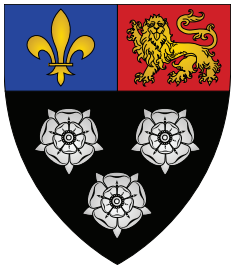
Detectors and Physics at a Future Linear Collider

Boruo Xu
of King's College



This dissertation is submitted to the University of Cambridge
for the degree of Doctor of Philosophy
on fifteenth day of September, two thousand and seventeen.

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 three studies related to the optimisation of highly granular calorimeters and one study on the sensitivity of Higgs couplings at CLIC. s

Photon reconstruction algorithms were developed for highly granular calorimeters of a future linear collider detector. 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 compared to previous reconstruction software.

The reconstruction and classification of tau lepton decay products was studied. Utilising highly granular calorimeters, the high resolution of energy and invariant mass of the tau decay products enabled a high classification rate. A hypothesis test was performed for expected decay final states. A multivariate analysis was trained to classify decay final states with a machine learning method. The performance of tau decay classification is used for the electromagnetic calorimeter optimisation at the ILC or CLIC. A proof-of-principle analysis 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 is presented.

Sensitivity of Higgs couplings at CLIC was studied using the double Higgs production process. Algorithms were developed for signal event

selection. The event selection relies on the jet reconstruction and the flavour tagging. A multivariate analysis is performed to select signal events. An attempt at extracting Higgs trilinear self-coupling and quartic coupling was conducted.

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 financial and moral support. When the PhD study became an intense and stressful exercise, they were able to put up with me and not abandon me. For that I am very grateful.

The next person I would like to thank is my supervisor, Mark Thomson. I was lucky to follow him to embark on an incredible journey on an exciting project. I have received much useful guidance from him on numerous occasions. One occasion, which influenced me greatly, was in the very early stage of my PhD study. I managed to make improvements to photon reconstruction 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 were 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. This 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 my coding skills is because of the help from John. There were a couple of months, where I had written my working algorithms, and had to rewrite the 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. 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 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 a usher on Steve's wedding. The wedding was great. And we should have more boardgame nights. I also need to thank Steve for helping generating samples in the tau analysis.

Before moving onto external collaborators, I would 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; 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 of 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 the time of writing, 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. I would like to express my gratitude 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 study.

My friends in Cambridge, whom I see on a daily basis, deserve a lot of my appreciations. It is them who made my PhD study in Cambridge lively and fun. I am again very lucky 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 suggestions 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

For the past 20 years, the high energy physics community has been considering a next-generation electron−positron collider. Measurements from the LHC helped to establish the Standard Model of particle physics. Yet there are issues that the Standard Model can not explain. For example, the origin of the masses of neutrinos and the particles that 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 International Linear Collider (ILC) [1], and the Compact Linear Collider (CLIC) [2], are the two most promising candidates of the next-generation electron−positron collider. The ILC is being designed to operate 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 the processes $e^+e^- \to ZH$ and $e^+e^- \to H\nu\nu$ and measure the 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 individual particles in an event, the event can be studied in detail. At the same time, physics simulation studies are important to demonstrate the physics reach of the future linear collider.

Chapter 2 starts with an overview of Standard Model of particle physics, including brief discussions on quantum electrodynamics, quantum chromodynamics, and the electroweak interaction. The focus of the Standard Model discussion is on the Higgs mechanism and the Higgs boson in the 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 of studying the correlation between the polarisations of the tau pair from a boson decay to determine statistically if the parent boson is a scalar or a vector.

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 detail. The chapter finishes with a discussion of the modified ILD detector concept for CLIC.

In chapter 4, the software for event simulation, event reconstruction, and event analysis is discussed. PandoraPFA, a world-leading pattern-recognition software for particle flow calorimetry is presented. A discussion of jet algorithms is provided followed by a discussion on multivariate analysis, where different fitting models, optimisation, and overfitting are described in detail.

Chapter 5 describes several new PandoraPFA algorithms for photon reconstruction. Sophisticated pattern recognition algorithms were developed using the topological properties of electromagnetic showers to identify photon candidates and separate them from nearby particles. The algorithms perform 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. The algorithms lead to a significant improvement in the reconstruction of both single photons and multiple photons in high energy jets, which in return improves the jet energy resolution at high energies.

