1 from 2.4; for a 500 GeV photon sample, 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 a two-photon-per-event sample with photon energies of 500 GeV and



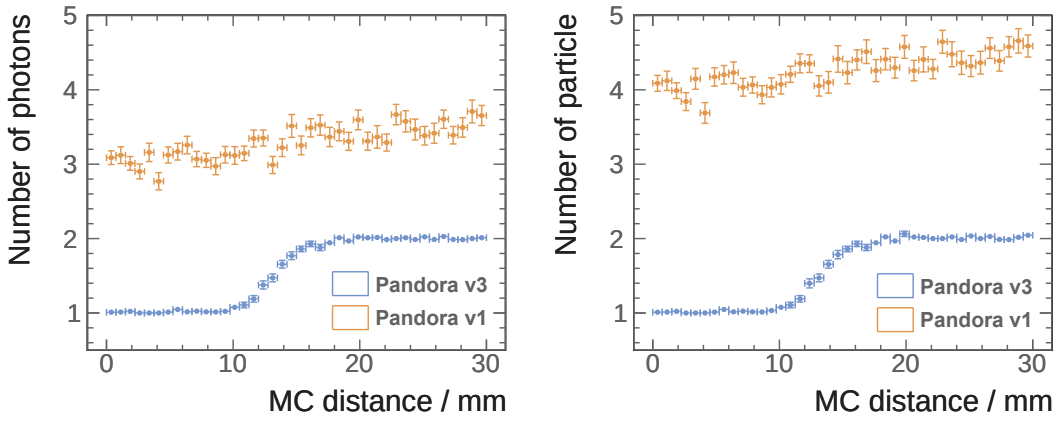
(a) (b)

Figure 5.12: Average number of reconstructed a) photons, and b) particles, as a function of their true energies using single-photon-per-event samples. 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.

50 GeV. The figures show the numbers of reconstructed photon and particles as a function of the Monte Carlo distance separation of the photon pair from 0 to 30 mm, which corresponds to approximately 6 ECAL square cell lengths of the default ILD detector model. The average numbers of photon and particle for reconstruction in PandoraPFA version 3 are both below 2.05 at 30 mm apart, which is significantly lower than the reconstruction in PandoraPFA version 1. For reconstruction with PandoraPFA version 3, two photons start to be resolved at 10 mm apart, and fully resolved at 20 mm apart.

Another metric to reflect the improvement in photon reconstruction is the fraction of the fragment energy to the total energy in a event. In a two-photon-per-event sample, the fragment energy is defined as the total energy of particles excluding the two most energetic photons. Shown in figure 5.14, using two-photon-per-event sample with photon energies of 500 GeV and 50 GeV, a reduction in fragment energy can be seen clearly going from PandoraPFA version 1 to version 3. With the photon reconstruction in PandoraPFA version 3, the average fragment energy fraction is below 0.1% up to 30 mm apart, whilst around 5% energy would be in fragments with the reconstruction in PandoraPFA version 1.

The reduction in the fragments, as shown in the reconstruction of the single-photon-per-event and two-photon-per-event samples, leads to a small improvement in the jet energy resolution at a high energy. Using the same jet sample as in the previous section,



(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-photon-per-event samples with photon energies of 500 GeV and 50 GeV. For both figures, 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.

shown in figure 5.15, the jet energy resolutions are better at 360 and 500 GeV with the photon reconstruction in PandoraPFA version 3.

5.12 Understanding photon reconstruction improvement

To show the incremental improvement of the performance of individual photon algorithm, a two-photon-per-event sample with photon energies of 500 GeV and 500 GeV is used, with different photon algorithms turned on and off. Figure 5.16 shows the average number of reconstructed particle as a function of MC distance separation between the pair, reconstructed with full photon algorithms with PandoraPFA version 3 (blue), reconstructed with only fragment removal algorithms in the ECAL and photon reconstruction in PandoraPFA version 1 (orange), reconstructed with fragment removal algorithms in the ECAL and the HCAL and photon reconstruction in PandoraPFA version 1 (green), and reconstructed with PandoraPFA version 1 (red).

For the reconstruction with fragment removal algorithm in the ECAL (orange), the number of fragment is reduced significantly, compared with photon reconstruction in

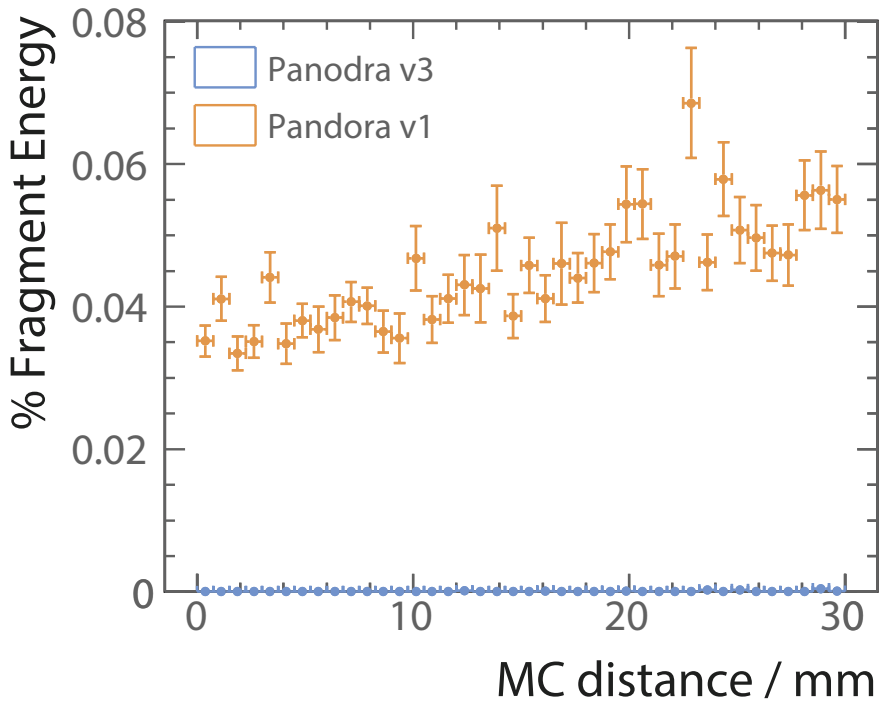


Figure 5.14: Average fraction of fragments energy to the total energy in the event, as a function of the Monte Carlo distance separation in the calorimeter, using a two-photon-per-event sample with photon energies of 500 GeV and 50 GeV. The top orange and bottom blue dots represent the reconstruction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

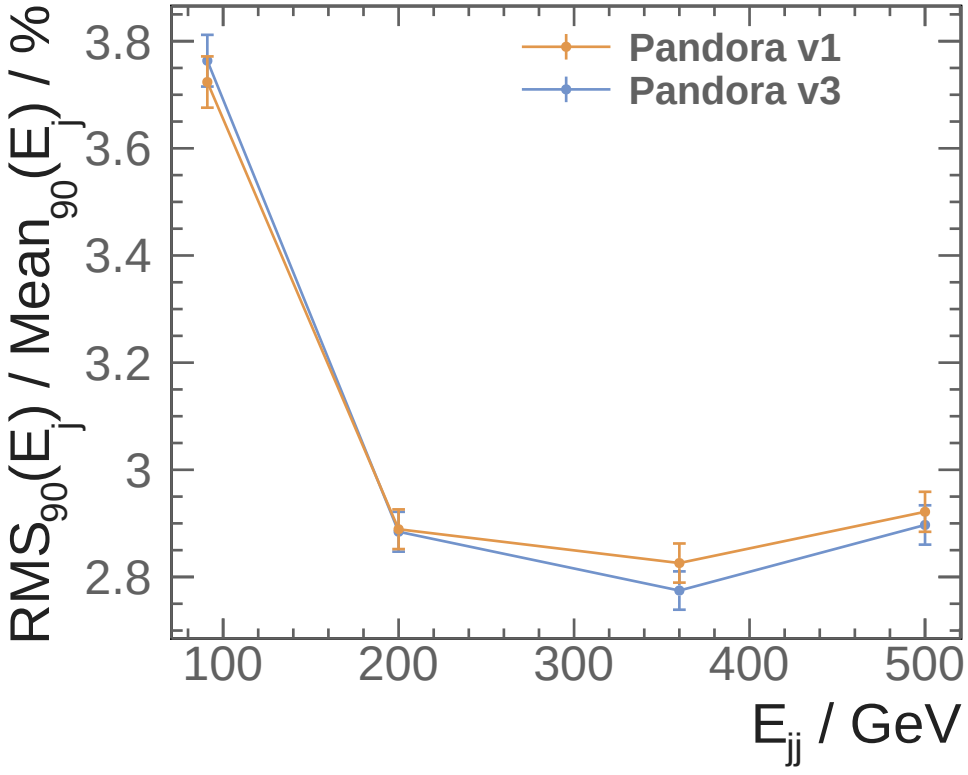


Figure 5.15: Jet energy resolutions as a function of the total jet energy using $e^+e^- \to Z'Z'$ samples, where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$, at barrel region. The top orange and bottom blue dots represent the reconstruction with PandoraPFA version 1 and version 3. The photon reconstruction is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

PandoraPFA version 1 (red). With the additional fragment removal algorithm in the HCAL (green), the number of fragments is reduced further. At 40 mm apart, for the reconstruction with fragment removal algorithms in the ECAL and the HCAL (green), there is on average less than 0.05 fragment per photon pair.

The introduction of the photon reconstruction and photon splitting algorithm (blue) resolves the photon pair at a much shorter distance separation between the pair. Photon pairs start to be resolved at 5 mm apart, and fully resolved at 15 mm apart when reconstructed with full photon algorithms.

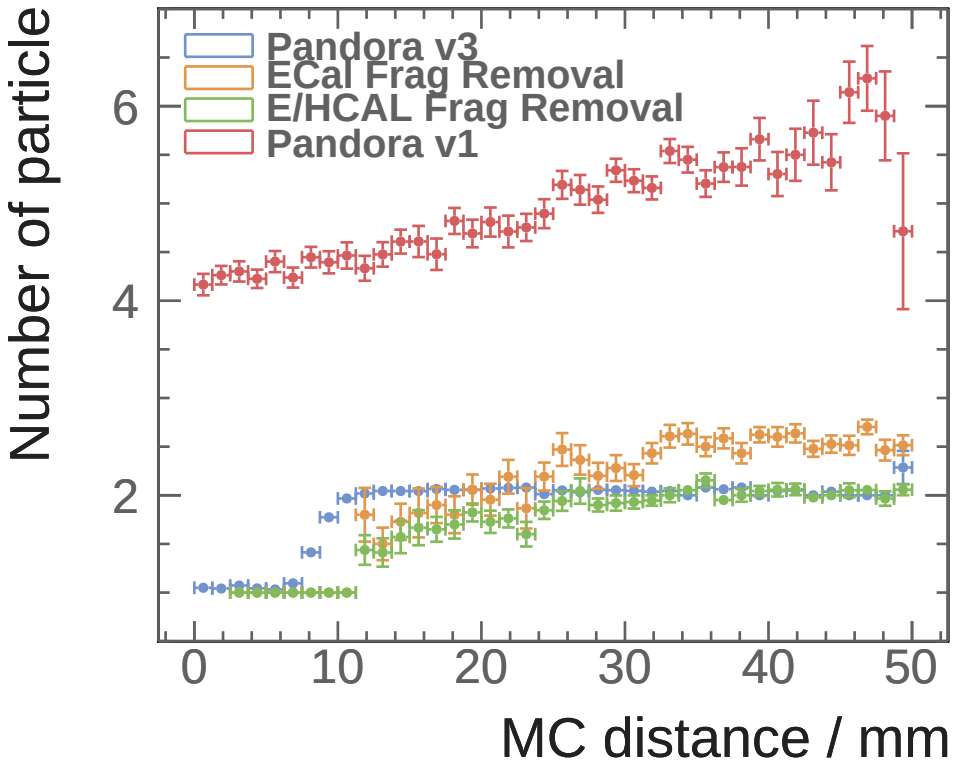
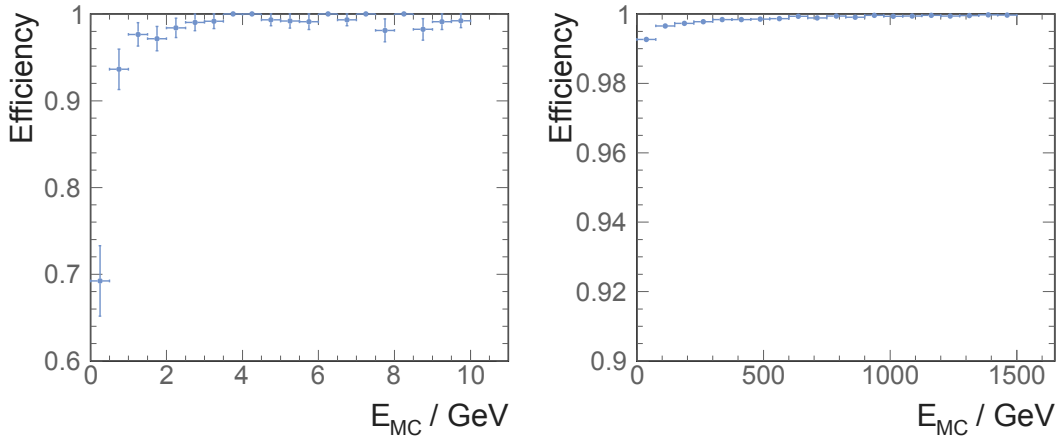


Figure 5.16: Average number of photons, as a function of the Monte Carlo distance separation between the photon pair in the calorimeter, using two-photon-per-event sample with photon energies of 500 GeV and 500 GeV. The blue, orange, green, and red dots represent the reconstruction with PandoraPFA version 3, the reconstruction with fragment removal in the ECAL and photon reconstruction in PandoraPFA version 1, the reconstruction with fragment removal in the ECAL and the HCAL and photon reconstruction in PandoraPFA version 1, the reconstruction with PandoraPFA version 1, respectively. The photon reconstruction is changed in PandoraPFA version 2.

5.13 Current photon reconstruction performance

Average single photon reconstruction efficiency is demonstrated in figure 5.17, using single-photon-per-event samples. In single-photon-per-event samples, an event can have an efficiency of 1, or 0, depending on whether there is a reconstructed photon corresponding to the MC photon. The average single photon reconstruction efficiency is above 98% for photons with energies above 2GeV, and above 99.5% for photons with energies above 100 GeV. The low efficiency in the first bin in figure 5.17a, for photon energies in the range from 0 to 0.25 GeV, is because photon reconstruction does not attempt to reconstruct photons with energies below 0.2 GeV.



(a) **(b)**

Figure 5.17: Single photon reconstruction efficiency as a function of true photon energies, using single-photon-per-event samples, for a) the low photon energy regime, and b) the high photon energy regime.

Figure 5.18 shows the average numbers of photons and particles as a function of the MC distance separation between the photon pair, using a two-photon-per-event sample with photon energies of 500 GeV and 500 GeV. The number of photons are particles are both fewer than 2.05 for a distance separation beyond 20 mm, less that 1 fragment produced per 20 events.

The ability to resolve of a photon pair depends on energies of two photons. Figure 5.19 shows the average number of photon reconstructed using two-photon-per-event samples, for different photon energies. When the energies of two photons are similar, the distance of two photons starting to be resolved is shorter. This is because that when the two photon showers have similar sizes, the 2D Peak Finding algorithm can exploit

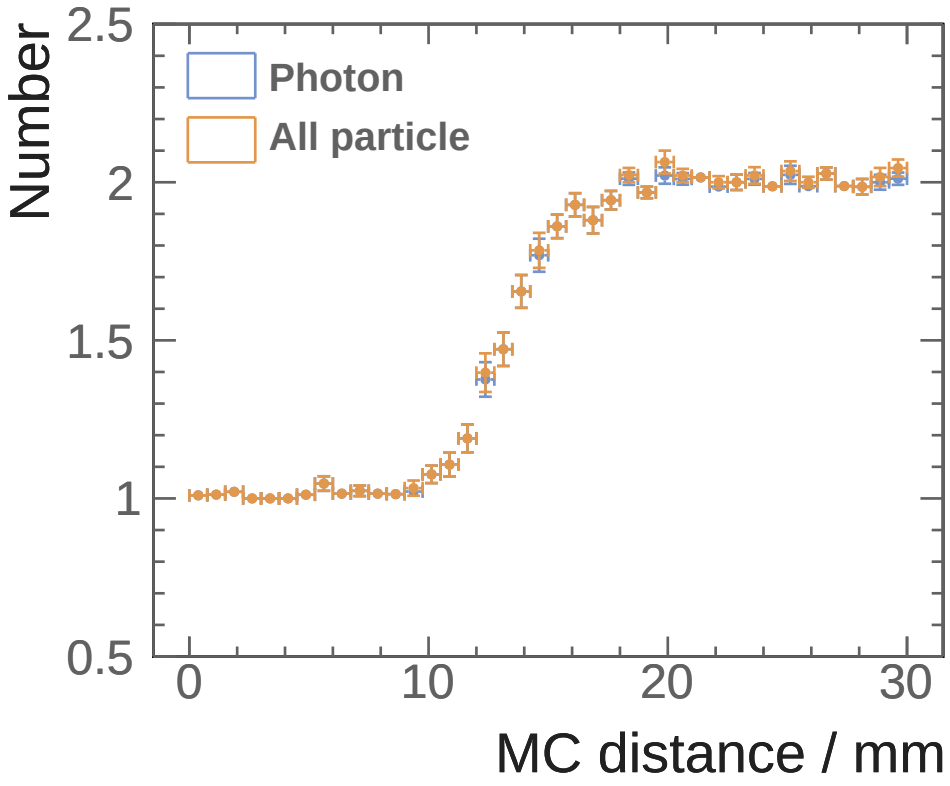


Figure 5.18: Average numbers of reconstructed photon (blue) and particle (orange), as a function of the Monte Carlo distance separation between the photon pair, using two photons of 500 GeV and 50 GeV per event sample.

the symmetry in the size of the EM showers. For example, 500 GeV−500 GeV photon pair and 10 GeV−10 GeV photon pair start to be resolved at 6 mm apart, which is about one ECAL cell length. In contrast, photon pairs with different energies, for example 500 GeV−50 GeV and 100 GeV−10 GeV pairs, start to be resolved at 10 mm apart, which is about two ECAL cells length.

For an energetic photon, it is easier to identify the photon, because the electromagnetic shower core is denser and contains more energies than the peripheral calorimeter hits. Therefore separating two energetic photons is easier than separating two low-energy photons. As shown in figure 5.19, at 20 mm apart, 500 GeV−500 GeV photon pairs are fully resolved, whereas approximately only 60% of 10 GeV−10 GeV photon pairs are resolved.

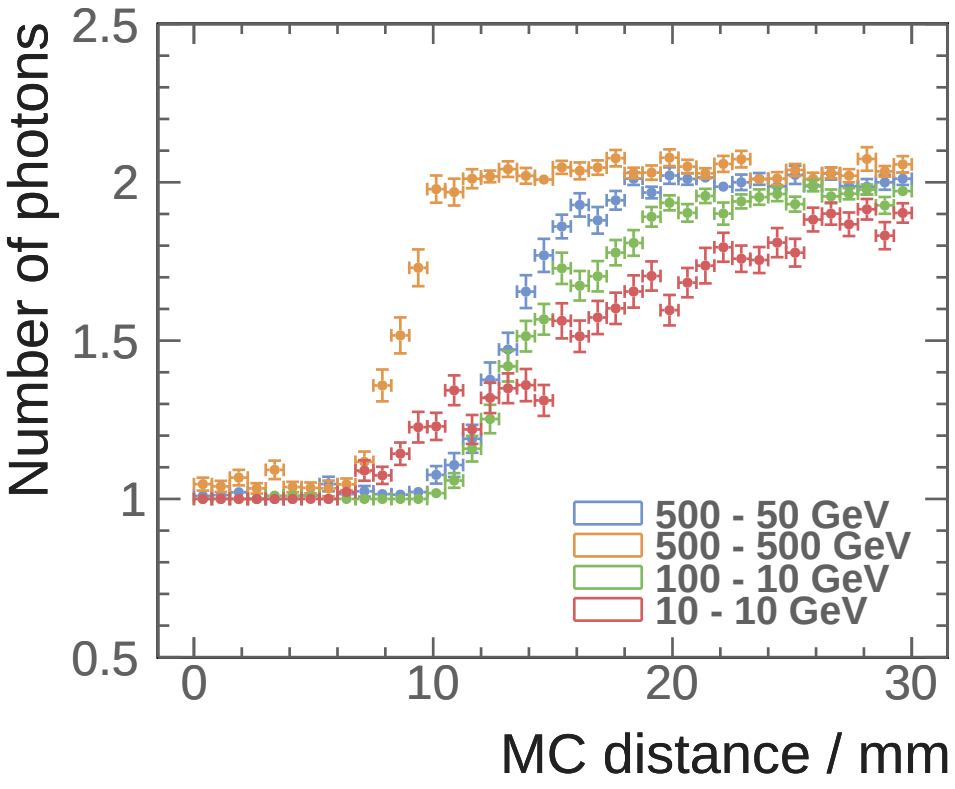


Figure 5.19: Average numbers of reconstructed photon for four different photon pairs: 500 GeV−50 GeV (blue), 500 GeV−500 GeV (orange), 100 GeV−10 GeV (green), and 10 GeV−10 GeV (red), as a function of the Monte Carlo distance separation between the photon pair.

Chapter 6

Tau Lepton Decay Modes Classification

'I once tried standing up on my toes to see far out in the distance, but I found that I could see much farther by climbing to a high place.'

— Xun Kuang, 313 BC – 238 BC

The tau pair polarisation correlation from a boson decay can be used to determine statistically if the parent boson is a scalar or a vector, and, for example, to differentiate a H boson from a Z boson [27]. It can also be used to measure the CP (the product of charge conjugation and parity symmetries) of the Higgs, via $H \to \tau^+\tau^-$ channel [88]. The CP of the Higgs can be inferred from the angle between the tau pair.

Since the tau lepton has a very short mean decay lifetime of 290 fs [26], only tau decay products can be detected with the calorimeters and tracking detectors. Therefore, the performance of the calorimetric and track systems determines the ability to reconstruct tau lepton decay products and classify different tau decay modes. The efficiency of tau decay mode classification can be used as a benchmark for detector performance.

The main challenge in the tau lepton decay modes classification is to reconstruct and to separate spatially close photons. For tau leptons with energies above tens of GeV, the visible decay products are often boosted. Electromagnetic showers from photons from π^0 decays often overlap each other in the calorimeters. Reconstructing these photons as separate entities requires good photon reconstruction. Hence the photon reconstruction algorithms described in chapter 5 are used in this study.

This chapter presents a study of the tau decay modes classification. The ability to classify tau decay modes in a highly granular linear collider detector is used for the ECAL optimisation study, where the impact of the ECAL cell sizes, as well as the tau lepton energy, is examined.

6.1 Event generation and simulation

Two million $e^+e^- \to \tau^+\tau^-$ events at a centre-of-mass energy of 100 GeV were generated with the WHIZARD software [49]. The TAUOLA software [54] was used to describe the tau lepton decay with correct spin correlations of the tau decay products. The study is focused on separating tau decay modes. Hence the beam specific effects are not generated, such as the initial state radiation and the beam induced background. The $e^+e^- \to \tau^+\tau^-$ events were simulated using the ILD detector model as described in chapter 4.

6.2 Event reconstruction

Events are reconstructed with iLCSoft version v01-17-07 [89] and PandoraPFA version 3 [65], using the photon reconstruction discussed in chapter 5. An event display of $e^+e^- \to \tau^+\tau^-$ interaction reconstructed in the ILD detector is shown in figure 6.1. The top half of the event shows a tau lepton decaying into a $\pi^-\pi^0\nu_\tau$ final state. The bottom half of the event shows a tau lepton decaying into a $\pi^+\pi^-\pi^-\pi^0\nu_\tau$ final state. Purple lines represent $\pi^\pm$ tracks of the tracking detectors. Purple dots represent calorimeter hits of the hadronic showers of $\pi^\pm$ in the ECAL and the HCAL. Yellow dots represent calorimeter hits of EM showers of photons from $\pi^0 \to \gamma\gamma$. The blue region is the transverse cross section of the ECAL barrel part.

6.2.1 Tau major decay modes

To study the main decay modes of the tau lepton, decay modes with branching ratio above 2% are classified. This results in seven tau lepton decay modes, which cover 92.58 % of the tau decay [6]. The seven tau decay modes, their branching ratios, and detectable final states are shown in table 6.1.

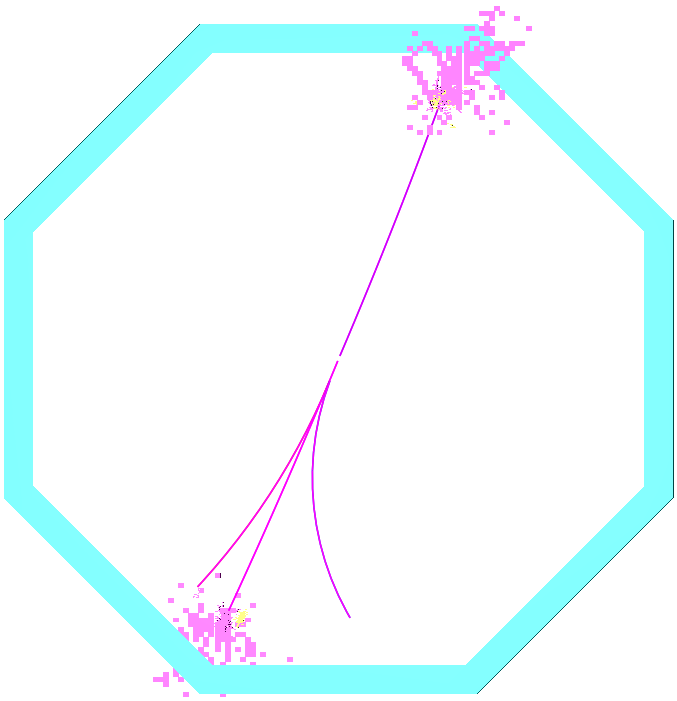


Figure 6.1: An event display of a simulated $e^+e^- \to \tau^+\tau^-$ event using the ILD detector model. The top half of the event shows a tau lepton decaying into a $\pi^-\pi^0\nu_\tau$ final state. The bottom half of the event shows a tau lepton decaying into a $\pi^+\pi^-\pi^-\pi^0\nu_\tau$ final state. Purple lines represent $\pi^\pm$ tracks of the tracking detectors. Purple dots represent calorimeter hits of the hadronic showers of $\pi^\pm$ in the ECAL and the HCAL. Yellow dots represent calorimeter hits of EM showers of photons from $\pi^0 \to \gamma\gamma$. The blue region is the transverse cross section of the ECAL barrel part.