In chapter 6, a classification of tau lepton decay modes is presented to illustrate the advantage of the highly granular linear collider detectors. The analysis contains the event generation, simulation, reconstruction, and the use of the multivariate classifier for the classification. Utilising highly granular calorimeters, the resolutions of energy and invariant mass of the tau decay products are improved. A hypothesis test was performed for expected decay final states. The performance of the tau decay mode classification is given, followed by an electromagnetic calorimeter optimisation study of the ILD detector

based on the tau decay mode classification. The tau decay mode classification is further used in a proof-of-principle analysis in chapter 7 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 $\tau^{\pm} \to \pi^{\pm}\nu_{\tau}$.

In chapter 8, 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. 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 CLIC. The event selection is further used to provide an estimate of the uncertainty on the measurements of trilinear Higgs self-coupling and quartic coupling at CLIC.

Analyses presented in chapter 5, chapter 6, and chapter 7 are solely my own work. The analysis in chapter 8 was collaborated with two researchers from CERN. Contribution from collaborators was clearly indicated in the text.

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 8. 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 7.

2.1 Overview of the Standard Model

The Standard Model (SM) [3–6] is a quantum field theory concerning the fundamental particles three of the 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 fermions have three generations. Each generation of fermions has the same set of quantum numbers,

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 [3].

but different masses. The measurements of the Z boson decay-width strongly suggested three generations of neutrinos [7].

Fermions come in two distinct categories: leptons and quarks. Neutral leptons (the neutrinos) only experience the weak force. Charged leptons ($e^{\pm}$, $\mu^{\pm}$, $\tau^{\pm}$) experience the weak force and the electromagnetic force. Quarks experience all three fundamental forces described by the SM. The 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 [8], the tau neutrino in 2000 [9], and the Higgs boson in 2012 [10, 11]. 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 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 [12, 13].

2.2 Quantum electrodynamics

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

Type	Generation	Fermion	Mass	Charge / e
Lepton	1	e^-	$0.5109989461(31)$ MeV	-1
		ν_e	-	0
	2	μ^-	$105.6583745(24)$ MeV	-1
		ν_μ	-	0
	3	τ^-	$1776.86(12)$ MeV	-1
		ν_τ	-	0
Quark	1	u	$2.2^{+0.6}_{-0.4}$ MeV	$+\frac{2}{3}$
		d	$4.7^{+0.5}_{-0.4}$ MeV	$-\frac{1}{2}$
	2	c	1270 ± 30 MeV	$+\frac{2}{3}$
		s	98^{+8}_{-4} MeV	$-\frac{1}{3}$
	3	t	$173210 \pm 510 \pm 710$ MeV	$+\frac{2}{3}$
		b	4180^{+40}_{-30} 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 masses from the observation of neutrino flavour oscillations. The upper bound on the neutrino mass is 2 eV. For the top quark mass, the statistical uncertainty is listed first, followed by the systematic uncertainty. Values are taken from [3].

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 [14] 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:

$$\left(i\gamma^\mu \partial_\mu - m\right)\psi = 0, \tag{2.2}$$

where the γ^μ are the Dirac gamma matrices with $\mu \in \{0, 1, 2, 3\}$; $\overline{\psi}$ is defined as $\psi^\dagger\gamma^0$; $F_{\mu\nu} = c_\mu A_\nu - \partial_\nu A_\mu$ is the electromagnetic field tensor; 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; and e is the coupling constant, which is equal to the electric charge. Invariance under phase transformations of the fermion fields, $\psi \to \psi' = e^{i\phi(x)}\psi$, require a gauge transformation of the electromagnsetic field:

$$A_\mu \to A'_\mu = A_\mu + \partial_\mu \chi = A_\mu - \partial_\mu \phi / e, \tag{2.4}$$

where $\chi(x) = -\phi(x)/e$. Thus gauge invariance of the whole Lagrangian is conserved when the phase of the fermion field changes according to:

$$\psi \to \psi' = e^{-ie\chi(x)}\psi \tag{2.5}$$

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 charges, usually labelled as red, green, and blue. Anti-particles carry anticolour. 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:

$$\mathcal{L}_{QCD} = \sum_{f \in u,d,s,c,b,t} \overline{\psi_i} \left(\left(i\gamma^\mu \partial_\mu - g_s \gamma^\mu G^a_\mu \frac{\lambda^a}{2} \right)_{ij} - m_f \delta_{ij} \right) \psi_j - \frac{1}{4} G^a_{\mu\nu} G^{a\mu\nu}, \tag{2.6}$$

where ψ represents a quark with a colour charge, indicated by i or j; m is 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:

$$G^a_{\mu\nu} = \partial_\mu \gamma^a_\nu - \partial_\nu \gamma^a_\mu - g_s f_{abc} G^b_\mu G^c_\nu, \tag{2.7}$$

where G^b_μ is the gluon field with colour charge b; and a, b, and c indicate the colour charges.