In the $\rho\nu_\tau$ decay mode, the ρ meson subsequently decays into $\pi^-\pi^0$. In the $a_1\nu_\tau$ neutral and charged decay modes, the a_1 meson subsequently decays into $\pi^-\pi^0\pi^0$, and $\pi^+\pi^-\pi^-$, respectively. The invariant masses of the ρ meson and a_1 meson are $775.11 \pm 0.34\,\text{MeV}$ and $1230 \pm 40\,\text{MeV}$, respectively [6].

Decay mode	Detectable final state	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^-	$10.83 \pm 0.06\%$
$\rho\nu_\tau$	$\pi^-\pi^0$	$25.52 \pm 0.09\%$
$a_1\nu_\tau$ neutral	$\pi^-\pi^0\pi^0$	$9.30 \pm 0.11\%$
$a_1\nu_\tau$ charged	$\pi^+\pi^-\pi^-$	$8.99 \pm 0.06\%$
$\pi^+\pi^-\pi^-\pi^0\nu_\tau$	$\pi^+\pi^-\pi^-\pi^0$	$2.70 \pm 0.08\%$

Table 6.1: Decay modes, detectable final state particles, and branching ratios of seven major τ^- decay modes. Values are taken from [6].

6.2.2 Selecting single tau decay

The $e^+e^- \to \tau^+\tau^-$ channel contains two tau leptons. Since the tau decay mode classification is applied on a per-tau-lepton basis, the decay products of the two tau leptons in one event are divided into two sets for individual tau decay mode classifications. By identifying the axis of the back-to-back taus in the $e^+e^- \to \tau^+\tau^-$ event, the detector space can be separated in two hemispheres, where particles in each hemisphere correspond to the decay products of one tau lepton.

Separating reconstructed particles in an event into two sets is achieved using the principle thrust axis vector of the event, which is the axis that most particle are aligned to. The principle thrust axis vector, $\hat{t}$, is determined by maximising the thrust [90], 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 particle i; vector $\hat{t}$ is the unit principle thrust axis vector; and index i is summed over all particles in an event. Two sets of particles are obtained based on the sign of the scalar product between the principle thrust axis vector

and the momentum vector of a particle: particles with a positive sign of the scalar products are in one set and particles with a negative sign are in another set.

6.3 Pre-selection

Three pre-selection cuts are used based on the MC information of the particles. The cuts and the number of tau decays passing the cuts are listed in table 6.2.

Since this study is focused on the photon reconstruction in the ECAL to classify tau decay modes, the tau decay products with photon conversion to electron pairs in the tracking detector are not considered.

Another pre-selection cut requires the total visible energy (i.e. not accounting for neutrinos) of tau decay products, $E_{vis,MC}$, to be greater than 5 GeV.

Lastly, tay decay products are discarded when tau decay products deposit energies in the gap region between barrel and endcap part of the calorimeter, because there is a degradation in the particle reconstruction efficiency in the gap region. Tau decay products with the generated polar angle of the tau lepton in the region of 0.6 rad and 0.9 rad are not considered.

Table 6.2 shows the fractions of tau decays passing successive cuts for different tau decay final states. As expected, the cut on the photon conversion only affects tau decay modes with photons in the final states. The cut on the total visible energy of the tau decay products has the greatest affect on the leptonic decay modes with two neutrinos in the final states. The cut on the tau polar angle affects different tau decay modes almost equally.

6.4 Variables used in MVA

The classification of different tau decays uses a MVA classification based on twenty seven discriminant variables. These are listed in table 6.3. The particle ID information comes from the output of the PandoraPFA reconstruction.

Detectable final state	No photon conversion	$E_{vis,MC} > 5\,\text{GeV}$	$\lvert\theta_{Z,MC}\rvert$
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 6.2: Fractions of tau decays passing successive pre-selection cuts for different tau decay final states.

Category	Variable
Particle numbers	N_C, N_μ, N_e, N_γ, N_{π^-}
Invariant masses	m_{vis}, m_C, m_N, m_γ, m_{π^-}
Energy variables	$\tilde{E}_{vis}$, $\tilde{E}_C$, $\tilde{E}_\mu$, $\tilde{E}_\text{e}$, $\tilde{E}_\gamma$, $\tilde{E}_{\pi^-}$
Calorimetric energy information	E_C^{ECAL}/E_C, E^{ECAL}/E
$\rho(\pi^-\pi^0)$ reconstruction	$m_{\pi^0}^{(\rho)}$, m_ρ^{reco}
$\text{a}_1(\pi^-\pi^0\pi^0)$ reconstruction	$m_{\pi^0}^{(\text{a}_1)}$, $m_{\pi^0}^{*(\text{a}_1)}$, $m_{\text{a}_1}^{reco}$
EM shower profile	δl, t_0, $\langle w \rangle$
Calorimeter hit information	$\bar{E}_{hit}$, MIP
Track information	E/p

Table 6.3: Variables used in the MVA classification for the tau lepton decay mode classification.

6.4.1 Particle number variables

The most crucial variables for classifying tau decay modes are the number of particle of different types of particles. There are five variables used in the MVA classification: the number of charged particles (N_C); 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 reconstructed charged particles for different tau decay modes. Over 98% of $\tau^- \to \mathrm{e}^-\overline{\nu}_\mathrm{e}\nu_\tau$, $\tau^- \to \mu^-\overline{\nu}_\mu\nu_\tau$, and $\tau^- \to \pi^-\nu_\tau$ final state events have exactly one reconstructed charged particle per event, and approximately 95% of $\mathrm{a}_1(\pi^+\pi^-\pi^-)$ final state events give exactly three reconstructed charged particles. Figure 6.2b and figure 6.2c show the distributions of the number of reconstructed muons and electrons respectively for different tau decay modes. 99% of $\tau^- \to \mu^-\overline{\nu}_\mu\nu_\tau$ final state events produce exactly one reconstructed muon, and 99% of $\tau^- \to \mathrm{e}^-\overline{\nu}_\mathrm{e}\nu_\tau$ final state events have one reconstructed electron. Figure 6.2d shows the distributions of the number of reconstructed photons for different tau decay modes, which distinguishes between final states with different numbers of π^0. Nearly 75% of $\rho(\pi^-\pi^0)$ decays give exactly two reconstructed photons, and over 60% of the $\mathrm{a}_1(\pi^-\pi^0\pi^0)$ decays have exactly four photons.

6.4.2 Invariant mass variables

Five invariant mass variables are used in the MVA classification: the invariant mass of all reconstructed particles (m_{vis}); the invariant mass of all reconstructed charged particles (m_C); the invariant mass of all reconstructed neutral particles (m_N); the invariant mass of all reconstructed photons (m_γ); and the invariant mass of all charged reconstructed pions (m_{π^-}).

Figure 6.3a shows the distributions of the invariant mass of all reconstructed particles for different tau decay modes. Peaks in the invariant mass distribution can be seen for the ρ and a_1 decay modes. Figure 6.3b shows the distributions of the invariant mass of all reconstructed neutral particles for different tau decay modes. Difference in the distributions for the ρ and a_1 decay modes can be seen.

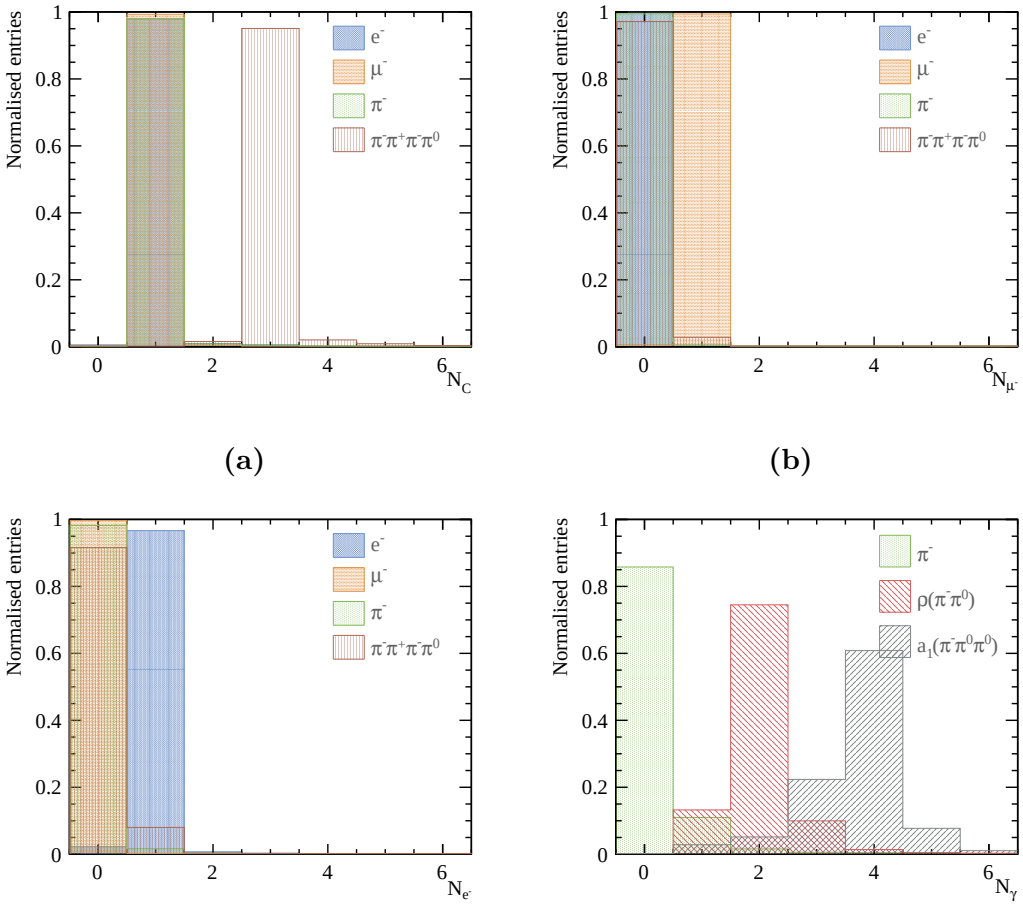
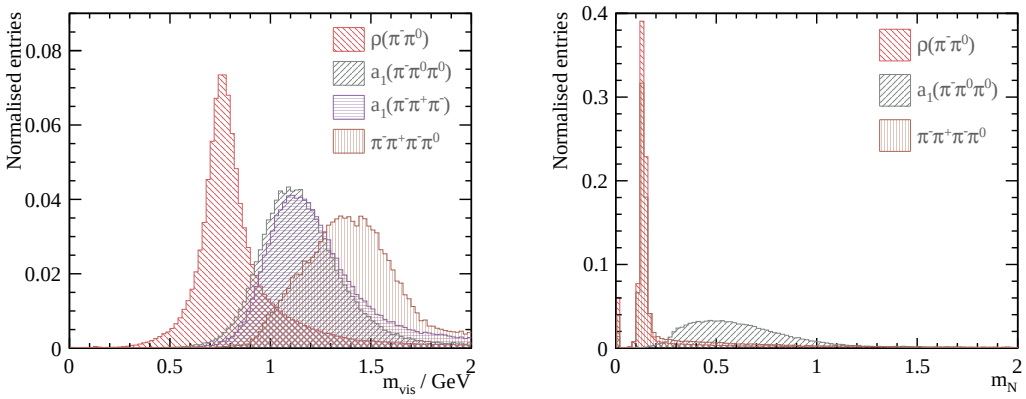


Figure 6.2: Distributions of the number of reconstructed a) charged particles (N_C); b) muons (N_μ); c) electrons (N_e); d) and photons (N_γ). The particle ID information comes from the output of the PandoraPFA reconstruction. The area under the curve for each decay mode is normalised to unity. Decay modes in all plots are selected using the truth information.



(a) **(b)**

Figure 6.3: Distributions of the invariant mass of all reconstructed a) particles (m_{vis}); b) and neutral particles (m_N). The particle ID information comes from the output of the PandoraPFA reconstruction. The area under the curve for each decay mode is normalised to unity. Decay modes in all plots are selected using the truth information.

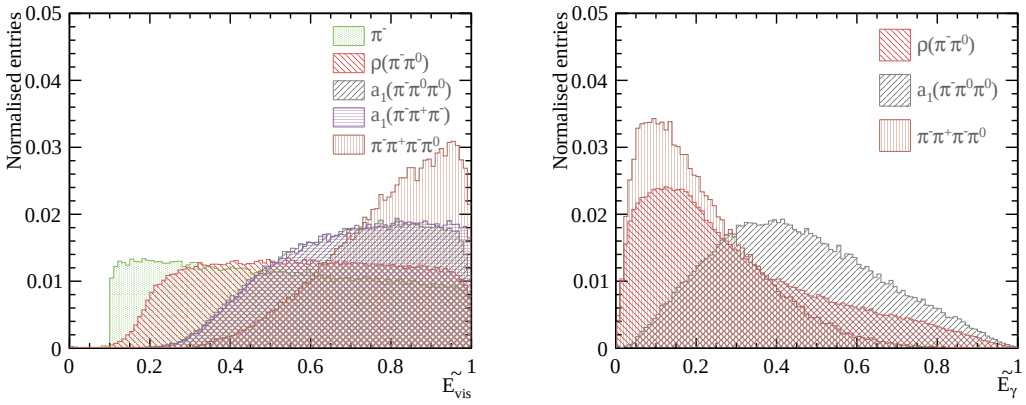
6.4.3 Energy variables

Energy information helps to further separate different tau decay modes. Six energy variables are used in the MVA classification: the normalised total energy of all reconstructed particles ($\tilde{E}_{vis}$); the normalised total energy of charged particles ($\tilde{E}_C$); the normalised total energy of muons ($\tilde{E}_\mu$); the normalised total energy of electrons ($\tilde{E}_\mathrm{e}$); the normalised total energy of photons ($\tilde{E}_\gamma$); and the normalised total energy of charged pions ($\tilde{E}_{\pi^-}$). All variables are normalised with respect to the energy of the associated tau lepton, i.e. $\tilde{E}_{vis} = E_{vis}/E_\tau$, where E_{vis} is the total energy of all reconstructed particles and E_τ is the energy of the associated tau lepton.

Figure 6.4a and figure 6.4b show the distributions of the normalised energy of all reconstructed particles and photons respectively for different tau decay modes. Difference in the distributions for different tau decay modes can be seen.

6.4.4 Calorimetric energy information variables

E_C^{ECAL}/E_C, E^{ECAL}/E



(a) **(b)**

Figure 6.4: Distributions of the normalised energy of all reconstructed a) particles ($\tilde{E}_{vis}$); b) and photons ($\tilde{E}_\gamma$). The particle ID information comes from the output of the PandoraPFA reconstruction. The area under the curve for each decay mode is normalised to unity. Decay modes in all plots are selected using the truth information.

Two calorimetric energy information variable are used in the MVA classification: the fraction of the energy deposited in the ECAL divided by the energy deposited in the ECAL and HCAL, where only calorimetric deposits associated with charged particles are considered (E_C^{ECAL}/E_C), and the fraction of the energy deposited in the ECAL divided by the energy deposited in the ECAL and HCAL for all particles (E^{ECAL}/E). These two variables help to identify electron and muon decay modes. For example, an electron typically deposits over 95% of its energy in the ECAL, and a muon typically deposits 5% to 20% of its energy in the ECAL. The difference between $\%E_C$ and $\%E$ is that photons and neutral hadrons, which deposit most of their energy in the ECAL, are not included in the calculation of E_C^{ECAL}/E_C.

6.4.5 $\rho(\pi^-\pi^0)$ and $a_1(\pi^-\pi^0\pi^0)$ resonances variables

By utilising the photon identification potential of the highly granular ECAL, the identification of the $\rho(\pi^-\pi^0)$ and $a_1(\pi^-\pi^0\pi^0)$ decay modes is enhanced by reconstructing the ρ and a_1 invariant mass. For events with at least one charged pion and one photon, the reconstruction selects the combination of charged pion and photon that have a invariant mass consistent with the ρ or a_1 mass.

For example, the final state of the $\rho(\pi^-\pi^0)$ decay mode contains a π^- and a π^0, where $\pi^0 \to \gamma\gamma$. The $\rho(\pi^-\pi^0)$ decay mode hypothesis test is performed by selecting the combination of the charged pion and photons that gives the smallest value of a χ^2 function:

$$\chi^2 = \left(\frac{m_{tot} - m_\rho}{\sigma_\rho}\right)^2 + \left(\frac{m_{\gamma_1\gamma_2} - m_{\pi^0}}{\sigma_{\pi^0}}\right)^2, \tag{6.2}$$

where $m_{\gamma_1\gamma_2}$ is the invariant mass of two photons; the variable m_{tot} is the total invariant mass of the two photons and one π^-; the variables m_ρ and m_{π^0} are the respective true masses of ρ and π^0, taken from [6]; and the mass resolution is assumed to be 20%, i.e. $\sigma_\rho/m_\rho = \sigma_{\pi^0}/m_{\pi^0} = 20\%$. Figure 6.5 shows the reconstructed invariant mass distributions for π^0 and ρ in the $\rho(\pi^-\pi^0)$ decay mode. The reconstructed masses are obtained using the MC truth information to find the corresponding reconstructed particles. The decay mode is selected using the MC truth information. A mass resolution of 20% is an good approximation for the invariant masses of π^0 and ρ.

The particle ID of charged pions and photons come from the output of the PandoraPFA reconstruction. The χ^2 function works naturally if there are two photons reconstructed in an event. If there are more than two photons in an event, combinations of two photons are iterated and the combination with the smallest value of χ^2 is chosen. If there is only one photon in an event, the second term in the equation 6.2 is dropped and the χ^2 function becomes:

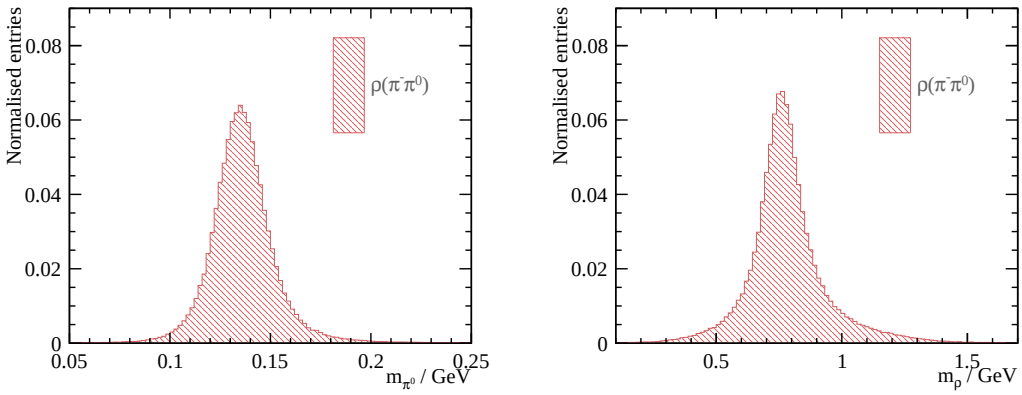
$$\chi^2 = \left(\frac{m_{tot} - m_\rho}{\sigma_\rho}\right)^2, \tag{6.3}$$

where m_{tot} is the total invariant mass of one photon and one π^-.

The χ^2 function in equation 6.2 is modified for the $a_1(\pi^-\pi^0\pi^0)$ decay mode hypothesis test:

$$\chi^2 = \left(\frac{m_{tot} - m_{a_1}}{\sigma_{a_1}}\right)^2 + \left(\frac{m_{\gamma_1\gamma_2} - m_{\pi^0}}{\sigma_{\pi^0}}\right)^2 + \left(\frac{m_{\gamma_3\gamma_4} - m_{\pi^0}}{\sigma_{\pi^0}}\right)^2, \tag{6.4}$$

where the ρ mass has been replace by the a_1 mass and other variables are defined in the same way as previously. Four photons and one π^- are required for this minimisation. To resolve the degeneracy between two photon pairs, the requirement of $\left|m_{\gamma_1\gamma_2} - m_{\pi^0}\right| < \left|m_{\gamma_3\gamma_4} - m_{\pi^0}\right|$ is imposed.



(a) **(b)**

Figure 6.5: Reconstructed invariant mass distributions for a) π^0; b) ρ, in the $\rho(\pi^-\pi^0)$ decay mode. The reconstructed masses are obtained using the MC truth information to find the corresponding reconstructed particles. The decay mode is selected using the MC truth information. The area under the curve is normalised to unity.

The particle ID of charged pions and photons come from the output of the PandoraPFA reconstruction. If there are at least four photons in an event, combinations of photons are iterated and the combination with the smallest value of χ^2 is chosen. If there are two (three) photons in an event, the last term in the equation 6.4 is dropped and the χ^2 function becomes:

$$\chi^2 = \left(\frac{m_{tot} - m_{\mathrm{a}_1}}{\sigma_{\mathrm{a}_1}}\right)^2 + \left(\frac{m_{\gamma_1\gamma_2} - m_{\pi^0}}{\sigma_{\pi^0}}\right)^2, \tag{6.5}$$

where m_{tot} is the invariant mass of the charged pion and two (three) photons. If there is only one photon in an event, the χ^2 function becomes:

$$\chi^2 = \left(\frac{m_{tot} - m_{\mathrm{a}_1}}{\sigma_{\mathrm{a}_1}}\right)^2, \tag{6.6}$$

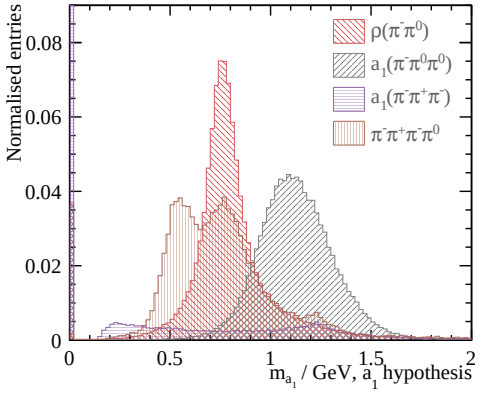
where m_{tot} is the invariant mass of the charged pion and the photon.

For the ρ invariant mass reconstruction of the $\rho(\pi^-\pi^0)$ decay mode, two variables are used in the MVA classification to help to identify $\rho(\pi^-\pi^0)$ decay mode: the ρ mass ($m_\rho^{reco} \equiv m_{tot}$ in equation 6.2) and the π^0 mass ($m_{\pi^0}^{(\rho)} \equiv m_{\gamma_1\gamma_2}$ in equation 6.2).

For the $\mathrm{a}_1(\pi^-\pi^0\pi^0)$ decay mode identification using the a_1 invariant mass resonance reconstruction, three variables are used in the MVA classification: the a_1 mass ($m_{\mathrm{a}_1}^{reco} \equiv$

m_{tot} in equation 6.4), the first π^0 mass ($m_{\pi^0}^{(a_1)} \equiv m_{\gamma_1\gamma_2}$ in equation 6.4), and the second π^0 mass ($m_{\pi^0}^{*(a_1)} \equiv m_{\gamma_3\gamma_4}$ in equation 6.4).

Figure 6.6a shows the distributions of $m_{a_1}^{reco}$ under $a_1(\pi^-\pi^0\pi^0)$ decay mode hypothesis test for four different tau decay modes. Only the distribution for $a_1(\pi^-\pi^0\pi^0)$ decay mode has a resonance peak at a_1 mass position.



(a)

Figure 6.6: Reconstructed invariant mass distributions for a_1 ($m_{a_1}^{reco}$), reconstructed under the $a_1(\pi^-\pi^0\pi^0)$ decay mode hypothesis. The reconstructed masses are obtained using the MC truth information to find the corresponding reconstructed particles. The decay mode is selected using the MC truth information. The area under the curve is normalised to unity.

6.4.6 Separating electrons from charged pions

Variables are used in this analysis to help further separating electrons from charged pions, obtained from a modified private version of PandoraPFA.

An electron develops a characteristic EM shower in the ECAL, whilst a charged pion develops a hadronic shower. Variables characterising the EM shower help to identify an electron. 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 EM shower profile (δl); and $\langle w \rangle$, a measure of the EM shower transverse width. These variables are defined in the same way as the variables used in the photon likelihood classifier in the photon reconstruction in PandoraPFA, described in section 5.5.

The calorimeter hit information can also be used to differentiate an EM shower from a hadronic shower. Two variables used in the MVA classification are: the average energy of a ECAL calorimeter hit ($\bar{E}_{hit}$), which is the total energy deposited in the ECAL and HCAL divided by the number of the ECAL and HCAL calorimeter hit, and the average fraction of minimum ionising calorimeter hits (MIP), which is the number of calorimeter hits in the ECAL and HCAL flagged as minimum ionising particles by the PandoraPFA reconstruction divided by the total number of calorimeter hits in the ECAL and HCAL.