2.4 The electroweak interaction

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

$$m\overline{\psi}\psi = \frac{1}{4}m\overline{\psi}\Big(1-\gamma^5\Big)\Big(1-\gamma^5\Big)\psi + \frac{1}{4}m\overline{\psi}\Big(1+\gamma^5\Big)\Big(1+\gamma^5\Big)\psi \tag{2.8}$$

$$= m\overline{\psi_R}\psi_L + m\overline{\psi_L}\psi_R \tag{2.9}$$

As ψ_L and $\overline{\psi_L}$ transform under $\mathrm{SU(2)_L}$ gauge symmetry while ψ_R and $\overline{\psi_R}$ do not, the mass term is not gauge invariant. Consequently, fermion mass terms are forbidden under $\mathrm{SU(2)_L}$ gauge symmetry.

The electroweak Lagrangian can be written as

$$\mathcal{L}_{Electroweak} = \mathcal{L}_{Boson} + \mathcal{L}_{Fermion} + \mathcal{L}_{Higgs} + \mathcal{L}_{Yukawa}\,. \tag{2.10}$$

The terms are:

1. $\mathcal{L}_{Boson}$, which is given by:

$$\mathcal{L}_{Boson} = -\frac{1}{4}W^i_{\mu\nu}W^{i\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}, \tag{2.11}$$

$$W^i_{\mu\nu} = \partial_\nu W^i_\mu - \partial_\mu W^i_\nu - g\varepsilon^{ijk}W^j_\mu W^k_\nu, \tag{2.12}$$

$$B_{\mu\nu} = \partial_\nu B_\mu - \partial_\mu B_\nu, \tag{2.13}$$

   where 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;

2. $\mathcal{L}_{Fermion}$, which 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 SU(2) singlets.

The left-handed fermions are in SU(2) doublets with the corresponding fermions of the same generation. The term $\mathcal{L}_{Fermion}$ is given by:

$$\mathcal{L}_{Fermion} = \sum_{\psi \in fermions} \overline{\psi}_L \gamma^\mu D_\mu^L \psi_L + \overline{\psi}_R \gamma^\mu D_\mu^R \psi_R, \tag{2.14}$$

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

$$D_\mu^L = \partial_\mu + ig\frac{\tau_i}{2}W_\mu^i + ig'Y_\psi B_\mu, \tag{2.15}$$

$$D_\mu^R = \partial_\mu + ig'Y_\psi B_\mu. \tag{2.16}$$

The structure of this Lagrangian allows the 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;

3. $\mathcal{L}_{Higgs}$, which describes the Higgs field. After electroweak symmetry breaking of the Higgs field, the mass terms of the gauge bosons are introduced; and

4. $\mathcal{L}_{Yukawa}$, which 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:

$$\mathcal{L} = \partial^\mu \psi^* \partial_\mu \psi - m^2|\psi|^2 = \partial^\mu \psi^* \partial_\mu \psi - V(\psi), \tag{2.17}$$

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:

$$V(\psi) = m^2|\psi|^2 + \lambda|\psi|^4, \tag{2.18}$$

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:

$$\frac{\partial V(\psi)}{\partial |\psi|} = 2m^2|\psi| + 4\lambda|\psi|^3 = 0, \tag{2.19}$$

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

$$|\psi| = \sqrt{\frac{-m^2}{2\lambda}} \equiv \frac{\nu}{\sqrt{2}}, \tag{2.20}$$

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 [15, 16]. The theorem states that spontaneous symmetry breaking always implies the existence of a massless particle.

To demonstrate Goldstone's theorem, the Lagrangian in equation 2.17 is used 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

$$\psi = \frac{1}{\sqrt{2}}(\nu + \psi_1 + i\psi_2), \tag{2.21}$$

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 2.17 gives:

$$\mathcal{L} = \frac{1}{2}\partial^\mu\psi_1\partial_\mu\psi_1 + \frac{1}{2}\partial^\mu\psi_2\partial_\mu\psi_2 - m^2\psi_1^2 + \dots. \tag{2.22}$$

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 Goldstone's theorem.