One variable used is the consistency of the track momentum with the total ECAL and HCAL energy. The variable used in the MVA classification is the energy in the ECAL and HCAL divided by the track momentum , averaged over all particles (E/p).

6.5 Multivariate Analysis

The MULTICLASS class of the TMVA package [91] was used to perform a multiple-class classification, which classifies seven tau lepton decay final states simultaneously. The MULTICLASS classification is an extension of a standard two-class signal-background classification. The Boosted Decision Tree classifier with Gradient boost (BDTG) is used. Half of the samples, randomly selected, were used in the training process and the other half were used for testing. The optimisation of the BDTG classifier followed the strategy outlined in section 4.5.1. The optimised parameters for the classifier are listed in table 6.4, where an explanation of the parameters can be found in section 4.5.6.1.

6.6 Tau decay mode classification efficiency

Two million $e^+e^- \to \tau^+\tau^-$ events at a centre-of-mass energy of 100 GeV were used in the classification. For tau decays passing pre-selection cuts, the correct classification and misidentification efficiencies for the seven tau decay modes are shown in table 6.5. The correct classification efficiencies, in bold numbers, are defined as:

$$\varepsilon_i = \frac{N_i^{correct}}{N_i^{MC}}, \tag{6.7}$$

where $N_i^{correct}$ is the number of correctly classified tau decays for tau decay mode i and the N_i^{MC} is the total number of true tau decays of tau decay mode i.

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 6.4: Optimised parameters for the Boosted Decision Tree with Gradient boost MULTICLASS classifier. A detailed explanation of variables can be found in section 4.5.6.1.

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 6.5: Classification efficiencies for the seven tau decay modes considered here. Bold numbers represent the correct classification efficiencies. The boxes highlight one-prong and three-prong tau hadronic decay modes. The entries marked with "-" represent numbers below 0.25%. The absolute statistical uncertainty for each entry is less than 0.25%.

The particle ID from the PandoraPFA reconstruction is effective, resulting in the classification efficiencies for $\tau^- \to e^- \overline{\nu}_e \nu_\tau$ and $\tau^- \to \mu^- \overline{\nu}_\mu \nu_\tau$ decays being 99.8% and 99.5% respectively.

For the $\tau^- \to \pi^- \nu_\tau$ decays, only 0.9% of decays are misclassified as $\tau^- \to e^- \overline{\nu}_e \nu_\tau$ decays, due to variables dedicated to separation between e^- and π^-.

For the separation of tau hadronic decay modes, the photon reconstruction is important as the number of photons is an essential variable to distinguish different hadronic decay modes. Failure to reconstruct photons in the $\tau^- \to a_1(\pi^-\pi^0\pi^0)\nu_\tau$ decays or extra

reconstructed photons in the $\tau^- \to \rho(\pi^-\pi^0)\nu_\tau$ decays leads to the misclassification between the two decays. Similarly, failure to reconstruct photons in the $\tau^- \to \rho(\pi^-\pi^0)\nu_\tau$ decays or extra reconstructed photons in the $\tau^- \to \pi^-\nu_\tau$ decays leads to the misclassification between the two decays. The misclassification between one-prong decays is highlighted in table 6.5.

6.7 Electromagnetic calorimeter optimisation

Photon reconstruction in a highly granular ECAL is an important metric for the ECAL performance. Since the classification of the tau hadronic decay modes depends on the ability to reconstruct photons, the tau hadronic decay modes classification is used as a metric to optimise the ECAL design. The tau decay mode classification was studied with ECAL square cell sizes of 3, 5, 7, 10, 15 and 20 mm, and at four centre-of-mass energies of 100, 200, 500, 1000 GeV. The other ECAL dimensions are kept the same as for the ILD nominal detector. The multivariate classifier was trained individually for each ECAL cell size and each centre-of-mass energy.

PandoraPFA was optimised for the nominal ILD detector. Therefore a re-optimisation is required for detector models with different ECAL cell sizes. In particular, the parameters used in PhotonFragmentRemoval algorithm need to be optimised for different ECAL cell sizes. The optimal ClosestHitDistance parameter in PhotonFragmentRemoval algorithm, which is a distance metric controlling the merging of the fragment, was chosen by selecting the value that gives the highest overall tau hadronic decay classification rate, ε_{had}, amongst values of 5, 10, 20, 30, 40, and 50 mm with $e^+e^- \to \tau^+\tau^-$ samples at a centre-of-mass energy of 100 GeV. The overall tau hadronic decay classification rate, ε_{had}, is the weighted average correct classification efficiency:

$$\varepsilon_{had} = \frac{\sum_i^5 B_i \varepsilon_i}{\sum_i^5 B_i}, \tag{6.8}$$

where B_i is the branching fraction of the tau hadronic decay mode i; ε_i is the correct classification efficiency of tau decay mode i (defined in equation 6.7); and the index i is summed over five tau hadronic decay modes considered here:$\tau^- \to \pi^-\nu_\tau$; $\tau^- \to \rho(\pi^-\pi^0)\nu_\tau$; $\tau^- \to a_1(\pi^-\pi^0\pi^0)\nu_\tau$; $\tau^- \to a_1(\pi^+\pi^-\pi^-)\nu_\tau$; and $\tau^- \to \pi^+\pi^-\pi^-\pi^0\nu_\tau$.

Table 6.6 shows the optimised values of ClosestHitDistance parameter in PhotonFragmentRemoval algorithm as a function of the ECAL square cell size. As expected, for larger cell sizes, the distance metric for merging photons becomes larger.

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 6.6: Optimised values of ClosestHitDistance parameter in PhotonFragmentRemoval algorithm as a function of the ECAL square cell size.

Figure 6.7 shows ε_{had} as a function of ECAL cell size for four different centre-of-mass energies. ε_{had} decreases with the increasing centre-of-mass energies and increasing ECAL cell sizes, because it is increasingly difficult to reconstruct boosted photons with lower ECAL transverse spatial resolutions.

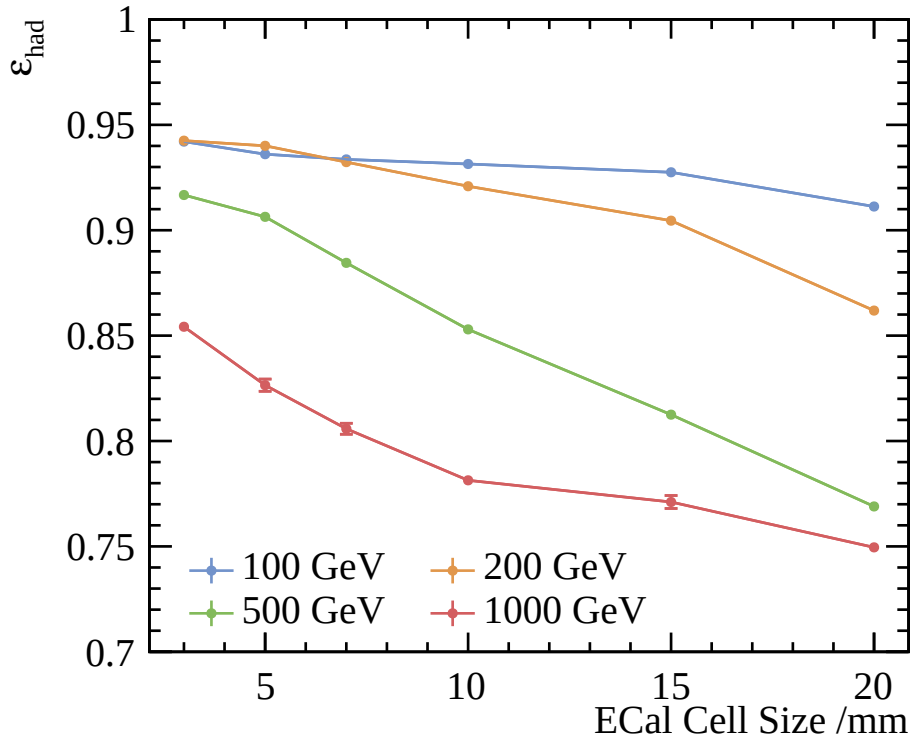


Figure 6.7: The weighted average tau hadronic decay efficiency, ε_{had}, as a function of the ECAL cell sizes for different centre-of-mass energies. The blue, orange, green, and red points show ε_{had} at $\sqrt{s}$ = 100, 200, 500 and 1000 GeV, respectively.

The tau decay mode correct classification efficiencies generally decrease with an increase of centre-of-mass energy. As the centre-of-mass energy increases, the tau decay products become more boosted, making it increasingly difficult to separate tau decay products, for example, the photon pair from π^0 decay. The reduction in the ability to separate photon pairs leads to a degradation of the classification performance.

The tau decay mode correct classification efficiencies decrease with an increase of the ECAL cell size. 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, leading to a reduction in the ability to separate a pair of photons. Consequently, a worse classification performance is expected for larger ECAL cell size.

table 6.7 lists the ε_{had} with a 3 mm and 20 mm ECAL cell sizes for four different centre-of-mass energies. The sensitivity of ε_{had} to different cell sizes is stronger 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 small ECAL cell size to reconstruct individual particles.

ε_{had}	3 mm	20 mm
100 GeV	94%	91%
200 GeV	94%	86%
500 GeV	92%	78%
1000 GeV	85%	75%

Table 6.7: ε_{had} with a 3 mm and 20 mm ECAL cell sizes for four different centre-of-mass energies.

Figure 6.8 shows 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. The tau decay mode correct classification efficiencies generally decrease with an increase of centre-of-mass energies and an increase of ECAL cell sizes.

For the $\tau^- \to \rho(\pi^-\pi^0)\nu_\tau$ decay mode, the efficiency for $\sqrt{s} = 1000\,\text{GeV}$ increases as the cell size increases. This is because the multivariate classifier optimises for the overall classification efficiency, which may balance 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 the efficiency for $\tau^- \to \rho(\pi^-\pi^0)\nu_\tau$ decay mode at $\sqrt{s} = 1000\,\text{GeV}$ is

compensated by the drastic decrease in the efficiency for $\tau^- \to a_1(\pi^-\pi^0\pi^0)\nu_\tau$ decay mode at the same centre-of-mass energy.

For the $\tau^- \to a_1(\pi^-\pi^0\pi^0)\nu_\tau$ decay mode, the loss of efficiency with an increasing ECAL cell size and an increasing centre-of-mass energy is most significant comparing to other decay modes. With most number of photons 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 $\tau^- \to a_1(\pi^+\pi^-\pi^-)\nu_\tau$ decay mode, the efficiencies are similar to that of the $\tau^- \to \pi^-\nu_\tau$ decay mode. Both final states contain charged particles only. Therefore it is most sensitive to the tracking performance, which is not affected by different ECAL cell sizes.

6.8 Tau pair polarisation correlations as a signature of Higgs boson

A spin-0 scalar Higgs boson can decay to $\tau_L^+\tau_L^-$ or $\tau_R^+\tau_R^-$, whereas a spin-1 vector boson Z with can decay to $\tau_L^+\tau_R^-$ or $\tau_R^+\tau_L^-$, where L, R denotes the tau lepton helicity, due to the helicity conservation. Therefore, by studying the tau pair polarisation correlation from a boson decay, one can determine statistically if the parent boson is a scalar or a vector.

This section follows the theoretical discussion in section 2.9 on using the correlation between the polarisations of the tau pair from a boson decay as a signature to differentiate the Higgs boson from the Z boson. A proof-of-principle analysis is performed 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 generation and simulation, 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 tau pair polarisation correlation obtained with generator-level Monte Carlo particles.

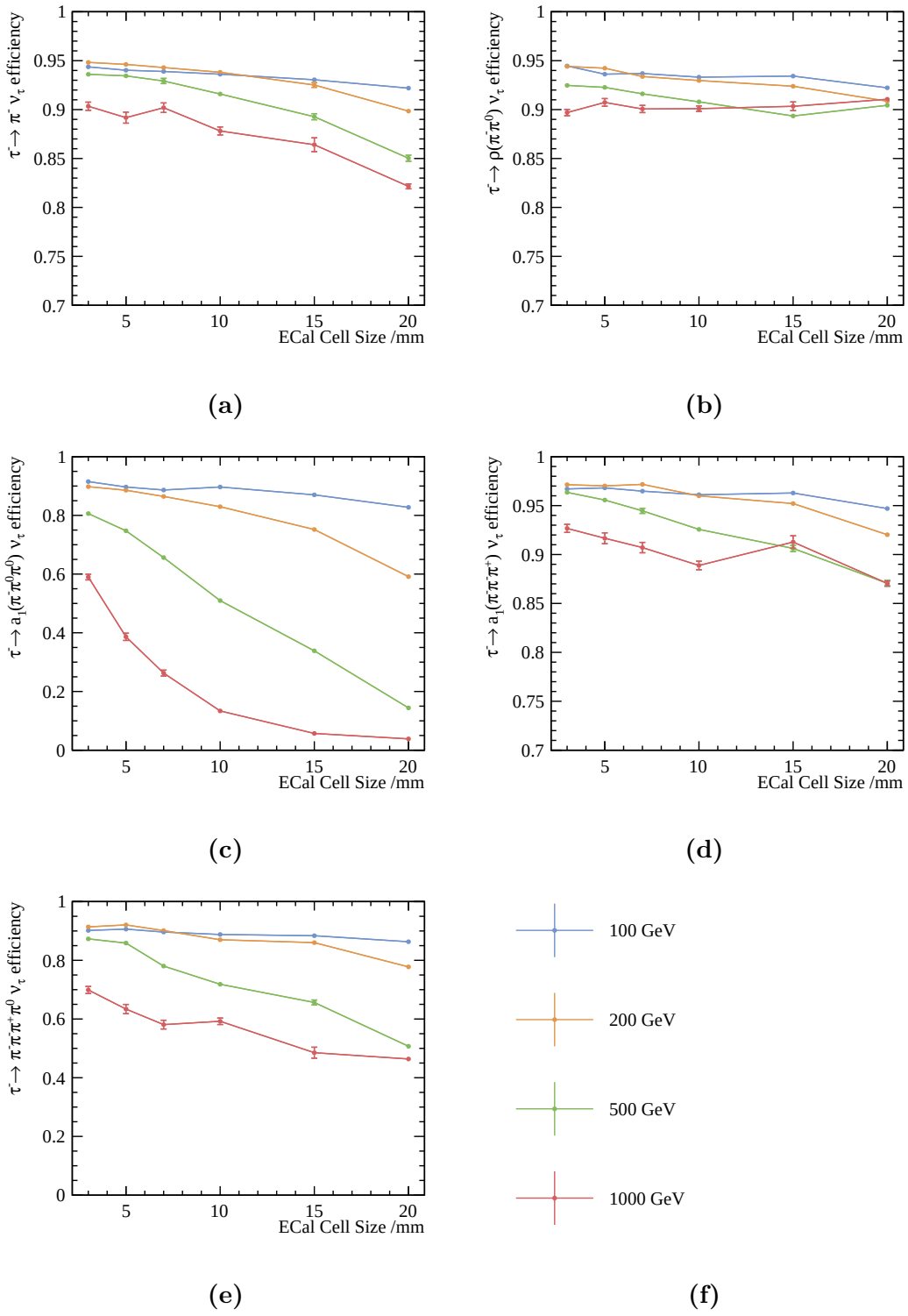


Figure 6.8: The correct classification efficiencies as a function of the ECAL square cell sizes for a) $\tau^- \to \pi^- \nu_\tau$ decay mode; b) $\tau^- \to \rho(\pi^-\pi^0)\nu_\tau$ decay mode; c) $\tau^- \to a_1(\pi^-\pi^0\pi^0)\nu_\tau$ decay mode; d) $\tau^- \to a_1(\pi^+\pi^-\pi^-)\nu_\tau$ decay mode; and e) $\tau^- \to \pi^+\pi^-\pi^-\pi^0\nu_\tau$ decay mode. The legend is shown in f). All plots are produced with the ILD detector model using $e^+e^- \to \tau^+\tau^-$ channel at $\sqrt{s}$ = 100, 200, 500, and 1000 GeV.

6.8.1 Event generation and simulation

The studied process is $e^+e^- \to ZZ$, where one Z boson decays hadronically and the other Z boson decays to a tau lepton pair. The samples were generated at a centre-of-mass energy of 350 GeV without ISR contribution for this proof-of-principle study.

The same seven tau decay modes defined in the previous analysis in section 6.1 are studied. The constraints on the simulated events are the same as those in the previous analysis in section **??**, except that the constraint on the total energy of non-neutrino decay products is not applied. A considerable fraction of $Z \to \tau^+\tau^-$ events, where $\tau^- \to \pi^-\nu_\tau$, have two low-energy charged pions. Therefore, the constraint on the total energy of non-neutrino decay products would distort the energy distribution of the charged pions, thus affecting the tau pair polarisation correlation extration.

6.8.2 Find tau decay products

The final state of the process, $e^+e^- \to ZZ \to \tau^+\tau^-qq$, contains two tau leptons and two quark jets. Therefore, tau decay products can either be found by direct tau lepton decay products searching with a tau finder processor, or by using jet algorithms to find tau decay products as jets. When a tau lepton decays to a few particles, such as the leptonic tau decay final states, the direct tau searching method is favoured. When a tau lepton decays hadronically, resulting in multiple charged and neutral particles, the jet cluster method is suitable to find tau decay methods. Hence, two approaches are combined to find the two tau leptons decay products.

6.8.2.1 Direct tau searching

Tau finder processer, IsolatedTauIdentifer, is a modified version of the one used in the double Higgs analysis in section 7.3.2.2. The idea is to find tau decay products consistent with tau decay topologies, and to require the tau decay products to be isolated from the rest of the particles.

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 (m_c) less than 3 GeV. Particles

are iteratively added to the search cone with a gradually widening opening angle of the search cone. 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 charged particle ($\theta_{c-2^{nd}X^+}$) is larger than 0.6 rad. If the criteria is satisfied, particles associated with the search cone is identified are the decay products of one tau lepton.

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

Table 6.8: Optimised parameters for the modified IsolatedTauIdentifer.

6.8.2.2 Jet clustering

Tau hadronically decay products can be also identified as a small jet. The Durham algorithm (see section 4.4.2.2) was used to form jets. The jet algorithm runs in the exclusive mode to find four jets for $\mathrm{e^+e^-} \to \mathrm{ZZ} \to \tau^+\tau^-\mathrm{qq}$ events.

6.8.2.3 Selecting best tau candidates for each tau finding method

The direct tau searching method may find more than two tau candidates, where each tau candidate corresponds to one tau lepton decay products. To identified the best two tau candidates, kinematic constraints are used. For example, in $\mathrm{e^+e^-} \to \mathrm{ZZ}$ events, the energy of the Z boson is half of the centre-of-mass energy. The invariant mass of two quarks from Z should be close to Z mass. Therefore, the χ^2 minimisation function utilising kinematic constraints is:

$$\chi^2 = \frac{\left(m_{\mathrm{qq}} - m_{\mathrm{Z}}\right)^2}{\sigma^2_{m_{\mathrm{qq}}}} + \frac{\left(E_{\mathrm{qq}} - \frac{\sqrt{s}}{2}\right)^2}{\sigma^2_{E_{\mathrm{qq}}}}, \tag{6.9}$$

where $\sqrt{s}$ is the centre-of-mass energy; the variable m_{Z} is the mass of Z boson from reference [6]; the variables $\sigma_{m_{\mathrm{qq}}}$ and $\sigma_{E_{\mathrm{qq}}}$ are the reconstructed mass resolution and energy resolution of the $\mathrm{Z} \to \mathrm{qq}$, respectively; the variables m_{qq} and E_{qq} are obtained

from the recoil momenta against two tau candidates, assuming that the collision occurs at $\sqrt{s}$. The m_{qq} is defined as the invariant mass of the recoil momenta, and the E_{qq} is defined as the energy of the recoil momenta.; and the minimisation is iterated over all tau candidates.

This minimisation produces two best tau candidates for the direct tau searching method. For the jet clustering method, to find the two jets corresponding to two tau lepton decay, the same χ^2 minimisation function is used. The variables m_{qq} and E_{qq} are defined as the total invariant mass and energy of two non-tau-candidate jets, respectively. All other variables in the minimisation function are defined in the same way. After applying the minimisation function, two jets will be identified as the best two tau candidates from the jet clustering method.

6.8.2.4 Selecting best tau candidates in an event

Having identified the best two tau candidates for each method, a set of conditions is used to determine the best overall pair of tau candidates in an event. If the best pairs of tau candidates from both methods satisfy the kinematic constraint:

$$\left|m_{\mathrm{qq}} - m_{\mathrm{Z}}\right| < \sigma_{m_{\mathrm{qq}}}, \quad \left|E_{\mathrm{qq}} - \frac{\sqrt{s}}{2}\right| < \sigma_{E_{\mathrm{qq}}}, \tag{6.10}$$

the pair of tau candidates with smallest χ^2 is selected as the best pair of tau candidates. Otherwise, if only one pair of tau candidates satisfies the constraint in equation 6.10, that pair is chosen. If none of the pairs satisfies the constraint, and if one jet from the jet clustering is close to the beam pipe and there are exactly two tau candidates obtained from IsolatedTauIdentifer, then these two tau candidates from IsolatedTauIdentifer are chosen. This is because if one jet is close to the beam pipe, it is likely that some particles close to the beam pipe are undetected, which leads to a failure in the kinematic constraint and/or the jet reconstruction. Lastly, if all conditions above are not satisfied, two smallest jets by the number of PFOs from the jet clustering method are chosen to be the best pair of tau candidates.

6.8.3 Boosting tau decay products

To use the tau decay mode classifier, it is necessary to know the tau lepton energy to calculate the variables used in the classifier. For the channel $\mathrm{Z} \to \tau^+\tau^-$, the energy of the

tau lepton can be obtained in $Z \to \tau^+\tau^-$ decay rest frame, which is half of the Z boson energy in the $Z \to \tau^+\tau^-$ rest frame. Hence the tau decay products need to be boosted to the $Z \to \tau^+\tau^-$ decay rest frame for the calculation of the variables used in the MVA classification.

The boosting of the tau decay product requires the four-momentum of the Z boson, where $Z \to \tau^+\tau^-$, which is calculated from the recoil momenta of non tau-decay-products:

$$p^{\mu}_{\tau\tau} = \begin{pmatrix} \sqrt{s} \\ \sqrt{s} \times \sin(\theta_{beam}) \\ 0 \\ 0 \end{pmatrix} - \sum_{i}^{non-\tau} p^{\mu}_{i}, \tag{6.11}$$

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 Z boson to be a half of $\sqrt{s}$:

$$p^{\mu}_{\tau\tau,correct} = p^{\mu}_{\tau\tau} \times \frac{\frac{1}{2}\sqrt{s}}{E_{\tau\tau}}, \tag{6.12}$$

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. The vector $p^{\mu}_{\tau\tau,correct}$ is then treated as the four-momentum vector of Z boson, 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 $Z \to \tau^+\tau^-$ decay rest frame.

6.8.4 Variables used in the MVA

Variables used in the MVA classifier are a subset of the ones listed in table 6.3. Variables regarding EM shower profiles, calorimeter hit information, and track information are not used (the last three rows in table 6.3) as the information was not available in the outputs of the standard version of PandoraPFA used in this analysis.

6.8.5 Multivariate analysis

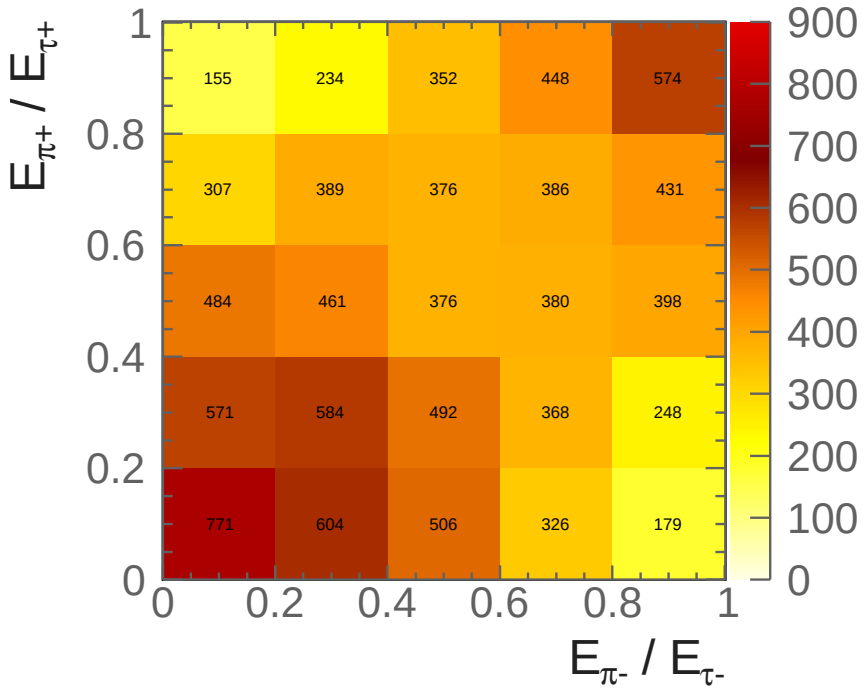
The training the multivariate classifier follows the procedure in section 6.5. The same classifier with the same parameters as in the previous tau decay mode classification is used. In the tau decay mode classifier applying stage, $\tau^- \to \pi^- \nu_\tau$ decay mode is selected with an additional criteria that there is at least one charged pion among the tau decay products.

6.8.6 Result

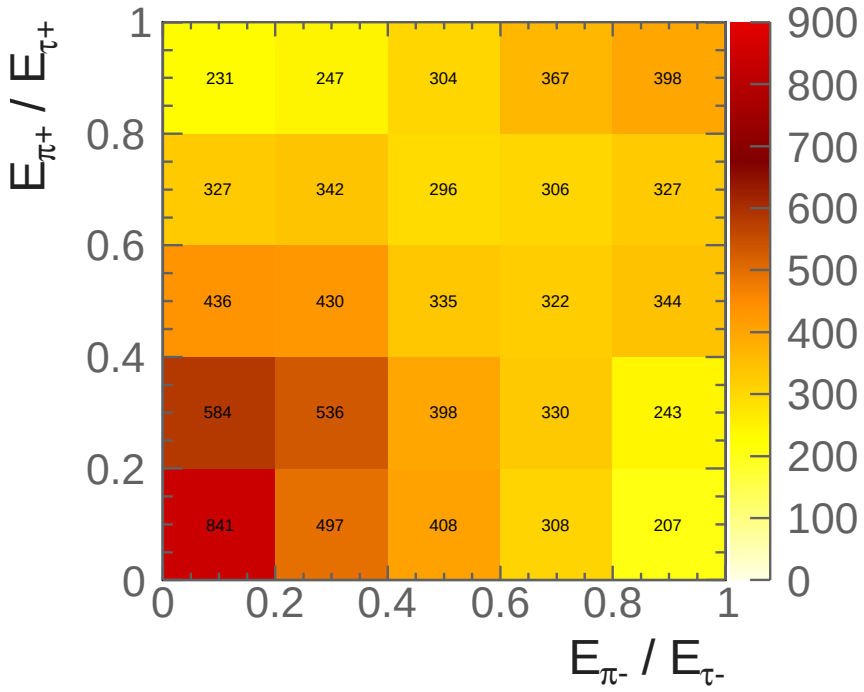
Figure 6.9 shows the two-dimensional distributions of tau decay product energy fractions, with $e^+e^- \to ZZ$ channel where one Z decays to a tau pair and the other Z decays hadronically, using both $\tau^- \to \pi^- \nu_\tau$ decay mode. The energy fractions of the tau decay product to the tau lepton (E_{π^+}/E_{τ^+} and E_{π^-}/E_{τ^-}) are the appropriate kinematic variables to study the tau pair polarisation correlation, motivated in the theoretical discussion in section 2.9.

Figure 6.9a shows the two-dimensional tau decay product energy fraction distribution obtained with the generator-level Monte Carlo particles. Figure 6.9b shows the distribution using the full detector simulation. A good match between the distributions obtained with the generator-level MC particles and the full detector simulation is achieved. Dark regions along the diagonal can be seen in both the distribution for the Monte Carlo particles and the distribution for the full detector simulation. In the $Z \to \tau^+\tau^-$ decays, where both $\tau^- \to \pi^- \nu_\tau$, an energetic $\pi^\pm$ is likely to be associated with an energetic $\pi^\mp$ and a low-energy $\pi^\pm$ is likely to be associated with a low-energy $\pi^\mp$. Comparing the two figures, some events in the top right quadrant, corresponding to both $\pi^\pm$ being energetic, are not reconstructed correctly in the full detector simulation. This is due to the incorrect finding of the tau pair decay products in section 6.8.2.

This proof-of-principle analysis shows the tau polarisation correlations, using the $e^+e^- \to ZZ$ channel where one Z decays hadronically and the other Z decays to a tau pair and subsequently both $\tau^- \to \pi^- \nu_\tau$, is possible to be observed with the ILD detector model at $\sqrt{s} = 350\,\text{GeV}$. With a similar study of the $e^+e^- \to HZ$ channel where H decays hadronically and Z decays to a tau pair and both $\tau^- \to \pi^- \nu_\tau$, the tau pair decay products energy distribution can be used to statistically identify if the parent boson is a Higgs boson or a Z boson.



(a) Generator-level Monte Carlo particles



(b) Full detector simulation

Figure 6.9: Two-dimensional distributions of E_{π^+}/E_{τ^+} as a function of E_{π^-}/E_{τ^-} from $Z \to \tau^+\tau^-$ decay where both $\tau^- \to \pi^-\nu_\tau$, with $e^+e^- \to ZZ$ channel where one Z decays to a tau pair and the other Z decays hadronically, for a) generator-level Monte Carlo particles, and b) the full detector simulation. The $\tau^- \to \pi^-\nu_\tau$ decay is selected using a) the truth information, and b) the MVA classifier.

Chapter 7

Double Higgs Boson Production Analysis

'Life is really simple, but we insist on making it complicated.'
— Confucius, 551 BC − 479 BC

Having discovered a Higgs-like particle the LHC in 2012 [3, 4],it became crucial to understand the interaction between the Higgs and other particles,and to determine whether it is the 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 [13, 14].The study of double Higgs production would prevail the measurement of the Higgs trilinear self coupling, g_{HHH}, and the quartic coupling, g_{WWHH}. The precision for the measurement of g_{HHH} achievable by the Compact Linear Collider (CLIC), is superior to that at the LHC and the HL-LHC [22].

In e^+e^- collisions, there are two main challenges with the study the double Higgs production, $e^+e^- \to HH\nu_e\overline{\nu}_e$. Firstly, the process has a small cross section; 0.149 fb at $\sqrt{s}$ = 1.4 TeV and 0.588 fb at $\sqrt{s}$ = 3 TeV. The other challenge is that at high centre-of-mass energies, events are often boosted. Consequently, many final-state particles are in the forward region of the detector, where the reconstruction performance is inferior to the barrel region. In addition, 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 [24].

7.1 Analysis Strategy Overview

The study of 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 the 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}. Figures 7.1c and 7.1d show the Feynman diagrams for irreducible background processes in the study of g_{HHH} and g_{WWHH}.

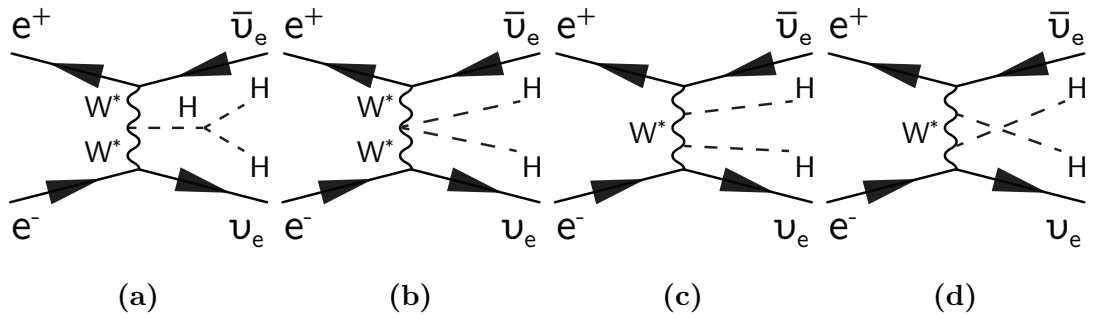


Figure 7.1: The main Feynman diagrams for the leading-order $e^+e^- \to HH\nu_e\overline{\nu}_e$ processes at CLIC.

Double Higgs production can be also proceded via $e^+e^- \to ZHH$, where the Z decays to $\nu\overline{\nu}$. The ZHH channel has been studied in e^+e^- collisions $\sqrt{s} = 500\,\text{GeV}$ [92]. However, for the 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 it can be neglected.

The two Higgs in the $e^+e^- \to HH\nu_e\overline{\nu}_e$ decay to a range of particles. Hence double Higgs production has several distinct final-state topologies. The sub-channel with the largest cross section, $HH \to b\overline{b}b\overline{b}$, has been studied by collaborators at CERN. In this chapter, the $HH \to b\overline{b}W^+W^-$ sub-channel is investigated. Firstly, the $HH \to b\overline{b}W^+W^-$ sub-channel is studied for fully hadronic decays of the W^+W^-; fully hadronic W^+W^-

decays have the largest branching fraction and the lack of neutrinos in the final states allows each W to be reconstructed. The semi-leptonic final state of the W^+W^- system in $HH \to b\bar{b}W^+W^-$ is also studied. Here the presence of the neutrino in the final state makes it difficult to reconstruct the two Higgs bosons.

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.

The chapter is organised as follows. Firstly, suitable signal and background channels are identified. Events with isolated high-energy leptons are discarded. Vertex information is used to identify b quark jets, in return to help to select signal events. The particles are clustered into jets and afterwards, the jets are used as inputs for pre-selection and multivariate analysis.

The event reconstruction was first performed for $\sqrt{s} = 1.4\,\text{TeV}$ and then $\sqrt{s} = 3\,\text{TeV}$, 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

A full list of generated samples with their cross sections can be found in table 7.1. All samples were generated with the CLIC_ILD detector model.

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 [93].

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 $e^+e^- \to qqqq\nu\bar{\nu}$ and $e^\pm\gamma \to \nu qqqq$. For the same reason, single Higgs boson production, such as $e^+e^- \to qqH\nu\bar{\nu}$, has a similar final state to the signal events and 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 fermions and usually correspond to the contributions from multiple Feynman diagrams. These diagrams are already accounted for in the generated samples for explicit Higgs production. 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, $HH \to b\bar{b}W^+W^-$, is scaled according to values listed in [94], as the values are more updated than the default Higgs branching ratios in the generator software.

The simulation and reconstruction chain is described in chapter 4. For some background processes, events are generated requiring that the invariant mass of the total momenta of all quarks is above 50 GeV or 120 GeV.This restricts the vent generation to the region of phase space that could be populated by the signal processes.

Finally, the beam induced background, $\gamma\gamma \to$ hadrons, is simulated and overlayed on all events. Details can be found in section 4.2.1.

7.3 Lepton identification

For the signal channel, $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$, 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 step in the event selection. Primary leptons deposit energies in the tracking detector. The impact parameter to the interaction point of the fitted track of the primary lepton is typically small. At the same time, the primary leptons often have energies above 10 GeV and are isolated from other particles. High-energy electrons and muons are stable enough to deposit energies in the calorimeters. However, tau leptons are short lived with a typical decay lifetime of 290 fs [26]. They decay before reaching the vertex detector. Therefore, only the decay products of the tau leptons can be reconstructed.

Channel $\sqrt{s} = 1.4\,\text{TeV}$	σ / fb
$e^+e^- \to HH\nu_e\overline{\nu}_e$	0.149
$e^+e^- \to HH\nu_e\overline{\nu}_e \to b\bar{b}W^+W^-$,hadronic	0.018
$e^+e^- \to HH\nu_e\overline{\nu}_e \to b\bar{b}b\bar{b}\nu_e\overline{\nu}_e$	0.047
$e^+e^- \to HH \to$ others	0.085
$e^+e^- \to q_{\parallel}q_{\parallel}H\nu\overline{\nu}$	0.86
$e^+e^- \to c\bar{c}H\nu\overline{\nu}$	0.36
$e^+e^- \to b\bar{b}H\nu\overline{\nu}$	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
$e^\pm\gamma(\text{BS}) \to e^\pm qqqq$	2317
$e^\pm\gamma(\text{EPA}) \to e^\pm qqqq$	574
$e^\pm\gamma(\text{BS}) \to \nu qqqq$	159.1†
$e^\pm\gamma(\text{EPA}) \to \nu qqqq$	34.7†
$e^\pm\gamma(\text{BS}) \to qqH\nu$	31.5*
$e^\pm\gamma(\text{EPA}) \to qqH\nu$	6.78*
$\gamma(\text{BS})\gamma(\text{BS}) \to qqqq$	21406.2*
$\gamma(\text{BS})\gamma(\text{EPA}) \to qqqq$	4018.7*
$\gamma(\text{EPA})\gamma(\text{BS}) \to qqqq$	4034.8*
$\gamma(\text{EPA})\gamma(\text{EPA}) \to qqqq$	753.0*

Table 7.1: List of signal and background samples used in the double Higgs analysis with the corresponding cross sections at $\sqrt{s} = 1.4\,\text{TeV}$. 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 labeled with * and †, events are generated with the invariant mass of the total momenta of all quarks above 50 and 120 GeV, respectively.

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.

7.3.1.1 IsolatedLeptonFinder

An optimised version of the existing IsolatedLeptonFinder reconstruction package is used. This algorithm identifies high energy electrons and muons that are isolated from other particles. The algorithm parameters were optimised by the collaborator using the $HH \to b\bar{b}b\bar{b}$ as the signal channel and the $e^+e^- \to qqqq\ell\nu$ as the background channel, as the background channels are the same for this analysis with $HH \to b\bar{b}W^+W^-$ channel.

Optimal values of the parameters 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; E_{cone} is the total energy within a cone of an opening angle of $\cos^{-1}(0.995)$ around the lepton; and the impact parameters, d_0, z_0, and r_0 are the closest Euclidean distance of the fitted track of the primary lepton 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	$d_0 < 0.02\,\text{mm};\ z_0 < 0.03\,\text{mm};\ r_0 < 0.04\,\text{mm}$
Isolation	$E^2_{cone} \leqslant 5.7\,\text{GeV} \times E - 50\,\text{GeV}^2$

Table 7.2: Optimised parameters for the IsolatedLeptonFinder processor.

7.3.1.2 IsolatedLeptonIdentifer

A complimentary electron finder, IsolatedLeptonIdentifer, was 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 reconstruction to identify leptons.

Table 7.3 lists the selection cuts for IsolatedLeptonIdentifer. The variables in the IsolatedLeptonFinder and the IsolatedLeptonIdentifer are defined in the same way. In addition: p_{T} is the transverse momentum; E_{cone1} and E_{cone2} are the total energy of PFOs within a cone around the lepton of an opening angle of $\cos^{-1}(0.995)$ and $\cos^{-1}(0.99)$ respectively.

The algorithm 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 energy above 10 GeV and $r_0 < 0.015$ mm. Afterwards, the lepton should either have $p_{\mathrm{T}} > 40\,\mathrm{GeV}$, or $E \geqslant 23\,\mathrm{GeV}^{\frac{1}{2}} \times \sqrt{E_{cone1}} + 5\,\mathrm{GeV}$. Figure 7.2a and 7.2b show the distributions of the p_{T} of identified electrons after E and r_0 cuts, for $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-} \to \mathrm{b\bar{b}qqqq}$ signal channel and $\mathrm{e^+e^-} \to \mathrm{qqqq}\ell\nu$ background channel respectively. A cut of $p_{\mathrm{T}} > 40\,\mathrm{GeV}$ preserves most signal events. Figure 7.2c and 7.2d show the distributions of the $23\,\mathrm{GeV}^{\frac{1}{2}} \times \sqrt{E_{cone1}} + 5\,\mathrm{GeV}$ as a function of E of identified electrons after E and r_0 cuts, for $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-} \to \mathrm{b\bar{b}qqqq}$ signal channel and $\mathrm{e^+e^-} \to \mathrm{qqqq}\ell\nu$ background channel respectively. A cut along the two-dimensional histogram would discard background events and leave signal events intact.

The second set of cuts is similar to the first set of cuts. Apart of the differences in the values of the cuts, lepton ID in the second set cuts is determined using the fraction of the energy deposited in the ECAL as a function of the total energy, $\frac{E_{ECAL}}{E}$: if $\frac{E_{ECAL}}{E}$ > 0.95 then the PFO is an electron; and if $0.2 > \frac{E_{ECAL}}{E} > 0.05$ then the PFO is a muon.

7.3.2 Tau lepton identification

The tau lepton has a short lifetime and decays before reaching the vertex detector and 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 decay modes.

The existing TauFinder [95] reconstruction package has been optimised. In addition, a package, IsolatedTauIdentifer, was developed to provide additional tau lepton identification.

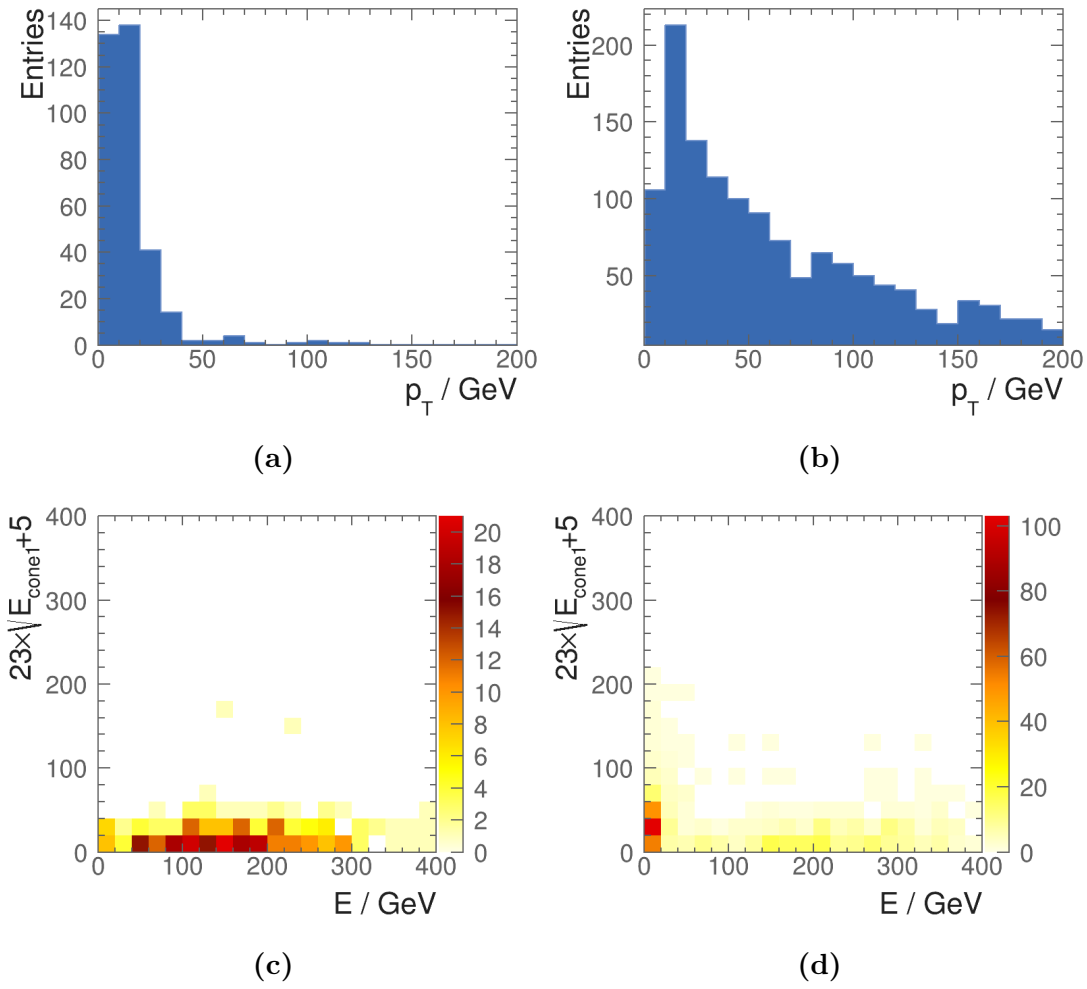


Figure 7.2: Distributions shown for p_{T} of identified electrons after E and r_0 cuts, for a) $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-} \to \mathrm{b\bar{b}qqqq}$ signal channel; and b) $\mathrm{e^+e^-} \to \mathrm{qqqq}\ell\nu$ background channel. Distributions shown for $23\,\mathrm{GeV}^{\frac{1}{2}} \times \sqrt{E_{cone1}} + 5\,\mathrm{GeV}$ as a function of E after E and r_0 cuts, for c) $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-} \to \mathrm{b\bar{b}qqqq}$ signal channel; and d) $\mathrm{e^+e^-} \to \mathrm{qqqq}\ell\nu$ background channel.

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, or	$p_T > 40\,\text{GeV}$
b) Isolation	$E \geqslant 23\,\text{GeV}^{\frac{1}{2}} \times \sqrt{E_{cone1}} + 5\,\text{GeV}$
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, or	$p_T > 40\,\text{GeV}$
b) Isolation	$E \geqslant 28\,\text{GeV}^{\frac{1}{2}} \times \sqrt{E_{cone2}} + 30\,\text{GeV}$

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

7.3.2.1 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 as a seed for the cone clustering algorithm. 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: no more than 3 charged particles in the cone; invariant mass of all PFOs in the cone 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, PFOs with low momentum (less than 1 GeV) are not used in tau finding, as they more likely come from $\gamma\gamma \to$ hadrons background. The identified tau lepton and associated decay products are then not used in further tau finding. This tau lepton finding procedure iterates with other high-energy tracks as seeds.

The optimised parameters are listed in table 7.4. The optimisation is performed by the collaborator using $HH \to b\bar{b}b\bar{b}$ signal channel and the $e^+e^- \to qqqq\ell\nu$ background channel, by scanning the parameters to obtain a good background rejection rate with lowest signal rejection rate. Variables are defined in the same way as in previous sections. In addition: 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 in the tau cone respectively; m_{tau} is the invariant mass of the sum of the PFOs in the tau candidate; and 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 track.

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

Table 7.4: Optimised parameters for the TauFinder processor.

7.3.2.2 IsolatedTauIdentifer

The IsolatedTauIdentifer works in a similar way to 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 candidate 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 the fake rate from $\gamma\gamma \to$ hadrons background, particles with energies less than 1 GeV are not considered.

Figure 7.3 show the isolation criterion of tau candidate for $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-} \to \mathrm{b\bar{b}qqqq}$ signal channel (blue) and $\mathrm{e^+e^-} \to \mathrm{qqqq}\ell\nu$ background channel (orange). Figure 7.3a shows the isolation criteria 1, where the cut at $p_{\mathrm{T}_{cone}} \geqslant 10\,\mathrm{GeV}$ selects more tau candidates

in background events than in the signal events. Similarly, in figure 7.3b, the cut at $r_0 > 0.01\,\text{mm}$, and in figure 7.3c, the cut at $p_{\text{T}cone} \geqslant 10\,\text{GeV}$ select more tau candidates in background events.

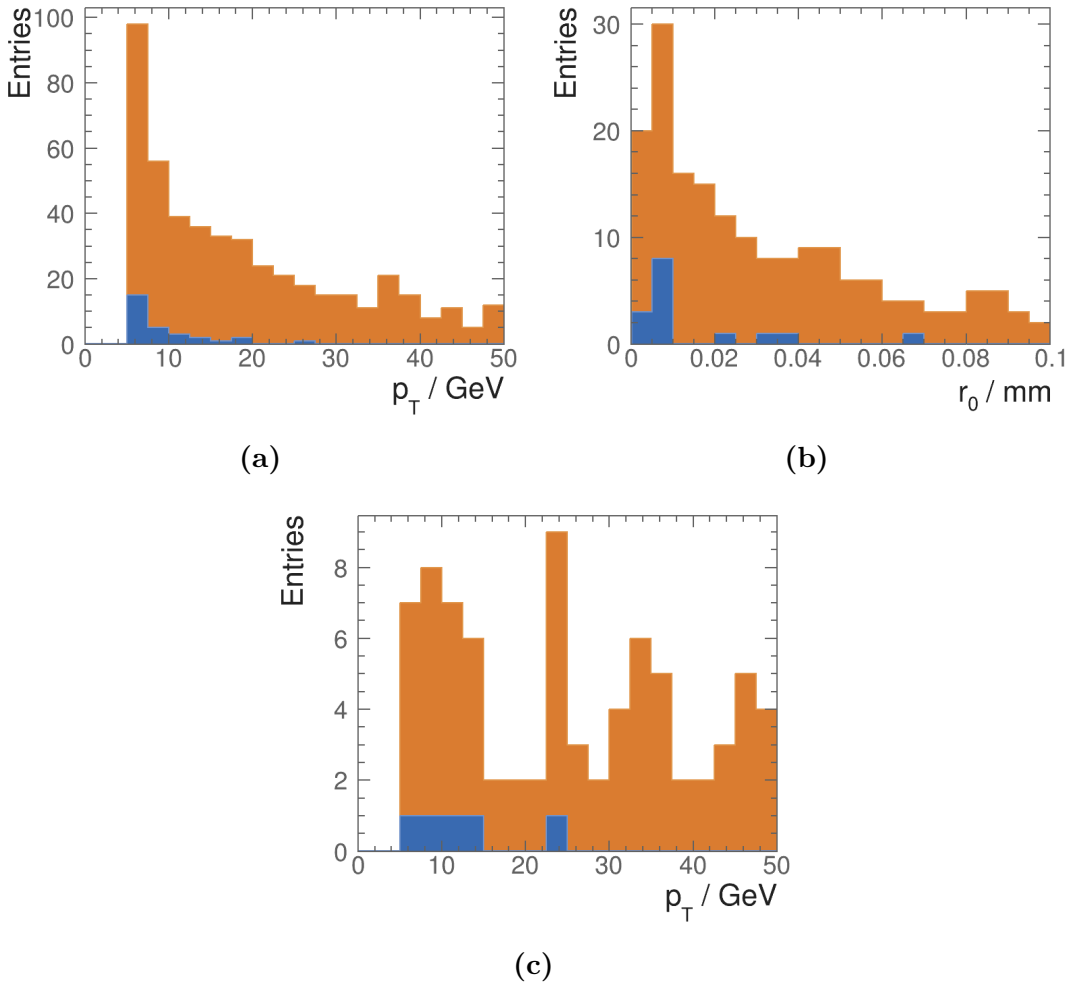


Figure 7.3: Distributions to show isolation criterion. a) Distribution of $p_{\text{T}cone}$ for isolation criteria 1 after $N_{cone1} = 0$; b) distribution of r_0 for isolation criteria 2 after $N_{X^+} = 1$, $N_{cone1} = 1$; and c) distribution of $p_{\text{T}cone}$ for isolation criteria 3 after $N_{X^+} = 3$, $N_{cone1} = 1$, $\theta_S < \cos^{-1}(0.9995)$.