The Lagrangian in equation 2.17 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 gauge invariance, the covariant

derivative becomes $D_\mu = \partial_\mu + ieA_\mu$. Hence the Lagrangian in equation 2.17 becomes:

$$\mathcal{L} = (D^\mu\psi)^*(D_\mu\psi) - m^2|\psi|^2 - \lambda|\psi|^4. \tag{2.23}$$

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

$$+\frac{e^2\nu^2}{2}A^\mu A_\mu, \tag{2.24}$$

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. Consider a complex scalar Higgs field, Φ_H, that transforms as a doublet of SU(2) with hypercharge $Y = \frac{1}{2}$. The Higgs Lagrangian is given by:

$$\mathcal{L}_{Higgs} = \left(D_\mu\Phi_\mathrm{H}\right)^\dagger \left(D^\mu\Phi_\mathrm{H}\right) - \mu^2\Phi_\mathrm{H}^\dagger\Phi_\mathrm{H} + \lambda\left(\Phi_\mathrm{H}^\dagger\Phi_\mathrm{H}\right)^2, \tag{2.25}$$

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

$$D_\mu = \left(\partial_\mu + ig\frac{\tau_i}{2}W_\mu^i + ig'\frac{1}{2}B_\mu\right), \tag{2.26}$$

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:

$$V(\mathrm{H}) = \mu^2\Phi_\mathrm{H}^\dagger\Phi_\mathrm{H} - \lambda\left(\Phi_\mathrm{H}^\dagger\Phi_\mathrm{H}\right)^2. \tag{2.27}$$

The Higgs potential is minimised when

$$\sqrt{\Phi_\mathrm{H}^\dagger\Phi_\mathrm{H}} = \frac{\nu}{\sqrt{2}} = \sqrt{\frac{\mu^2}{2\lambda}}. \tag{2.28}$$

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

$$\left\langle \Phi_{\mathrm{H}} \right\rangle = \begin{pmatrix} 0 \\ \frac{\nu}{\sqrt{2}} \end{pmatrix}, \tag{2.29}$$

with a real ν. Substituting the Higgs VEV into the $\mathcal{L}_{Higgs}$ in equation 2.25, the $\left(D_\mu \Phi_{\mathrm{H}}\right)^\dagger \left(D^\mu \Phi_{\mathrm{H}}\right)$ term becomes

$$-\frac{1}{8}\begin{pmatrix} 0 & \nu \end{pmatrix} \begin{pmatrix} gW_\mu^3 + g'B_\mu & g\left(W_\mu^1 - iW_\mu^2\right) \\ g\left(W_\mu^1 + iW_\mu^2\right) & -gW_\mu^3 + g'B_\mu \end{pmatrix} \begin{pmatrix} gW_\mu^3 + g'B_\mu & g\left(W_\mu^1 - iW_\mu^2\right) \\ g\left(W_\mu^1 + iW_\mu^2\right) & -gW_\mu^3 + g'B_\mu \end{pmatrix} \begin{pmatrix} 0 \\ \nu \end{pmatrix}. \tag{2.30}$$

Ignoring the negative sign, equation 2.30 simplifies to

$$\frac{\nu^2 g^2}{8}\left(W_\mu^1 - iW_\mu^2\right)\left(W_\mu^1 + iW_\mu^2\right) + \frac{\nu^2}{8}\left(gW_\mu^3 - g'B_\mu\right)^2. \tag{2.31}$$

The physical fields W_μ^+ and W_μ^- can be identified with the first part of equation 2.31, as

$$W_\mu^+ = \frac{1}{\sqrt{2}}\left(W_\mu^1 - iW^2\mu\right), \tag{2.32}$$

$$W_\mu^- = \frac{1}{\sqrt{2}}\left(W_\mu^1 + iW^2\mu\right). \tag{2.33}$$

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 equation 2.31 should give rise to Z boson mass term only, with no mass term for the photon. This can be achieved by rearranging the second part of the equation 2.31:

$$\frac{\nu^2}{8}\left(gW_\mu^3 - g'B_\mu\right)^2 = \frac{\nu^2\left(g^2 + g'^2\right)}{8}\left(\frac{g}{\sqrt{g^2 + g'^2}}W_\mu^3 - \frac{g'}{\sqrt{g^2 + g'^2}}B_\mu\right)^2. \tag{2.34}$$