Table 7.4 lists the optimised parameters for IsolatedTauIdentifer. Variables are defined in the same way as those in previous sections In addition, θ_S is the opening angle of the search cone in rad; *cone*1 and *cone*2 are defined as a cone around the tau seed of an opening angle of $\cos^{-1}(0.95)$, and $\cos^{-1}(0.99)$ respectively.

Relative to the TauFinder algorithm, the main difference is that the IsolatedTauIdentifer adopts an iterative approach to build up a tau candidate, which allows a dynamic

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 or	$N_{cone1} = 0;\ p_{\text{T}cone} \geqslant 10\,\text{GeV}$
Isolation 2 or	$N_{X^+} = 1;\ N_{cone1} = 1;\ r_0 > 0.01\,\text{mm}$
Isolation 3 or	$N_{X^+} = 3;\ N_{cone1} = 1;\ p_{\text{T}cone} \geqslant 10\,\text{GeV};\ \theta_S < \cos^{-1}(0.9995)$
Isolation 4 or	$N_{X^+} = 1;\ N_{cone2} = 0;\ r_0 > 0.01\,\text{mm};\ p_{\text{T}cone} \geqslant 10\,\text{GeV}$
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 processor

tau search cone size. The IsolatedTauIdentifer also has smaller cut values on the minimum p_T and invariant mass of the tau candidate, but stricter isolation criterions.

7.3.3 Very forward electron identification

At the high centre-of-mass energy of CLIC, particles produced are often highly boosted. Because of this, it is important to identify leptons in the forward calorimeters to aid the signal selection. In particular, photon-electron interactions, can have energetic primary electrons in the forward calorimeters, the LumiCAL and/or the BeamCAL.

Because of the large background in the forward region, it is challenging to identify primary leptons. In the Monte Carlo production, particles including the primary leptons and beam induced background in the forward calorimeters are not simulated, due to the high demand on the computational resources. Instead, studies have been performed with particles simulated in the forward calorimeters to understand the primary lepton identification efficiencies [48, 96, 97]. The studied primary lepton identification efficiencies are then parameterised as a function of lepton energies. The parametrisation approach is adopted in this analysis.

Figure 7.4a shows the primary electron identification efficiencies in the BeamCAL as a function of polar angle for a 500 GeV electron. An external processor [96] has developed to parameterise the primary electron identification efficiencies in the BeamCAL at $\sqrt{s} = 3\,\text{TeV}$ as a function of electron energy and the polar angle. The full simulation study

to obtain the primary electron identification efficiencies in the BeamCAL assumes a background integrated over 40 bunch crossings. The same primary electron identification efficiency is assumed for $\sqrt{s} = 1.4\,\text{TeV}$ and $\sqrt{s} = 3\,\text{TeV}$. In the analysis for $\sqrt{s} = 1.4\,\text{TeV}$, the momenta of the electron is scaled down by a ratio of the centre-of-mass energy to use the external processor.

Figure 7.4b shows the primary electron identification efficiencies in the LumiCAL as a function of electron energy for a polar angle $\theta = 50\,\text{mrad}$. The efficiency is obtained from a full simulation study [97], assuming a background integrated over 100 bunch crossings. In this analysis, the primary electron identification efficiency is assumed to be parasitised as a function of electron energy by the curve in figure 7.4b. The polar angle dependency of the efficiency is not considered, due to the lack of study. The primary electron identification efficiency curve in figure 7.4b takes the functional form of:

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

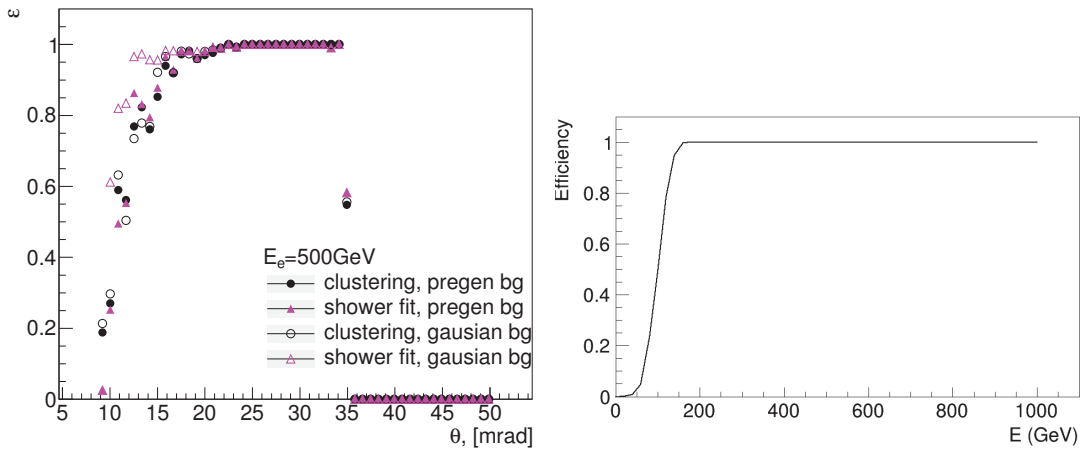
where E is the energy of the electron and erf is the error function.

Due to lack of tracking ability in the forward region, electrons and photons can not be differentiated. Therefore, both photons and electrons are identifer in the forward calorimeters. Events with identified high-energy electrons and/or photons in the BeamCAL and/or LumiCAL are rejected.

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 for $\sqrt{s} = 1.4\,\text{TeV}$. 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 for signal events and the $e^-\gamma(BS) \to e^-qqqq$ background events. Only 1% of signal events



(a) **(b)**

Figure 7.4: a) shows the 500 GeV electron identification efficiency in the BeamCAL as a function of polar angles, with different methods to model backgrounds: pregenerated and Gaussian, and two methods to identify electrons: clustering algorithm and shower fitting algorithm, obtained from a full simulation study in [96]. b) shows the electron tagging efficiency in the LumiCAL as a function of the electron energy, for a polar angle $\theta = 50$ mrad, obtained from a full simulation study in [97].

are rejected, but 47.4% of the $e^-\gamma(BS) \to e^-qqqq$ background events are rejected. Table 7.9 list the number of events surviving lepton rejection for signal and all background channels.

Efficiency (1.4 TeV)	Signal	$e^+e^- \to qqqq\ell\nu$	$e^-\gamma(\mathrm{BS}) \to e^-qqqq$
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 the lepton finding algorithms for the signal events and selected background events at $\sqrt{s} = 1.4$ TeV. Numbers represent the fractions of events where no leptons are identified by the individual lepton finder.

The lepton finding processors were optimised with events at $\sqrt{s} = 1.4$ TeV. It was found that the same set of parameters is also effective for $\sqrt{s} = 3$ TeV. The performances of the lepton finders at $\sqrt{s} = 3$ 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 ten to be boosted more and the spatial separation between particles is smaller due to the higher multiplicities. Coonsequently particles are less isolated from each other. The higher centre-of-mass energy also affects 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.

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 the lepton finding algorithms for 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.

7.4 Jet reconstruction

The signal channel, $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$, is a six-quark final state, which will result in multiple reconstructed jets. The pairing of jets to form the H, W^+ and W^- in the event is an essential part of the event reconstruction. 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 4.4.2. The free parameter for k_t algorithm is the R parameter, which controls the radius of the jet. The jet clustering will also depend on the centre-of-mass energy of the event. This is

particularly important at the CLIC because of the large beam induced background from relative low p_{T} particles. Hence a suitable level of background suppression needs to be chosen, which is incorporated in the choice of the PFO collection.

The use of the k_t jet algorithm in exclusive modes allows some particles to be clustered into beam jet, which is not used in the subsequent event reconstruction.

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, $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-} \to \mathrm{b\bar{b}qqqq}$ events are processed through k_t jet algorithm in the 6-jet exclusive mode. The six jets are paired using the MC truth information by examining the decay chain of MC particles. Four invariant mass distributions are obtained: two Higgs masses ($m_{\mathrm{H_{bb}}}$ and $m_{\mathrm{H_{WW^*}}}$) and two W masses (m_{W} and $m_{\mathrm{W^*}}$). Here $\mathrm{W^*}$ indicates the off-mass-shell W boson. The MC paring is old used to optimised the choice of parameters. It is not used in the subsequent analysis.

Three invariant mass distributions are considered: $m_{\mathrm{H_{bb}}}$, $m_{\mathrm{H_{WW^*}}}$, and m_{W}. The optimal jet reconstruction should produce sharp mass peaks around the simulated particle masses. For example, figure 7.5 shows the $m_{\mathrm{H_{bb}}}$ invariant mass distribution for $R = 1.3$ using the loose PFO collection for samples at $\sqrt{s} = 3\,\mathrm{TeV}$. An analytical functional form is fitted to describe the shape. The fitting function is a Gaussian-like function. Additional parameters are used in the fitting function to describe the tails of the distribution. The fitting function takes the form of

$$f(m) = A \exp\left\{-\frac{(m-\mu)^2}{g}\right\}, \tag{7.2}$$

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

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

To parameterise the performance of different jet algorithm settings, the overall relative width is used, 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, and with the three possible PFO collections: loose, normal, and tight.

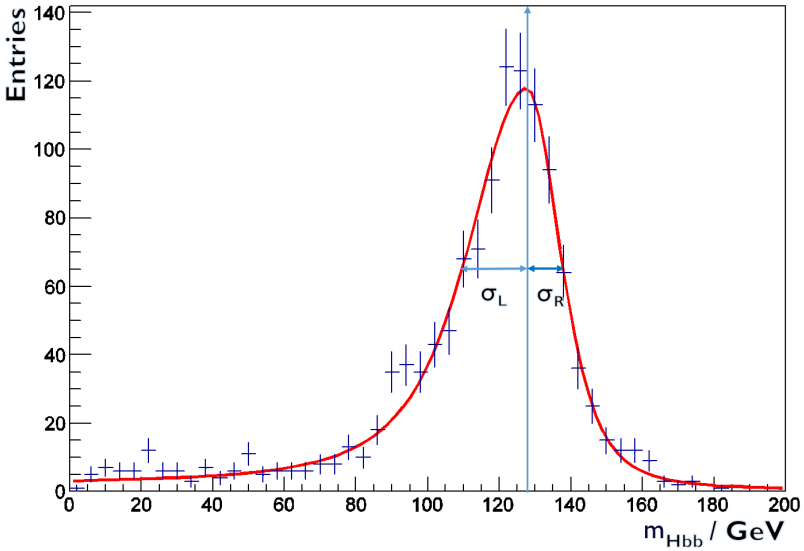


Figure 7.5: A typical example of the reconstructed $m_{\mathrm{H_{bb}}}$ mass distribution for $R = 1.3$ using loose PFO collection for $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-} \to \mathrm{b\bar{b}qqqq}$ samples at $\sqrt{s} = 3\,\mathrm{TeV}$. The fitting function is superimposed in red. The arrow shows the fitted peak position.

Figure 7.6 shows the variation of the mass peak position and its 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, whereas 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 7.6 for this analysis, it was decided to use $R = 0.7$ with the selected PFO collection. This choice gives good fitted mass peak positions for $\mathrm{H_{bb}}$, $\mathrm{H_{WW^*}}$ and W. The extracted fitted parameters of optimal jet reconstructions are summarised in table 7.8.

A separate jet reconstruction optimisation is performed for $\sqrt{s} = 3\,\mathrm{TeV}$ analysis. Figure 7.7 shows the variation of fitted mass peak positions and the relative mass

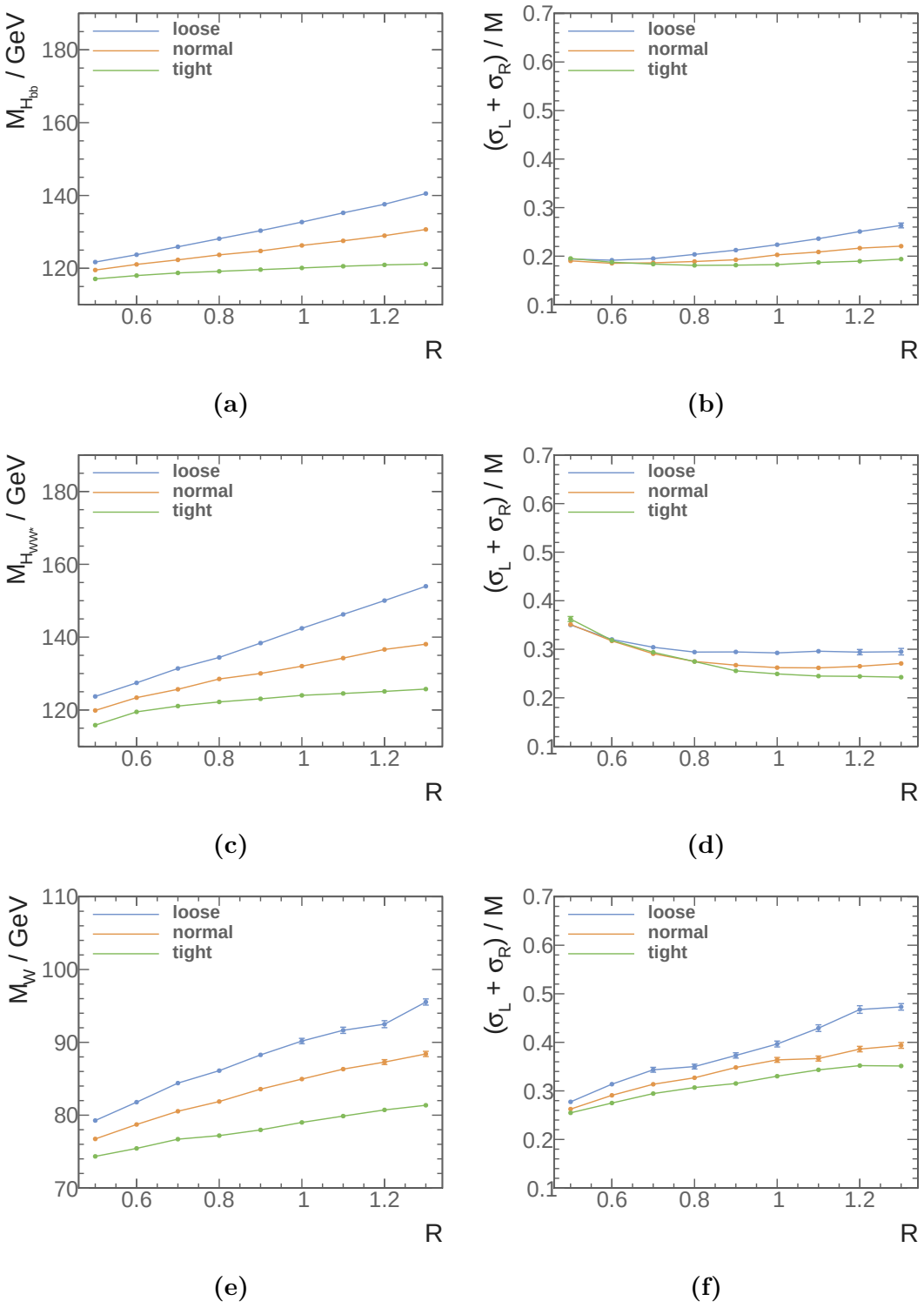


Figure 7.6: Distributions of a) fitted mass peak positions for H_{bb}, b) relative mass peak width for H_{bb}, c) fitted mass peak positions for H_{WW^*}, and f) relative mass peak width for H_{WW^*}, e) fitted mass peak positions for W, b) relative mass peak width for W. All plots show the variation of the fitted masses and mass resolutions as a function of R for loose, normal, and tight selected PFO collections at $\sqrt{s} = 1.4\,\text{TeV}$, using $\text{HH} \to \text{b}\bar{\text{b}}\text{W}^+\text{W}^- \to \text{b}\bar{\text{b}}\text{qqqq}$.

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

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.

resolutions for H_{bb}, $\text{H}_{\text{WW*}}$, and W as function of R and PFO collections. The relative mass resolution of W boson quickly degrades with increasing R. The fitted mass peak positions also increases more rapidly with the increase of R, compared with the fitted positions at $\sqrt{s} = 1.4\,\text{TeV}$. This is because 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. Based on this study, $R = 0.7$ with the tight selected PFO collection was chosen for the $\sqrt{s} = 3\,\text{TeV}$ analysis. With chosen parameters, the better relative mass resolutions compensate for the invariant masses being 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.

7.5 Jet flavour tagging

As the signal channel, $\text{HH} \to \text{b}\bar{\text{b}}\text{W}^+\text{W}^- \to \text{b}\bar{\text{b}}\text{qqqq}$, contains two b quarks in the final state, identifying jets originated from b quarks is an important part of the event selection.

The flavour tagging processor, LCFIPlus [98] is used. The processor is based on the LCFIVertex package [99], which was used in the simulation studies for the ILC Letter of Intent [29, 100] and the CLIC Concept Design Report [2].

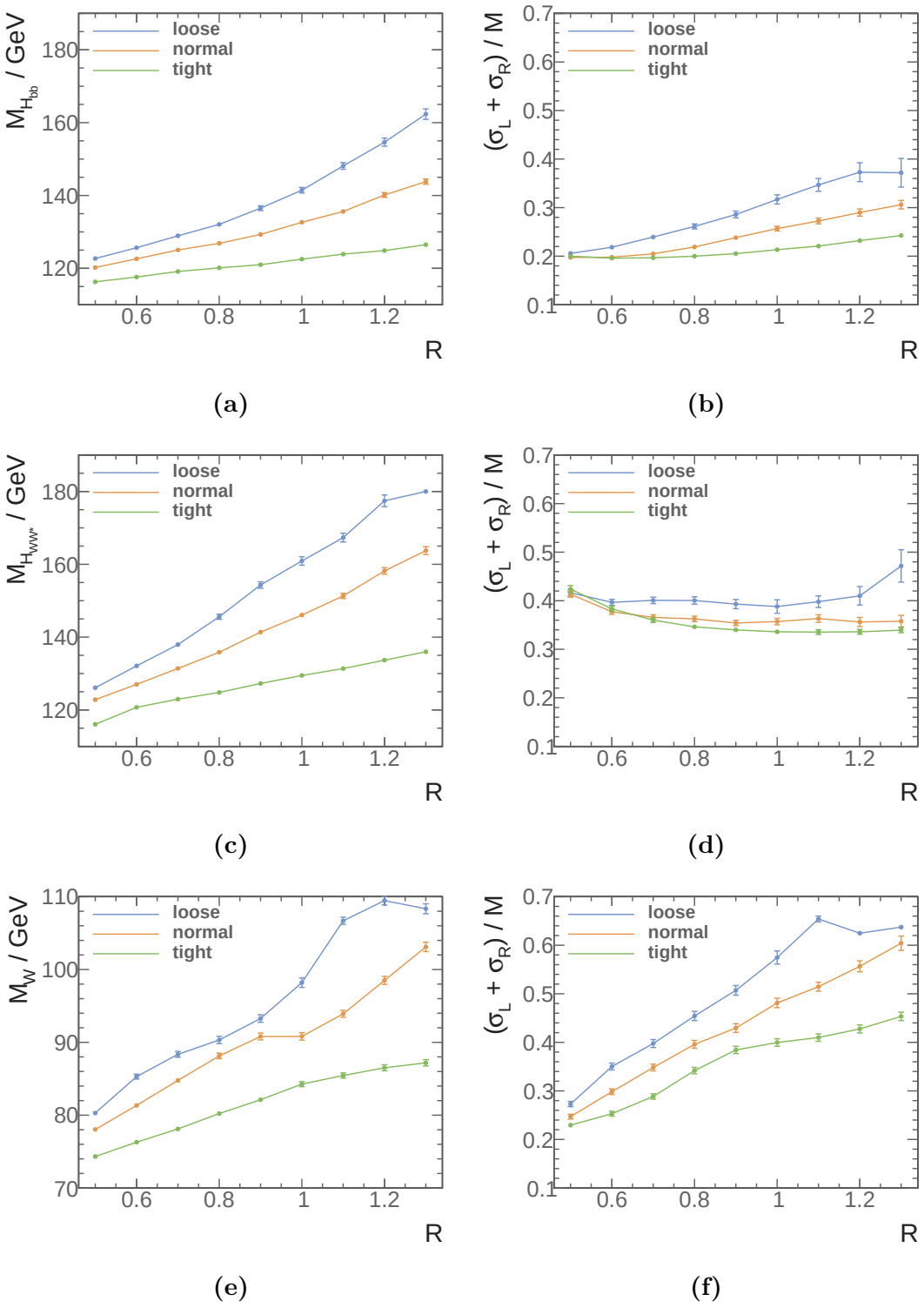


Figure 7.7: Distributions of a) fitted mass peak positions of H_{bb}, b) relative mass peak widths of H_{bb}, c) fitted mass peak positions of H_{WW^*}, d) relative mass peak widths of H_{WW^*}, e) fitted mass peak positions of W, and f) relative mass peak widths of W. All plots show the variation of fitted masses and mass resolutions as a function of R for loose, normal, and tight selected PFO collections at $\sqrt{s} = 3\,\text{TeV}$, using $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$ channel.

After the previous jet clustering step, PFOs, which are not in the beam jet, are used as inputs for the flavour tagging. The flavour tagging algorithm identifies vertices, then re-cluster PFOs into jets. Lastly, the algorithm decides if a jet us a b jet or a c jet.

The vertex finding algorithms perform vertex fitting and identify primary and secondary vertices. There are two vertex refining algorithm. First algorithm rejects the topology of a neutral particle that decays into pairs of charged particles, which can be mistaken as the decay of b or c quarks. The second algorithm is performed after the re-clustering step to reconstruct more secondary vertices, with additional information from the jet clustering.

The jet re-clustering algorithm is a modified Durham algorithm, with additional constraint of the secondary vertices and the muons, which ares identified from semi-leptonic decay of the quarks, falling into the same jet as the quarks. This ensures the topology of the jet remains consistent with the hadronic decays of heavy quarks.

Having obtained re-cluster jets, the next step is the flavour tagging, which uses a multivariate classifier to determine if a jet is from b quark or c quark. LCFIPlus uses the Boosted Decision Tree MVA multiclass classifier as implemented in the TMVA software package [77]. There are four categories for classification: jets with zero, one, two properly reconstructed vertices, or a single-track pseudo-vertex. A jet can be classified into one of three classes: a b jet, a c jet, or a light flavour quark jet (u, d or s).

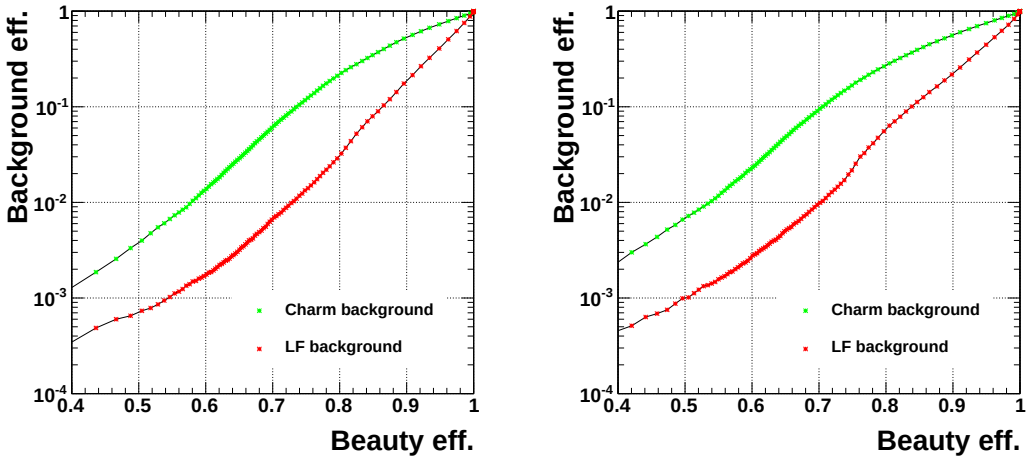
The MVA multiclass classifier was trained with $e^+e^- \to Z\overline{\nu}\nu$ event at $\sqrt{s} = 1.4\,\text{TeV}$, where Z decays to $b\overline{b}$, $c\overline{c}$, or $u\overline{u}/d\overline{d}/s\overline{s}$. The training sample contains missing momentum, which is similar to the signal sample. The training sample only has two quarks in the final states, which reduces the error in jet clustering and provides a good ground truth for training. The MVA classicisation efficiency with the training samples is shown in figure 7.8a.

Having trained the MVA classifier, the MVA classifier is applied to the samples. Under the signal hypothesis, the re-cluster algorithm is set to find six jets. The normalised distribution of the highest b-jet tag value for the $HH \to b\overline{b}W^+W^- \to b\overline{b}qqqq$ sample is shown in figure 7.9.

For the $\sqrt{s} = 3\,\text{TeV}$ analysis, the MVA classifier is re-trained with $e^+e^- \to Z\overline{\nu}\nu$ event at $\sqrt{s} = 3\,\text{TeV}$. The performance of the flavour tagging with training samples is shown in figure 7.8b. The performance at $\sqrt{s} = 3\,\text{TeV}$ is slightly worse than at $\sqrt{s} = 1.4\,\text{TeV}$,

because at the higher centre-of-mass energy, jets are more collimated and more difficult to separate.

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 7.8: Performance of b-jet tagging with $e^+e^- \to Z\overline{\nu}\nu$ samples, where Z decays to $b\overline{b}$, $c\overline{c}$, or $u\overline{u}/d\overline{d}/s\overline{s}$ at a) $\sqrt{s} = 1.4\,\text{TeV}$, and b) $\sqrt{s} = 3\,\text{TeV}$.

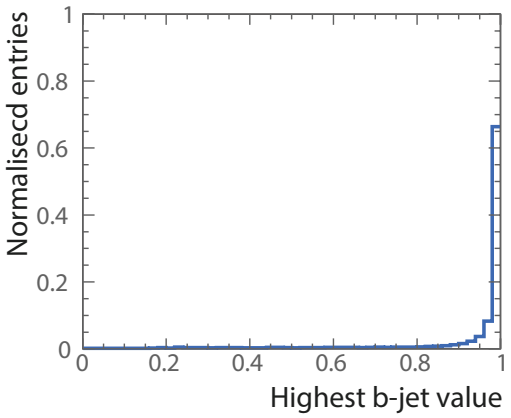


Figure 7.9: The distribution of the highest b-jet value for the $HH \to b\overline{b}W^+W^- \to b\overline{b}qqqq$ events at $\sqrt{s} = 1.4\,\text{TeV}$. The area under the curve is normalised to unity.

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

The two $e^+e^- \to HH\nu_e\bar{\nu}_e$ final states with the largest branching fractions are $HH \to b\bar{b}b\bar{b}$ (31.5%) and $HH \to b\bar{b}W^+W^-$ (25.9%). These 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.Because the results of the two studies are subsequently combined, a set of cuts are designed to separate samples, for both signal and background events, into two mutually exclusive sets for two independent analyses. This ensures there are no correlations between two analyses.

The most distinctive difference between the $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$ sub-channels is the different jet multiplicity and the different number of b-jets in the final state. Consequently variables relating to the number of b-jets and total number of jets are suitable for separating the two sub-channels.

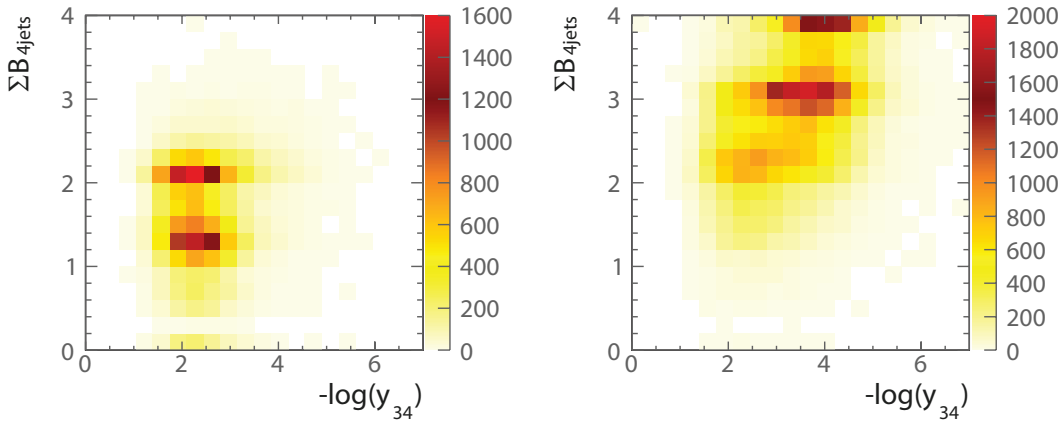
Figure 7.10 shows the sum of b-jet tag values, when the event is clustered into four jets, as a function of $-\log(y_{34})$ for the hadronic W^+W^- decay in $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$. As expected, the two sub-channels can be clearly separated in this two dimensional phase space. A rectangular cut can be used to 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 $\neg S$.

The optimal cuts are chosen such that they maximise:

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

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

Several combinations of pairs of variables were considered. In each case, the product of the fraction of the sub-channel events in each space, ε was maximised. This procedure identified $\Sigma B_{4jets} < 2.3$, $-\log(y_{34}) < 3.7$ as the best choice with 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 full list of fraction of events after passing mutually exclusive cuts for individual background processes are listed in table 7.9.



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

Figure 7.10: The two-dimensional distribution of sum of b-jet tag values against $-\log(y_{34})$. The plots show a) a) hadronic $\mathrm{W^+W^-}$ decay of $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$, and b) $\mathrm{HH} \to \mathrm{b\bar{b}b\bar{b}}$ events at $\sqrt{s} = 1.4\,\mathrm{TeV}$. The sum of b-jet tag values is calculated for the case where events are clustered into four jets.

7.6 Jet pairing

Events are reconstructed assuming the $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$ signal topology. The six jets are obtained from the jet re-clustering 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, and two jets for hadronic decay of a W*. In addition, the two Ws should be from the H boson decay.

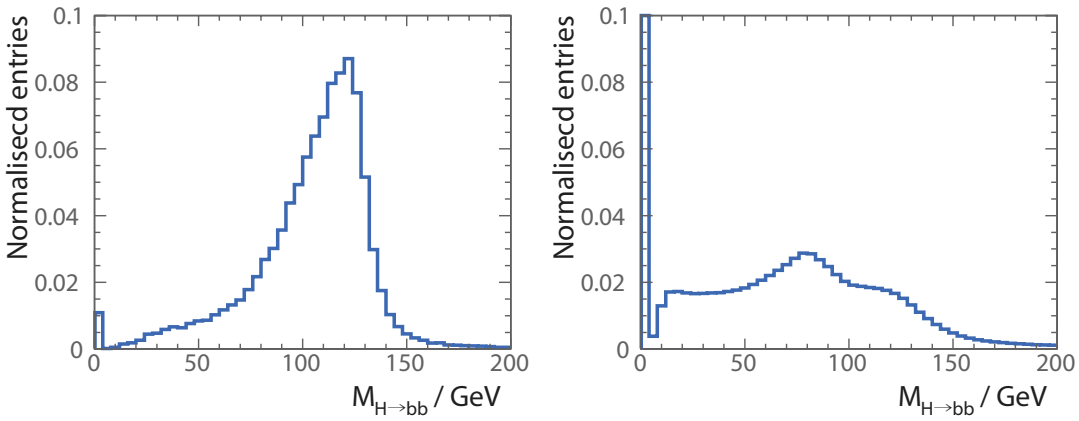
Six jets are associated to $\mathrm{H_{bb}}$, W and W*. There are 90 possible permutations for associating six jets to $\mathrm{H_{bb}}$, W and W*. The best permutation is obtained by minimising a χ^2 quantity representing the consistency of the hypothesis with the signal topology:

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

where the indices represent the six jets. The parameter μ and σ' are the expected peak and (asymmetric) width of the reconstructed mass distributions given in table 7.8, defined as:

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

A jet pairing is only considered when at least one of the jets associated to the H_{bb} decay has a b-jet tag > 0.2. Of these combinations of jets, the jet pairing giving smallest χ^2 is selected. Figure 7.11 shows the normalised distribution of $m_{H_{bb}}$ after jet pairing, for the signal channel, $HH \to b\bar{b}W^+W^-$ and the sum of all background channels. For the signal channel, the distribution peaks around the expected mass of $m_{H_{bb}}$. Around I% of signal events have no solutions for the jet pairing, as no jet has a b-jet tag > 0.2. These events are no longer considered in the analysis. The full list of fraction of events surviving after this jet pairing selection are listed in table 7.9 for signal $HH \to b\bar{b}W^+W^-$ and all background channels.



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

Figure 7.11: The distribution of $m_{H_{bb}}$ for a) the signal channel, hadronic W^+W^- decay of $HH \to b\bar{b}W^+W^-$, and b) the sum of all background channels normalised to the respective cross sections. The area under the curve is normalised to unity. All plots are shown for $\sqrt{s} = 1.4\,\text{TeV}$.

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 $m_{HH} > 150\,\text{GeV}$, as shown in Figure 7.12,

$\sqrt{s} = 1.4\,\text{TeV}$	N	Lepton evto	$b\bar{b}W^+W^-$/ $b\bar{b}W^+W^-$ separation	Valid jet Pairing
$e^+e^- \to HH\nu_e\bar{\nu}_e \to$				
$b\bar{b}W^+W^-\nu_e\bar{\nu}_e$, hadronic	27.9	89.7%	79.1%	78.3%
$e^+e^- \to HH\nu_e\bar{\nu}_e \to$				
$b\bar{b}b\bar{b}\nu_e\bar{\nu}_e$	67.6	90.8%	18.0%	18.0%
$e^+e^- \to HH\nu_e\bar{\nu}_e \to$ other	128.0	40.8%	35.8%	31.2%
$e^+e^- \to q_l q_l H\nu\bar{\nu}$	1290	72.8%	69.7%	57.7%
$e^+e^- \to c\bar{c}H\nu\bar{\nu}$	540	74.7%	59.8%	52.7%
$e^+e^- \to b\bar{b}H\nu\bar{\nu}$	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%
$e^+e^- \to qqqq\nu\bar{\nu}$	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%
$e^\pm\gamma(\text{BS}) \to e^\pm qqqq$	2606625	34.2%	33.5%	22.9%
$e^\pm\gamma(\text{EPA}) \to e^\pm qqqq$	861000.0	16.4%	15.8%	10.7%
$e^\pm\gamma(\text{BS}) \to \nu qqqq$	178987.5	85.6%	81.3%	54.4%
$e^\pm\gamma(\text{EPA}) \to \nu qqqq$	52050	44.5%	42.0%	27.4%
$e^\pm\gamma(\text{BS}) \to qqH\nu$	35437.5	70.7%	65.0%	55.4%
$e^\pm\gamma(\text{EPA}) \to qqH\nu$	10170	37.0%	33.8%	28.8%
$\gamma(\text{BS})\gamma(\text{BS}) \to qqqq$	2054951.5	85.6%	81.3%	54.0%
$\gamma(\text{BS})\gamma(\text{EPA}) \to qqqq$	4521037.5	49.6%	48.5%	32.9%
$\gamma(\text{EPA})\gamma(\text{BS}) \to qqqq$	4539150	49.6%	48.5%	32.9%
$\gamma(\text{EPA})\gamma(\text{EPA}) \to qqqq$	1129500	31.0%	30.1%	20.5%

Table 7.9: The table show the expected number of events, before cuts and after successive cuts: the lepton veto, $HH \to b\bar{b}W^+W^-/HH \to b\bar{b}b\bar{b}$ separation, and valid jet pairing. Table shows the signal and background events at $\sqrt{s} = 1.4\,\text{TeV}$, assuming an integrated luminosity of $1500\,\text{fb}^{-1}$. q can be u, d, s, b or t. Unless specified, q, ℓ and ν represent either particles or the corresponding anti-particles.

removes a small amount of signal events, but discard lots of background events, especially $\gamma\gamma \to \mathrm{qqqq}$ events.

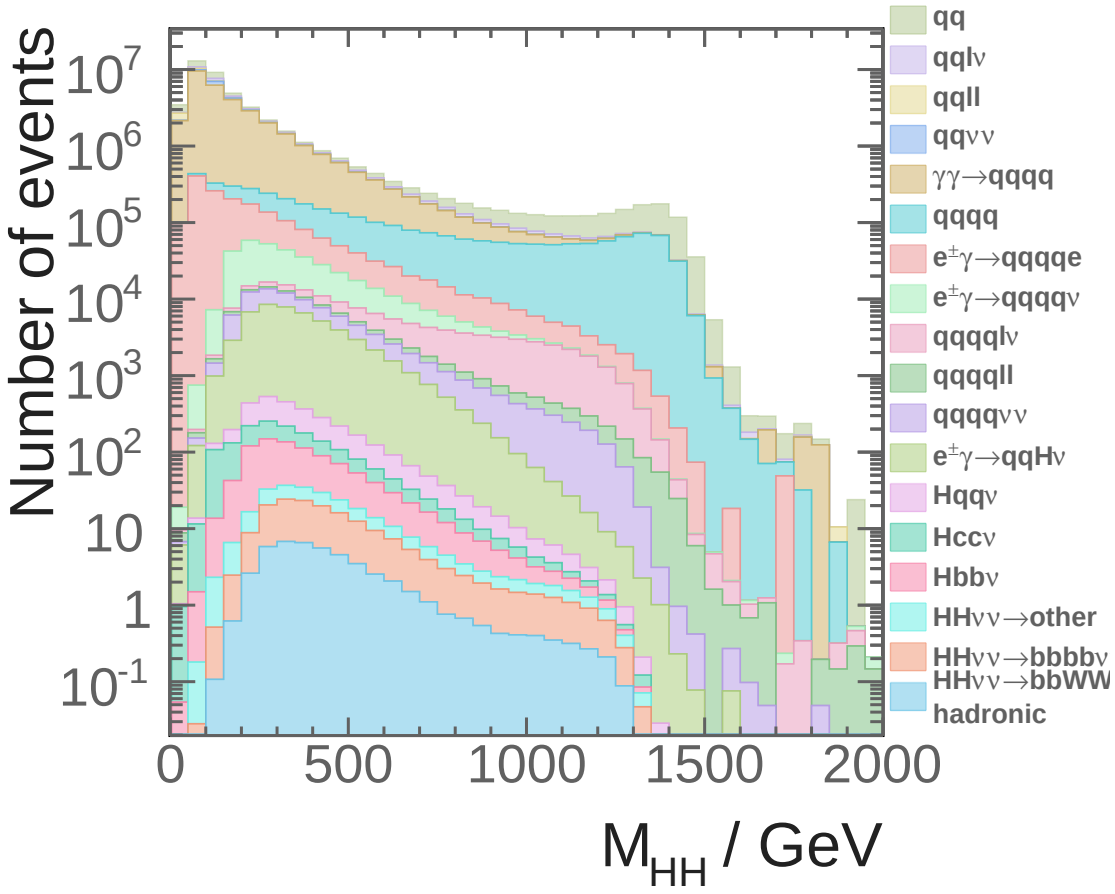


Figure 7.12: Distributions of the invariant mass of the two Higgs system for $\sqrt{s} = 1.4\,\mathrm{TeV}$, assuming an intergraded luminosity of $1500\,\mathrm{fb}^{-1}$.

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 7.13, 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 $p_{\mathrm{T}} > 30\,\mathrm{GeV}$, as shown in figure 7.14, 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.

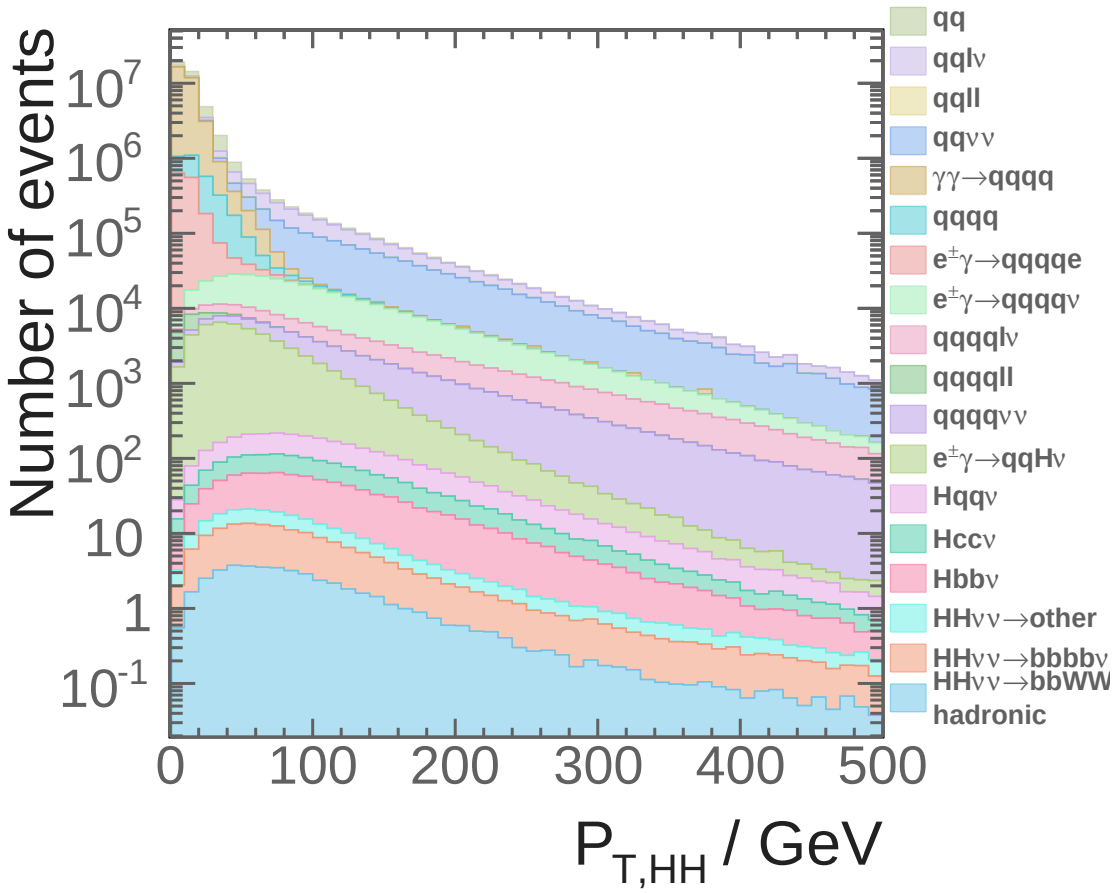


Figure 7.13: Distributions of the second highest b-jet tag value for $\sqrt{s} = 1.4\,\text{TeV}$, assuming an intergraded luminosity of $1500\,\text{fb}^{-1}$.

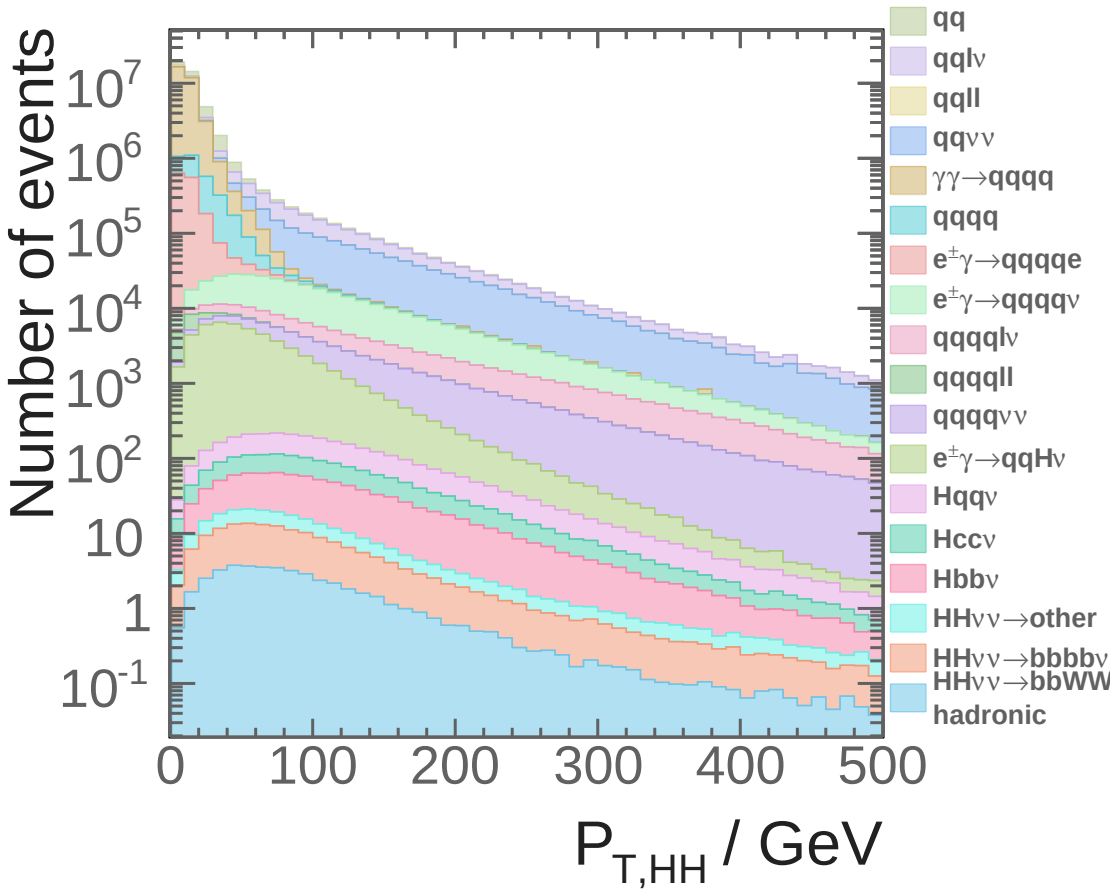


Figure 7.14: Distributions of the transverse momentum of the two Higgs system for $\sqrt{s} = 1.4\,\text{TeV}$, assuming an intergraded luminosity of $1500\,\text{fb}^{-1}$.

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: The table shows the expected number of events after successive cuts: invariant mass of the two Higgs system > 150 GeV, the second highest b-jet tag value > 0.2, and the transverse momentum of the two Higgs system > 30 GeV. All cuts include the lepton veto, $HH \to b\overline{b}W^+W^-/HH \to b\overline{b}b\overline{b}$ separation, and valid jet pairing. Table shows the signal and background events at $\sqrt{s} = 1.4$ TeV, assuming an integrated luminosity of $1500\,\mathrm{fb}^{-1}$. q can be u, d, s, b or t. Unless specified, q, ℓ and ν represent either particles or the corresponding anti-particles.

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 7.15.

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 7.15a, 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 7.15b 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 7.15c, 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 ($\cos\left(\theta^*_{W^*}\right)$), and the angle between two Higgs bosons in two Higgs bosons decay rest frame ($\cos\left(\theta^*_{HH}\right)$).

These variables are some of the most powerful variables. For example, $\cos\left(\theta^*_{H_{bb}}\right)$ for the signal events has a unform distribution, shown in figure 7.15d, as it is equally likely for two quarks to decay in any open angle in the H_{bb} decay rest frame. For the background events, by pairing jets under the signal event hypothesis, $\cos\left(\theta^*_{H_{bb}}\right)$ does not have a flat distribution. The $\cos\left(\theta^*_{H_{bb}}\right)$ 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 [101]. 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.

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 $\mathrm{H_{bb}}$ ($B_{1,\mathrm{H_{bb}}}$), the lowest b-jet tag value of the two jets associated with $\mathrm{H_{bb}}$ ($B_{2,\mathrm{H_{bb}}}$), the highest b-jet tag value of the two jets associated with W ($B_{1,\mathrm{W}}$), the highest b-jet tag value of the two jets associated with W* ($B_{1,\mathrm{W^*}}$), the highest c-jet tag value of the two jets associated with $\mathrm{H_{bb}}$ ($C_{1,\mathrm{H_{bb}}}$), and the highest c-jet tag value of the two jets associated with W ($C_{1,\mathrm{W}}$).

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 $\mathrm{H_{bb}}$ ($\mathrm{H_{bb}}$), the number of PFOs associated with $\mathrm{H_{WW^*}}$ ($\mathrm{H_{WW^*}}$), 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.

Category	Variable
Invariant mass	$m_{\mathrm{H_{bb}}}$, $m_{\mathrm{H_{WW^*}}}$, m_{W}, m_{HH}
Energy and momentum	$E_{\mathrm{W^*}}$, E_{mis}, $p_{\mathrm{TH_{bb}}}$, $p_{\mathrm{TH_{WW^*}}}$, p_{TW}, p_{THH}
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 $\sqrt{s} = 1.4\,\mathrm{TeV}$

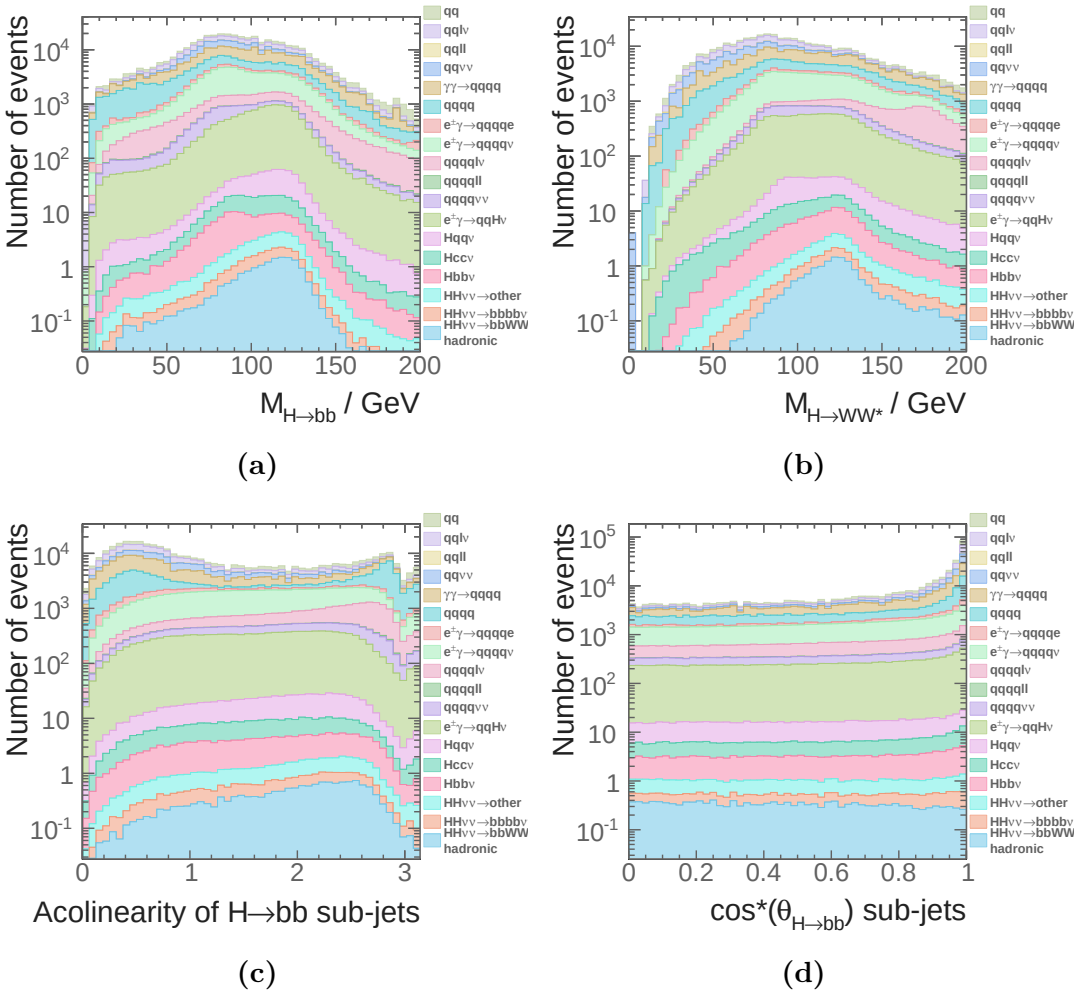


Figure 7.15: Distributions of the four variables with highest discriminating power: a) the invariant mass of H_{bb}, b) the invariant mas of H_{WW^*}, c) the acolinearity of the two jets associated with H_{bb}, and d) the opening angles of the two jets associated with H_{bb} in the decay rest frame of the H_{bb}. All plots assumes an intergraded luminosity of $1500\,\mathrm{fb}^{-1}$ at $\sqrt{s} = 1.4\,\mathrm{TeV}$ after all pre-selection cuts applied before the MVA.