A convenient way to connect g and g' is to use the Weinberg mixing angle [17], denoted as θ_{W}. The Weinberg mixing angle is defined as

$$\cos\theta_{\mathrm{W}} = \frac{g}{\sqrt{g^2 + g'^2}}, \tag{2.35}$$

$$\sin\theta_{\mathrm{W}} = \frac{g'}{\sqrt{g^2 + g'^2}}. \tag{2.36}$$

Equation 2.34 can be rewritten using the Weinberg mixing angle:

$$\frac{\nu^2\left(g^2 + g'^2\right)}{8}\left(\cos\theta_{\mathrm{W}} W_\mu^3 - \sin\theta_{\mathrm{W}} B_\mu\right)^2. \tag{2.37}$$

The physical field Z_μ can be immediately identified as:

$$Z_\mu = \cos\theta_{\mathrm{W}} W_\mu^3 - \sin\theta_{\mathrm{W}} B_\mu. \tag{2.38}$$

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

$$A_\mu = \sin\theta_{\mathrm{W}} W_\mu^3 + \cos\theta_{\mathrm{W}} B_\mu. \tag{2.39}$$

Equation 2.31 can be written in terms of the physical fields W_μ^+, W_μ^-, Z_μ, and A_μ:

$$\frac{(g\nu)^2}{4} W_\mu^+ W^{-\mu} + \frac{\left(g^2 + g'^2\right)\nu^2}{8} Z_\mu Z^\mu. \tag{2.40}$$

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:

$$m_{\mathrm{W}^+} = m_{\mathrm{W}^-} = \frac{g\nu}{2}, \quad m_{\mathrm{Z}} = \frac{\nu\sqrt{g^2 + g'^2}}{2} = \frac{m_{\mathrm{W}}}{\cos\left(\theta_{\mathrm{W}}\right)}. \tag{2.41}$$

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:

$$H(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \nu + h(x) \end{pmatrix}, \tag{2.42}$$

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 2.40:

$$m_W^2 \left(\frac{2h}{\nu} + \frac{h^2}{\nu^2} \right) W_{\mu}^{+} W^{-\mu} + \frac{m_Z^2}{2} \left(\frac{2h}{\nu} + \frac{h^2}{\nu^2} \right) Z_{\mu} Z^{\mu}. \tag{2.43}$$

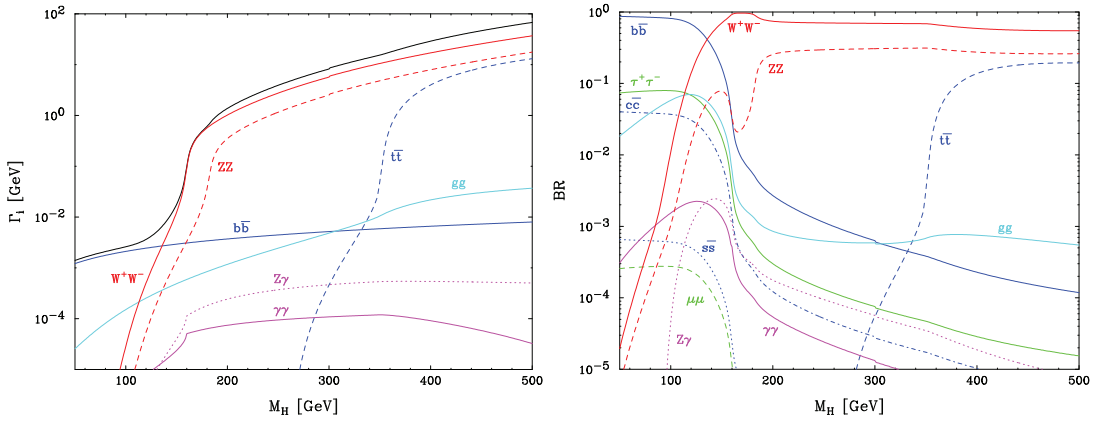
The Higgs boson self-interaction terms are obtained by replacing ν with $\nu + h(x)$ in the Higgs field potential in equation 2.27:

$$\frac{\mu^2}{2} (\nu + h)^2 - \frac{\lambda}{4} (\nu + h)^4 \supset -\lambda \nu^2 h^2 - \lambda \nu h^3 - \frac{\lambda}{4} h^4 \tag{2.44}$$

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 by ATLAS and CMS experiments to be 125.09 ± 0.24 GeV [3]. Because the Higgs boson is lighter than a pair of heavier particles such as $\mathrm{W^+W^-}$ or ZZ, the processes $\mathrm{H} \to \mathrm{W^+W^-}$ and $\mathrm{H} \to \mathrm{ZZ}$ are forbidden kinematically. However, in quantum field theory, such processes are allowed 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 [18].