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 require 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 $< 1400\,\text{GeV}$.

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 [77]. The parameters for boosted decision tree were optimised and checked for overtraining, following the strategy outlined in section 4.5. 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 W^+W^- decay of the $HH \to b\bar{b}W^+W^-$ 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 g_{HHH} and 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.

7.10 Signal selection results

Number of events passed the MVA event selection at $\sqrt{s} = 1.4\,\text{TeV}$, assuming a luminosity of $1500\,\text{fb}^{-1}$ are listed in table 7.13 for individual channels. A few background channels have non-zero events after the MVA event selection. $e^+e^- \to q\bar{q}H\nu\bar{\nu}$ 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

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 4.5.6.1 for detailed explanations of variables.

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 $e^+e^- \to HH\nu_e\overline{\nu}_e \to b\overline{b}W^+W^-\nu_e\overline{\nu}_e$ are presented.

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

The the hadronic W^+W^- decay of the $e^+e^- \to HH\nu_e\overline{\nu}_e \to b\overline{b}W^+W^-\nu\overline{\nu}$ at $\sqrt{s} = 3\,\text{TeV}$ analysis follows the same strategy as the analysis at $\sqrt{s} = 1.4\,\text{TeV}$. Lepton finding, jet 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 $\sqrt{s} = 1.4\,\text{TeV}$ analysis. Figure 7.16 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 $HH \to b\overline{b}W^+W^-$ and $HH \to b\overline{b}b\overline{b}$ sub-channels. The optimised cuts are

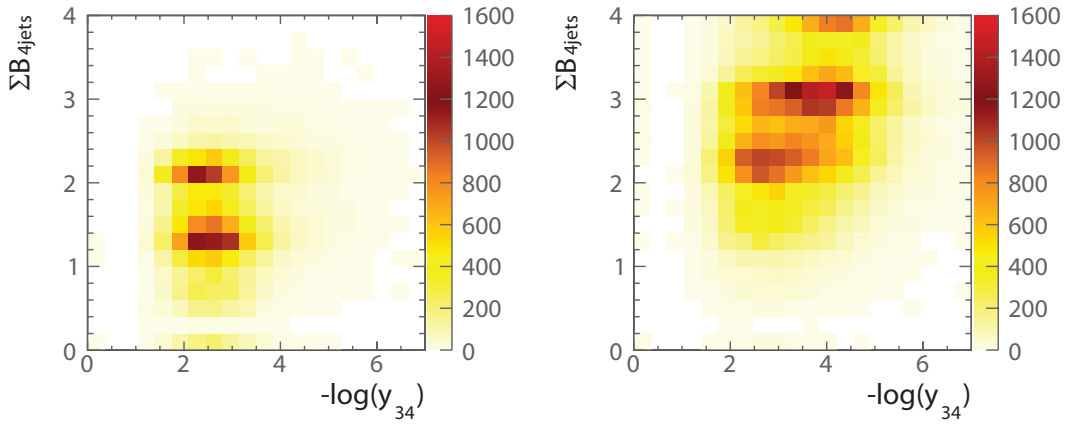
$\sqrt{s} = 1.4\,\text{TeV}$	N	ε_{presel}	ε_{MVA}	N_{MVA}
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$				
$b\bar{b}W^+W^-\nu_e\overline{\nu}_e$, hadronic	27.9	59.8%	8.2%	1.29
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$				
$b\bar{b}b\bar{b}\nu_e\overline{\nu}_e$	67.6	15.4%	0.5%	0.05
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$ other	128.0	20.4%	1.7%	0.45
$e^+e^- \to q_lq_lH\nu\overline{\nu}$	1290	39.5%	0.05%	0.29
$e^+e^- \to c\bar{c}H\nu\overline{\nu}$	540	31.6%	0.1%	0.16
$e^+e^- \to b\bar{b}H\nu\overline{\nu}$	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 an integrated of $1500\,\text{fb}^{-1}$. The number of events (N), the selection efficiencies of pre-selection cuts (ε_{presel}), the selection efficiencies of the MVA event selection after pre-selection cuts (ε_{MVA}), and the number of events after the MVA event selection (N_{MVA}) are shown. - represents a number less than 0.01. q can be u, d, s, b or t.

$\Sigma B_{4jets} < 2.3$, $-\log(y_{34}) < 3.6$. 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.

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
$e^+e^- \to q_{\parallel}q_{\parallel}H\nu\overline{\nu}$	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*
$\gamma(BS)\gamma(BS) \to qqqq$	13050.3*
$\gamma(BS)\gamma(EPA) \to qqqq$	2420.6*
$\gamma(EPA)\gamma(BS) \to qqqq$	2423.1*
$\gamma(EPA)\gamma(EPA) \to qqqq$	402.7*

Table 7.14: 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.



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

Figure 7.16: The two-dimensional distribution of sum of b-jet tag values against $-\log(y_{34})$. The plots show a) a) hadronic W^+W^- decay of $HH \to b\bar{b}W^+W^-$, and b) $HH \to b\bar{b}b\bar{b}$ events at $\sqrt{s} = 3\,\text{TeV}$. The sum of b-jet tag values is calculated for the case where events are clustered into four jets.

The pre-selection cuts at $\sqrt{s} = 3\,\text{TeV}$ use the same cut on m_{HH}. The cut on b-jet tag is different because the performance of flavour tagging is worse at $\sqrt{s} = 3\,\text{TeV}$ in comparison to the performance at $\sqrt{s} = 1.4\,\text{TeV}$. Figure 7.18 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 7.17 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 $\sqrt{s} = 3\,\text{TeV}$ are largely the same as the ones at $\sqrt{s} = 1.4\,\text{TeV}$, apart from the difference on the cut of the invariant mass of HH due to a higher $\sqrt{s}$. 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.

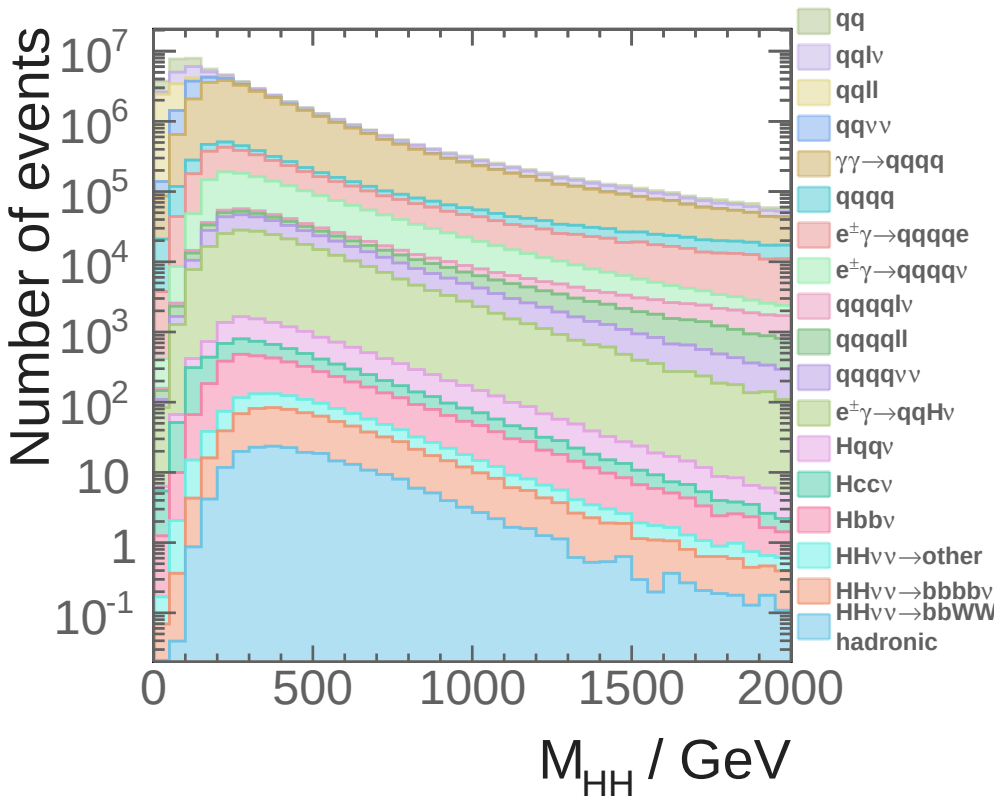


Figure 7.17: Distributions of the invariant mass of the two Higgs system for $\sqrt{s} = 3\,\text{TeV}$, assuming an intergraded luminosity of $2000\,\text{fb}^{-1}$.

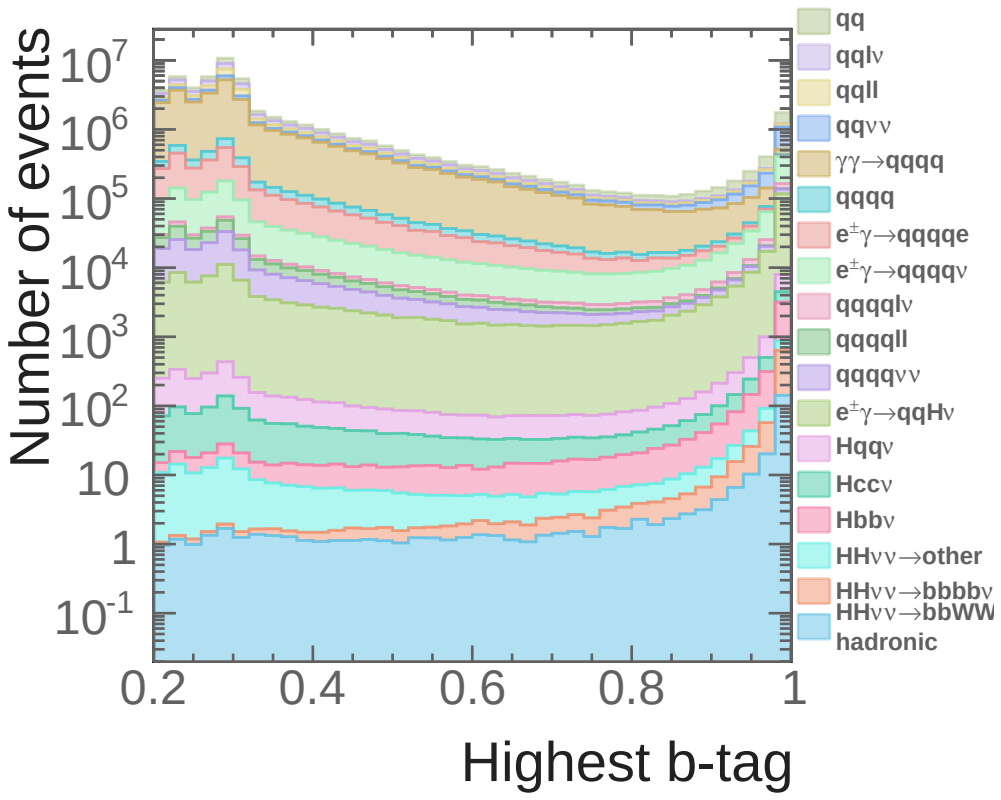


Figure 7.18: Distributions of the highest b-jet tag value for $\sqrt{s} = 3\,\text{TeV}$, assuming an intergraded luminosity of $2000\,\text{fb}^{-1}$.

$\sqrt{s} = 3\,\text{TeV}$	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 an integrated luminosity of $2000\,\text{fb}^{-1}$. The number of events (N), the selection efficiencies of pre-selection cuts (ε_{presel}), the selection efficiencies of the MVA event selection after pre-selection cuts (ε_{MVA}), and the number of events after the MVA event selection (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.

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

The final analysis is the semi-leptonic W^+W^- decay of $e^+e^- \to HH\nu_e\overline{\nu}_e \to b\overline{b}W^+W^-\nu\overline{\nu}$ 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 $HH \to b\overline{b}b\overline{b}$ 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 reduced to the invariant mass of $H_{bb} < 500\,\text{GeV}$ and the invariant mass of the double Higgs system $< 3000\,\text{GeV}$.

Figure 7.17 lists the selection efficiency and number of events after the MVA event selection for individual channel at $\sqrt{s} = 3\,\text{TeV}$, assuming a luminosity of $2000\,\text{fb}^{-1}$. 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 $\sqrt{s} = 3\,\text{TeV}$. Hence discussion of the MVA event selection is provided in section 7.10.

Category	Variable
Invariant mass	$m_{\mathrm{H_{bb}}}$, m_{W}, m_{HH}
Energy and momentum	E_{mis}, $p_{\mathrm{TH_{bb}}}$, p_{TW}, p_{THH}
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 $\sqrt{s} = 3\,\mathrm{TeV}$.

7.13 Result interpretation

The numbers of signal events and background events and significance after the MVA event selection for analyses at the $\sqrt{s} = 1.4\,\mathrm{TeV}$ and $\sqrt{s} = 3\,\mathrm{TeV}$ are listed in table 7.18. The significance is defined as the number of signal events divided by the square root of the sum of the number of the signal and background events. The $\mathrm{e^+e^-}$ collision at $\sqrt{s} = 1.4\,\mathrm{TeV}$ assumes an integrated luminosity of $1500\,\mathrm{fb^{-1}}$, whilst that of the $\sqrt{s} = 3\,\mathrm{TeV}$ assumes an integrated luminosity of $2000\,\mathrm{fb^{-1}}$. For the hadronic $\mathrm{W^+W^-}$ decay of $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$ analysis at $\sqrt{s} = 1.4\,\mathrm{TeV}$, the number of selected signal events is 1.79, and the number of selected background events is 8.41. For the hadronic analysis at $\sqrt{s} = 3\,\mathrm{TeV}$, the number of the signal is 15.45, and the number of the background is 242.28. For the semi-leptonic $\mathrm{W^+W^-}$ decay of $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$ analysis at $\sqrt{s} = 3\,\mathrm{TeV}$, the number of the signal is 31.24, and the number of the background is 3612.39.

The expected uncertainties on the measurement of the double Higgs production cross sections are estimated to be the inverse of the significance [6]:

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

The value at $\sqrt{s} = 3\,\mathrm{TeV}$ combines the results from analyses for the hadronic and semi-leptonic decay sub-channels.

$\sqrt{s} = 3\,\text{TeV}$	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
$e^+e^- \to q_lq_lH\nu\overline{\nu}$	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%	-	-

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 an integrated luminosity of $2000\,\text{fb}^{-1}$. 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.

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

Table 7.18: Number of signal (N_S) and background (N_B) events, and significance ($N_S/\sqrt{N_S+N_B}$) after MVA event selections for $HH \to b\bar{b}W^+W^-$ analyses at $\sqrt{s} = 1.4\,\text{TeV}$ and 3 TeV. The e^+e^- collision at $\sqrt{s} = 1.4\,\text{TeV}$ assumes an integrated luminosity of $1500\,\text{fb}^{-1}$, whilst that of the $\sqrt{s} = 3\,\text{TeV}$ assumes an integrated luminosity of $2000\,\text{fb}^{-1}$.

The Higgs trilinear self coupling g_{HHH} is related to the double Higgs production cross section via [24]:

$$\frac{\Delta g_{HHH}}{g_{HHH}} \approx \kappa \cdot \frac{\Delta\left[\sigma\left(HH\nu_e\overline{\nu}_e\right)\right]}{\sigma\left(HH\nu_e\overline{\nu}_e\right)}, \tag{7.12}$$

The coefficient κ can be determined by parameterising the $e^+e^- \to HH\nu_e\overline{\nu}_e$ cross section as a function of the coupling g_{HHH}. Figure 7.19 shows the $e^+e^- \to HH\nu_e\overline{\nu}_e$ cross sections as a function of the coupling g_{HHH} for $\sqrt{s} = 1.4\,\text{TeV}$ and $\sqrt{s} = 3\,\text{TeV}$. The negative gradient indicates that for all processes producing double Higgs, the process that is sensitive to the g_{HHH} experiences destructive interferences with other SM processes with the same final state. At the SM g_{HHH} value, the coefficient κ is 1.22 at $\sqrt{s} = 1.4\,\text{TeV}$, and 1.47 at $\sqrt{s} = 3\,\text{TeV}$.

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

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

Since the leading-order Feynman diagrams for the double Higgs boson production include a t-channel WW-fusion process, the cross section can be enhanced by using a polarised electron beam. For a electron beam polarisation of $P(e^-) = 80\%$, the

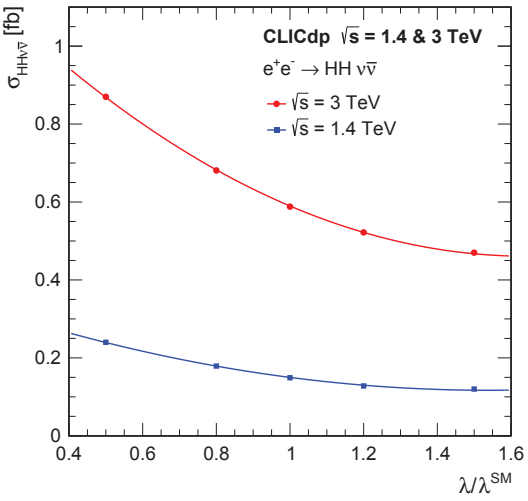


Figure 7.19: Cross section for the $e^+e^- \to HH\nu_e\overline{\nu}_e$ process as a function of the ratio λ/λ_{SM} at $\sqrt{s} = 1.4\,\text{TeV}$ and 3 TeV, taken from [24]. Here λ is the Higgs trilinear self coupling, g_{HHH}.

uncertainties of the coupling g_{HHH} become:

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

When the analyses at both $\sqrt{s} = 1.4\,\text{TeV}$ and $\sqrt{s} = 3\,\text{TeV}$ are combined, the uncertainty of the coupling g_{HHH} improves to 99% with the unpolarised beam, and to 87% with the polarised beam of $P(e^-) = 80\%$.

When the analyses for $HH \to b\overline{b}W^+W^-$ and $HH \to b\overline{b}b\overline{b}$ sub-channels are combined, the expected uncertainties on the double Higgs production cross section measurements are:

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

This translates to uncertainties on the measurement of the Higgs trilinear self coupling g_{HHH}, via equation 7.12, with unpolarised beams:

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

7.14 Simultaneous couplings extraction

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}. A two-dimensional g_{HHH} and g_{WWHH} couplings extraction is performed using the results of the hadronic analysis at $\sqrt{s} = 3\,\text{TeV}$. The integrated luminosity in this section is assumed to be $3000\,\text{fb}^{-1}$ at $\sqrt{s} = 3\,\text{TeV}$ to reflect the updated CLIC running scenario [102]. A simple scaling is applied to the results of the analysis at $\sqrt{s} = 3\,\text{TeV}$ to adapt to the change in the luminosity.

The $e^+e^- \to HH\nu_e\overline{\nu}_e$ events with non-SM couplings are generated and reconstructed. The normalised cross sections of the $e^+e^- \to HH\nu_e\overline{\nu}_e$ as a function of g_{HHH} and g_{WWHH} are shown in figure 7.20. Around the SM coupling values, the cross section increases with the decrease of g_{HHH} and with the increase of g_{WWHH}.

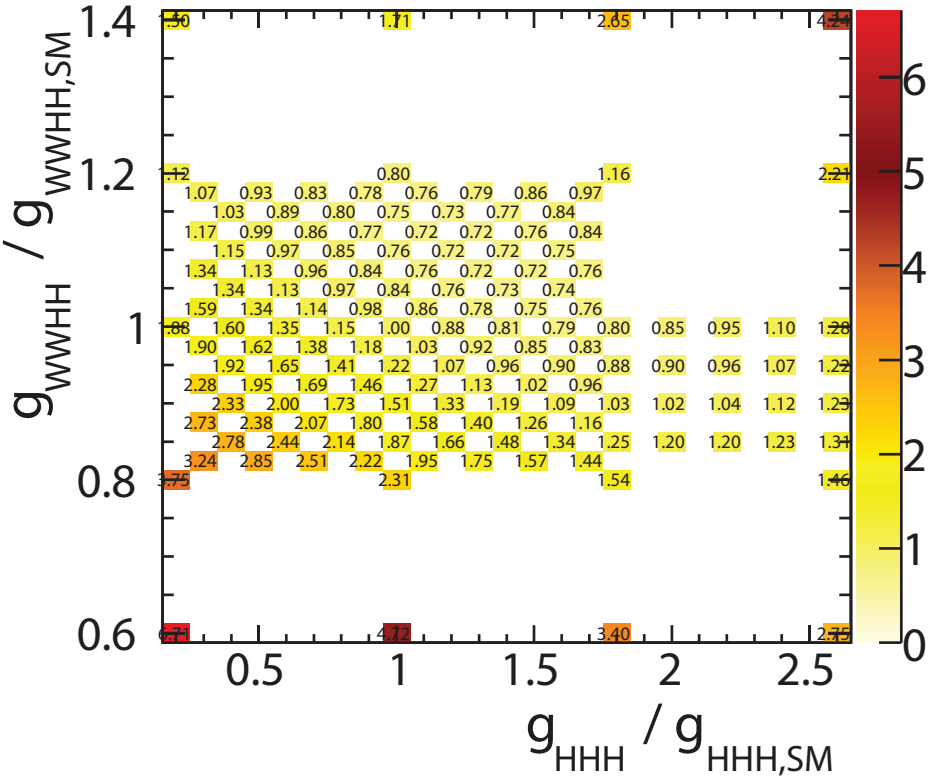


Figure 7.20: Normalised cross section for the $e^+e^- \to HH\nu_e\overline{\nu}_e$ process as a function of the g_{HHH}/g_{HHHSM} and g_{WWHH}/g_{WWHHSM} at $\sqrt{s} = 3\,\text{TeV}$. All cross sections are normalised to the cross section at the SM couplings value.

These generated non-SM coupling $e^+e^- \to HH\nu_e\overline{\nu}_e$ events went through the analysis chain discussed in this chapter with the same cuts and the same MVA classifier applied. The same of background events from the $\sqrt{s} = 3\,\text{TeV}$ analysis were added to the signal events with non-SM couplings. Figure 7.21 shows the signal significance of the double Higgs events with hadronic W^+W^- decay of $HH \to b\overline{b}W^+W^-$ sub-channel as a function of g_{HHH} and g_{WWHH}.

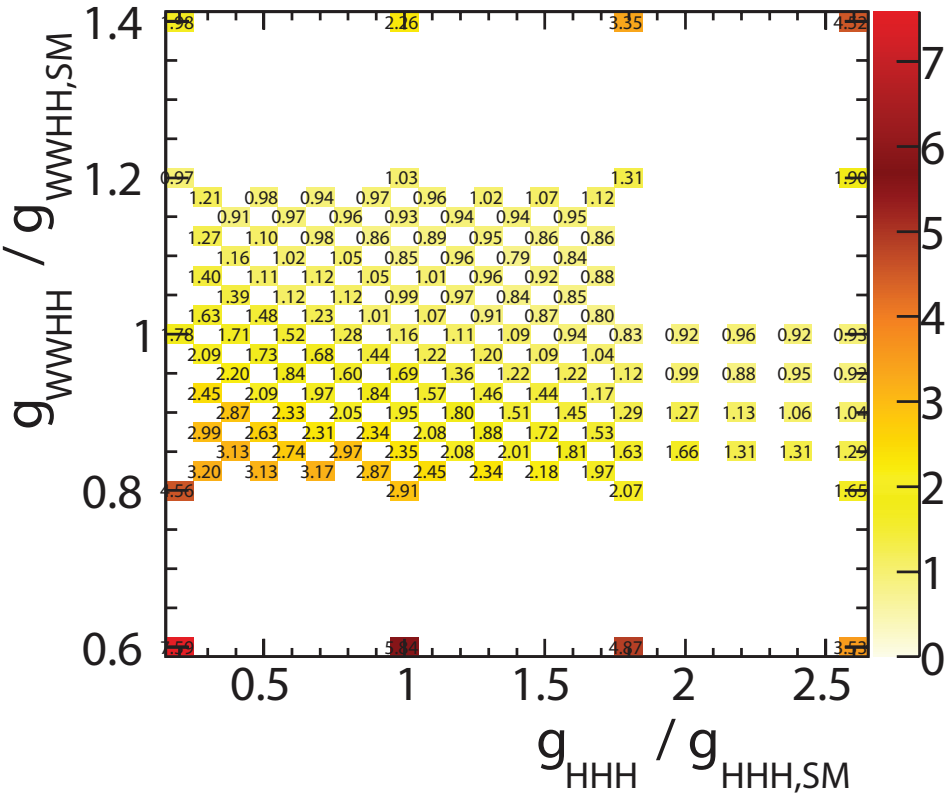


Figure 7.21: The significance for the $e^+e^- \to HH\nu_e\overline{\nu}_e$ process as a function of the g_{HHH}/g_{HHHSM} and g_{WWHH}/g_{WWHHSM} at $\sqrt{s} = 3\,\text{TeV}$, using hadronic W^+W^- decay of $HH \to b\overline{b}W^+W^-$ sub-channel, assuming an integrated luminosity of $3000\,\text{fb}^{-1}$.

Two kinematic variables that are sensitive to the change of the couplings are used to improve the extraction of the couplings (see section 2.8): the invariant mass of the two Higgs system, m_{HH}, and the scalar sum of the two Higgs transverse momentum, H_T.

The kinematic variables are subsequently binned. 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. This results in events being divided into 8 kinematic bins.

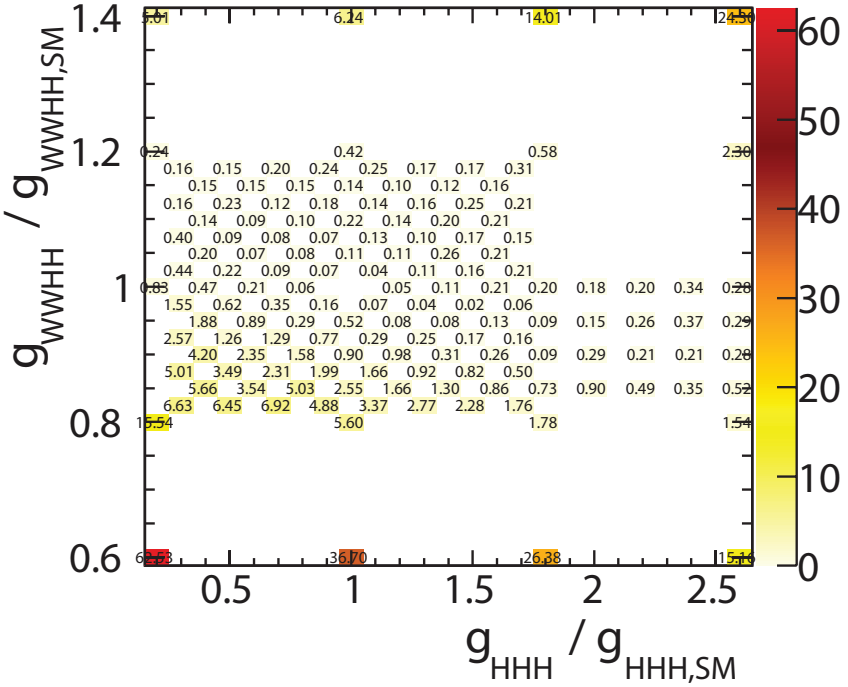
To access the change in the m_{HH} and H_T distributions for non-SM coupling samples, the χ^2 function is used:

$$\chi^2 = \sum_{i}^{8} \frac{(N_i - N_{i,observed})^2}{N_i}, \tag{7.17}$$

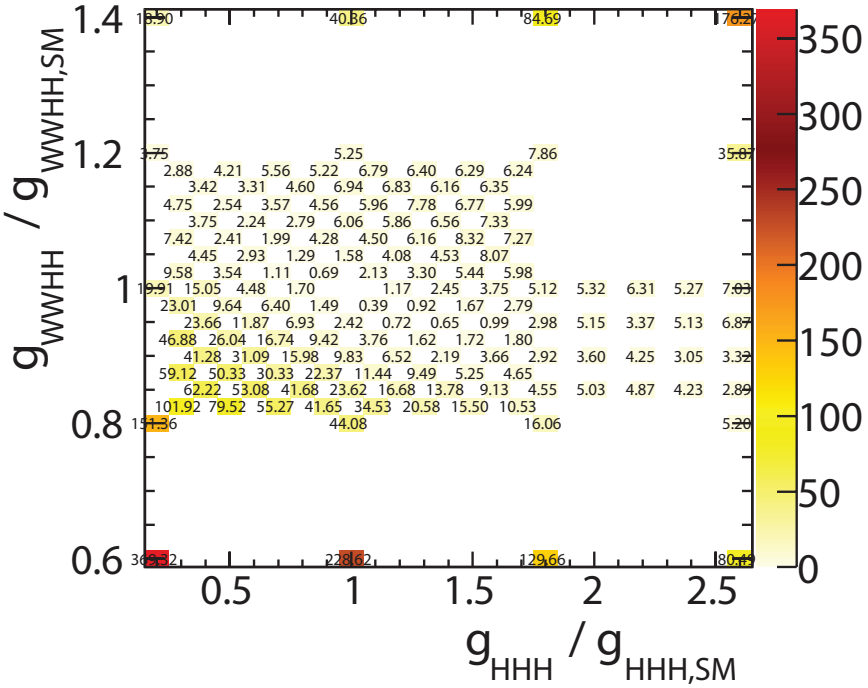
where N_i is the number of expected events in the kinematic bin i in a non-SM coupling sample; and $N_{i,observed}$ is the number of observed event in the kinematic bin i. Here the observed sample is assumed to be the SM coupling sample. The χ^2 is summed over all kinematic bins. By construction, the SM coupling sample has a χ^2 of 0. Figure 7.22 shows the χ^2 as a function of $\mathrm{g_{HHH}}$ and $\mathrm{g_{WWHH}}$ using samples from two sub-channels: hadronic $\mathrm{W^+W^-}$ decay of $\mathrm{HH \to b\bar{b}W^+W^-}$ and $\mathrm{HH \to b\bar{b}b\bar{b}}$. The χ^2 values for the $\mathrm{HH \to b\bar{b}b\bar{b}}$ sub-channel are larger because the signal significance of $\mathrm{HH \to b\bar{b}b\bar{b}}$ sub-channel is much higher than that of the hadronic $\mathrm{W^+W^-}$ decay of $\mathrm{HH \to b\bar{b}W^+W^-}$ channel.

Two sub-channels, hadronic $\mathrm{W^+W^-}$ decay of $\mathrm{HH \to b\bar{b}W^+W^-}$ and $\mathrm{HH \to b\bar{b}b\bar{b}}$, are combined to increase the statistical precision on the coupling measurements. To minimise the statistical fluctuations when generating samples with different non-SM couplings, a toy MC experiment is performed. The SM coupling sample is treated as a data template set. 100000 data sets are generated by fluctuating the event number in each kinematic bin in the data template according to the Poisson distribution with a mean that is equal to the event number in the bin. The χ^2 is calculated using these generated data sets as the observed data ($N_{i,observed}$ in equation 7.17). The χ^2 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 of χ^2 between samples with non-SM couplings and SM couplings is used for the couplings extraction, this normalisation does not affect the couplings extraction and helps to ease the visualisation. Figure 7.23 shows the normalised χ^2 after averaging over 100000 toy MC experiments as a function of $\mathrm{g_{HHH}/g_{HHH}}_{SM}$ and $\mathrm{g_{WWHH}/g_{WWHH}}_{SM}$.

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 χ^2 surface in figure 7.23. Figure 7.24 shows the contour of 68% confidence of the measurements of the $\mathrm{g_{HHH}}$ and $\mathrm{g_{WWHH}}$. The counter can be sliced one dimensionally to extract the uncertainty of the measurements of one coupling for a given value of the



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



(b) $\mathrm{HH} \to \mathrm{b\bar{b}b\bar{b}}$

Figure 7.22: The χ^2 for the $\mathrm{e^+e^-} \to \mathrm{HH}\nu_\mathrm{e}\overline{\nu}_\mathrm{e}$ process as a function of the $\mathrm{g_{HHH}}/\mathrm{g_{HHH}}_{SM}$ and $\mathrm{g_{WWHH}}/\mathrm{g_{WWHH}}_{SM}$ at $\sqrt{s} = 3\,\mathrm{TeV}$, using a) hadronic $\mathrm{W^+W^-}$ decay of $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$, b) and $\mathrm{HH} \to \mathrm{b\bar{b}b\bar{b}}$ sub-channel, assuming an integrated luminosity of $3000\,\mathrm{fb}^{-1}$.

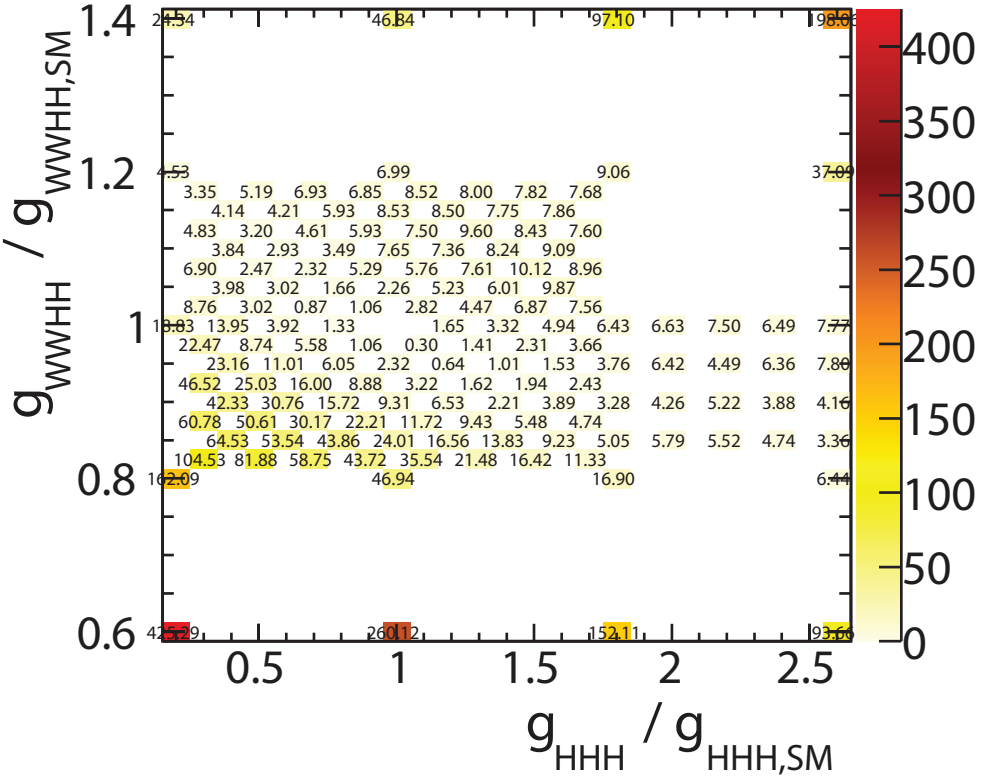


Figure 7.23: Normalised χ^2, after averaging over 100000 the toy MC experiments, as a function of $g_{HHH}/g_{HHH SM}$ and $g_{WWHH}/g_{WWHH SM}$, combining hadronic decay $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$ sub-channels, assuming an integrated luminosity of $3000\,fb^{-1}$. Normalisation is set such that the χ^2 at the SM coupling point is 0.

other coupling. For example:

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

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

The statistical precisions on the measurements of g_{WWHH} and g_{HHH} are much better at CLIC than at current LHC or high luminosity upgraded LHC [21].

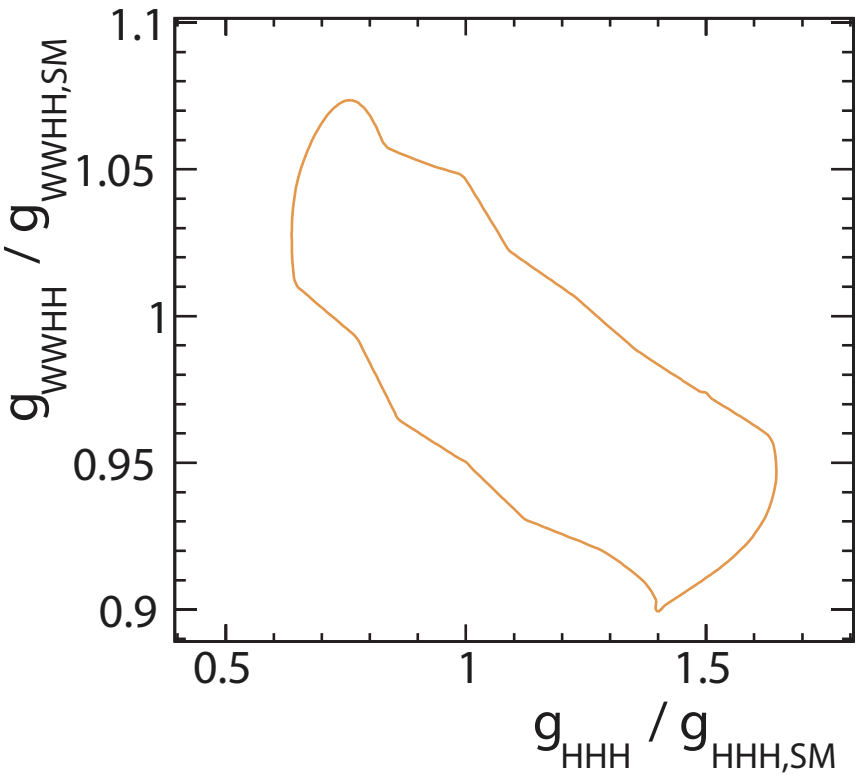


Figure 7.24: Contour plot of 68% confidence ($\chi^2 = 2.3$) , after averaging the toy MC experiments, as a function of $g_{HHH}/g_{HHH,SM}$ and $g_{WWHH}/g_{WWHH,SM}$, combining hadronic W^+W^- decay of $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$ sub-channels, assuming an integrated luminosity of 3000 fb^{-1}.

Chapter 8

Summary

'To know what you know and what you do not know, that is true knowledge.'

— Confucius, 551 BC − 479 BC

This chapter summarises key results presented in analyses in this thesis. In chapter 5, a set of photon reconstruction algorithms developed in PandoraPFA are discussed. The photon fragments produced during the event reconstruction have been greatly reduced. The ability to separate spatially close photons and the jet energy resolution have improved, as a result of a better photon reconstruction.

Using the ILD detector model, the single photon reconstruction efficiency is above 98% for photons with energies above 2GeV, and above 99.5% for photons with energies above 100 GeV. For the photon fragment reduction performance using a two-photon-per-event sample with photon energies of 500 GeV and 50 GeV, the average number of photons and particles beyond 20 mm distance separation are both less than 2.05, where the true value is 2. Photon pairs with the same energy, for example, 500 GeV−500 GeV photon pair and 10 GeV−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 GeV−50 GeV and 100 GeV−10 GeV pairs, start to be resolved at 10 mm apart, which is about two ECAL cells length. At 20 mm apart, 500 GeV−500 GeV photon pairs are fully resolved, whereas approximately only 60% of 10 GeV−10 GeV photon pairs are resolved.

In chapter 6, a high classification rate of the tau lepton seven major decay modes is achieved using the $e^+e^- \to \tau^+\tau^-$ events in the ILD detector. The tau decay mode

classification is used for the ECAL optimisation study with different ECAL cell sizes and different centre-of-mass energies. At a centre-of-mass energy of 100 GeV, the tau hadronic decay correct classification efficiency, ε_{had}, decreases from 94% at 3 mm ECAL cell size, to 91% at 20 mm cell size. Most significant decrease in the ε_{had} occurs at a centre-of-mass energy of 500 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 on the performance of the tau decay classification for centre-of-mass energies above 200 GeV, and it is more beneficial to have a small ECAL cell size at these high centre-of-mass energies.

The tau decay mode classification is also used in a proof-of-principle analysis to show that the tau polarisation correlations, using the $e^+e^- \to ZZ$ channel where one Z decays hadronically and the other Z decays to a tau pair and subsequently both $\tau^- \to \pi^-\nu_\tau$, is possible to be observed with the ILD detector model at $\sqrt{s} = 350$ GeV. With a similar study of the $e^+e^- \to HZ$ channel where H decays hadronically and Z decays to a tau pair and both $\tau^- \to \pi^-\nu_\tau$, the tau pair decay products energy distribution can be used to statistically identify if the parent boson is a Higgs boson or a Z boson.

In chapter 7, the analyses of the $e^+e^- \to HH\nu_e\overline{\nu}_e \to b\bar{b}W^+W^-\nu_e\overline{\nu}_e$ channel for CLIC at $\sqrt{s} = 1.4$ TeV and $\sqrt{s} = 3$ 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 TeV and $\sqrt{s} = 3$ TeV respectively. 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:

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

When analysis at both $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$ sub-channels are combined at $\sqrt{s} = 3$ TeV, the simultaneous extraction of the uncertainty on the measurement of the g_{HHH} and g_{WWHH} yields:

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

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

The statistical precisions on the measurements of g_{WWHH} and g_{HHH} are much better at CLIC than at current LHC or high luminosity upgraded LHC [21].

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 [103], 2011

Here are extra tables and plots for the chapter 7.

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

$\sqrt{s} = 3\,\text{TeV}$	N	Lepton evto	$b\bar{b}W^+W^-$/ $b\bar{b}W^+W^-$ separation	Valid jet Pairing
$e^+e^- \to HH\nu_e\bar{\nu}_e \to$				
$b\bar{b}W^+W^-\nu_e\bar{\nu}_e$, hadronic	146.0	80.9%	72.8%	72.1%
$e^+e^- \to HH\nu_e\bar{\nu}_e \to$				
$b\bar{b}b\bar{b}\nu_e\bar{\nu}_e$	355.0	83.5%	20.5%	20.5%
$e^+e^- \to HH\nu_e\bar{\nu}_e \to$ other	675.0	40.1%	34.3%	20.5%
$e^+e^- \to q_l q_l H\nu\bar{\nu}$	6120	67.7%	61.9%	61.9%
$e^+e^- \to c\bar{c}H\nu\bar{\nu}$	2300	69.1%	53.0%	48.8%
$e^+e^- \to b\bar{b}H\nu\bar{\nu}$	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%
$e^+e^- \to qqqq\nu\bar{\nu}$	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%
$e^\pm\gamma(\text{BS}) \to e^\pm qqqq$	4007354	31.0%	28.2%	21.1%
$e^\pm\gamma(\text{EPA}) \to e^\pm qqqq$	1151200	15.9%	14.5%	10.9%
$e^\pm\gamma(\text{BS}) \to \nu qqqq$	829184	78.3%	68.8%	53.3%
$e^\pm\gamma(\text{EPA}) \to \nu qqqq$	216800	39.6%	35.0%	26.9%
$e^\pm\gamma(\text{BS}) \to qqH\nu$	185018.0	64.0%	55.4%	49.8%
$e^\pm\gamma(\text{EPA}) \to qqH\nu$	46800	32.9%	28.8%	25.9%
$\gamma(\text{BS})\gamma(\text{BS}) \to qqqq$	18009414	71.6%	65.5%	49.4%
$\gamma(\text{BS})\gamma(\text{EPA}) \to qqqq$	3824548	44.3%	40.6%	30.6%
$\gamma(\text{EPA})\gamma(\text{BS}) \to qqqq$	3828498	44.3%	40.7%	30.7%
$\gamma(\text{EPA})\gamma(\text{EPA}) \to qqqq$	805400	29.0%	26.7%	20.1%

Table A.1: The table shows the expected number of events, before cuts and after successive cuts: the lepton veto, $HH \to b\bar{b}W^+W^-/HH \to b\bar{b}b\bar{b}$ separation, and valid jet pairing, for the signal and background events at $\sqrt{s} = 3\,\text{TeV}$, assuming an integrated luminosity of $2000\,\text{fb}^{-1}$. q can be u, d, s, b or t. Unless specified, q, ℓ and ν represent either particles or the corresponding anti-particles.

Channel	m_{HH}>150GeV	B_1>0.7
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$		
$b\overline{b}W^+W^-\nu_e\overline{\nu}_e$, hadronic	71.7%	61.8%
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$		
$b\overline{b}b\overline{b}\nu_e\overline{\nu}_e$	20.2%	18.8%
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$ other	30.2%	20.0%
$e^+e^- \to q_l q_l H\nu\overline{\nu}$	53.1%	36.0%
$e^+e^- \to c\overline{c}H\nu\overline{\nu}$	43.8%	26.3%
$e^+e^- \to b\overline{b}H\nu\overline{\nu}$	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%
$e^+e^- \to qqqq\nu\overline{\nu}$	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: The table shows the expected number of events after successive cuts: invariant mass of the two Higgs system $> 150\,\mathrm{GeV}$, and the highest b-jet tag value > 0.7. All cuts include the lepton veto, $HH \to b\overline{b}W^+W^-/HH \to b\overline{b}b\overline{b}$ separation, and valid jet pairing. Table shows the signal and background events at $\sqrt{s} = 3\,\mathrm{TeV}$, assuming an integrated luminosity of $2000\,\mathrm{fb}^{-1}$. q can be u, d, s, b or t. Unless specified, q, ℓ and ν represent either particles or the corresponding anti-particles.

Colophon

This thesis was made in LaTeX 2_ε using the "hepthesis" class [104].

Bibliography

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

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

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

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

[5] CLIC Detector and Physics Study, H. Abramowicz *et al.*, Physics at the CLIC e+e- Linear Collider – Input to the Snowmass process 2013, in *Proceedings, 2013 Community Summer Study on the Future of U.S. Particle Physics: Snowmass on the Mississippi (CSS2013): Minneapolis, MN, USA, July 29-August 6, 2013*, 2013, 1307.5288.

[6] C. Patrignani and P. D. Group, Chinese Physics C **40**, 100001 (2016).

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

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

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

[10] 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.

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

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

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

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

[15] M. Peskin and D. Schroeder, *An Introduction to Quantum Field Theory*Advanced book classics (Avalon Publishing, 1995).

[16] Y. Nambu, Phys. Rev. **117**, 648 (1960).

[17] J. Goldstone, Nuovo Cim. **19**, 154 (1961).

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

[19] 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.

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

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

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

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

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

[25] ALEPH collaboration, S. Schael *et al.*, Phys. Rept. **421**, 191 (2005).

[26] DELPHI collaboration, P. Abreu *et al.*, Phys. Lett. **B267**, 422 (1991).

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

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

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

[30] SiD, H. Aihara *et al.*, (2010).

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

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

[33] O. S. Bruning *et al.*, (2004).

[34] R. Placakyte, Parton Distribution Functions, in *Proceedings, 31st International Conference on Physics in collisions (PIC 2011): Vancouver, Canada, August 28-September 1, 2011*, 2011, 1111.5452.

[35] M. Thomson, Eur. Phys. J. **C33**, S689 (2004).

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

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

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

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

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

[41] CALICE, M. Ramilli, J. Phys. Conf. Ser. **404**, 012050 (2012).

[42] C. Adloff *et al.*, Nucl. Instrum. Meth. **A608**, 372 (2009).

[43] CALICE, C. Adloff *et al.*, JINST **9**, P01004 (2014), 1311.3505.

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

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

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

[47] C. Grah and A. Sapronov, Journal of Instrumentation **3**, P10004 (2008).

[48] 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).

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

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

[51] 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.

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

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

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

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

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

[57] M. Drees and R. M. Godbole, Phys. Rev. Lett. **67**, 1189 (1991).

[58] P. Chen, T. L. Barklow, and M. E. Peskin, Phys. Rev. **D49**, 3209 (1994), hep-ph/9305247.

[59] P. Chen, Beamstrahlung and the QED, QCD backgrounds in linear colliders, in *9th International Workshop on Photon-Photon Collisions (PHOTON-PHOTON '92) San Diego, California, March 22-26, 1992*, pp. 0418–429, 1992.

[60] D. Schulte, (1999).

[61] G. A. Schuler and T. Sjostrand, Z. Phys. **C73**, 677 (1997), hep-ph/9605240.

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

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

[64] F. Gaede, S. Aplin, R. Glattauer, C. Rosemann, and G. Voutsinas, J. Phys. Conf. Ser. **513**, 022011 (2014).

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

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

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

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

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

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

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

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

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

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

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

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

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

[78] 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).

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

[80] 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.

[81] 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.

[82] W. R. Nelson, T. M. Jenkins, R. C. McCall, and J. K. Cobb, Phys. Rev. **149**, 201 (1966).

[83] G. Bathow, E. Freytag, M. Koebberling, K. Tesch, and R. Kajikawa, Nucl. Phys. **B20**, 592 (1970).

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

[85] E. Longo and I. Sestili, Nucl. Instrum. Meth. **128**, 283 (1975), [Erratum: Nucl. Instrum. Meth.135,587(1976)].

[86] M. J. Berger and S. M. Seltzer, NASA Special Publication **3012** (1964).

[87] B. Rossi, *High-energy Particles* (Prentice-Hall, New York, 1952).

[88] S. Berge, W. Bernreuther, and S. Kirchner, Phys. Rev. **D92**, 096012 (2015).

[89] F. Gaede and J. Engels, EUDET Report (2007).

[90] E. Farhi, Phys. Rev. Lett. **39**, 1587 (1977).

[91] TMVA Core Developer Team, J. Therhaag, AIP Conf.Proc. **1504**, 1013 (2009).

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

[93] D. Lyth, Journal de Physique Colloques **35**, C2 (1974).

[94] S. Dittmaier *et al.*, (2012), 1201.3084.

[95] A. MÍźnnich, CERN Report No. LCD-Note-2010-009, 2010 (unpublished).

[96] CLICdp, A. Sailer and A. Sapronov, (2017), 1702.06945.

[97] 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.

[98] T. Suehara and T. Tanabe, Nucl. Instrum. Meth. **A808**, 109 (2016), 1506.08371.

[99] LCFI, D. Bailey *et al.*, Nucl. Instrum. Meth. **A610**, 573 (2009), 0908.3019.

[100] H. Aihara *et al.*, (2009), 0911.0006.

[101] G. Hanson *et al.*, Phys. Rev. Lett. **35**, 1609 (1975).

[102] CLICdp, CLIC, M. J. Boland *et al.*, (2016), 1608.07537.

[103] Bethesda Game Studios, *The Elder Scrolls V: Skyrim* (Bethesda Softworks, Rockville, MD, 2011).

[104] A. Buckley, The hepthesis LaTeX class.