2.7 Yukawa couplings of fermions

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:

$$\mathcal{L}_{Yukawa} = -\lambda^u \overline{q_L} \Phi^c_{\mathrm{H}} u_R - \lambda^d \overline{q_L} \Phi_{\mathrm{H}} d_R - \lambda^e \overline{l_L} \Phi_{\mathrm{H}} e_R + h.c., \tag{2.45}$$

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 associated with each fermion field; $\Phi^c_{\mathrm{H}} \equiv i\sigma^2 \mathrm{H}^*$ is a 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:

$$m_u = \frac{\lambda^u \nu}{\sqrt{2}}, \; m_d = \frac{\lambda^d \nu}{\sqrt{2}}, \; m_e = \frac{\lambda^e \nu}{\sqrt{2}}. \tag{2.46}$$

Thus the masses of fermions and bosons in SM are generated after the spontaneous symmetry breaking of the Higgs field.

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 a new strongly-interacting sector at the TeV scale. If the composite Higgs is a pseudo Nambu-Goldstone boson from spontaneous global symmetry breaking, the Higgs can be naturally light [12]. 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 double Higgs production via vector boson fusion at the TeV scale [19–21]. 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 [20]. 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 [22].

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 2.44. The coupling g_{WWHH} is associated with the terms in h^2 in Higgs interaction with other fields in equation 2.43. 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.

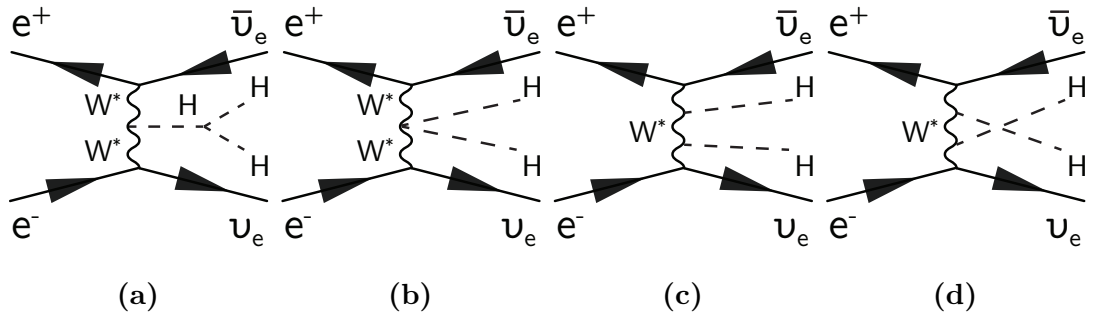


Figure 2.2: The main Feynman diagrams for the leading-order $e^+e^- \to HH\nu_e\overline{\nu}_e$ processes.

Following the assumption made in [20, 21] that mass scale at which new states appear is large, $m_\rho \gg m_h$, 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 [21]. In this

description, after the electroweak symmetry breaking, the Lagrangian is given by:

$$\mathcal{L} = \frac{1}{2}(\partial_\mu h)^2 - V(h) + \left(m_W^2 W_\mu^+ W^{-\mu} + \frac{m_Z^2}{2} Z_\mu Z^\mu\right)\left[1 + 2a\frac{h}{\nu} + b\frac{h^2}{\nu^2} + \dots\right], \quad (2.47)$$

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

$$V(h) = \frac{1}{2}m_h^2 h^2 + d_3\left(\frac{m_h^2}{2\nu}\right)h^3 + d_4\left(\frac{m_h^2}{8\nu^2}\right)h^4 + \dots, \quad (2.48)$$

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 strengths respectively. Comparing with the $\mathcal{L}_{Higgs}$ in the SM (see equation 2.43 and equation 2.44), it can be seen that $a = b = d_3 = d_4 = 1$ in the SM, and all higher order terms vanish. However, BSM Higgs models 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:

$$A = a^2(A_{SM} + A_1\delta_b + A_2\delta_{d_3}), \quad (2.49)$$

where A_{SM} is the SM amplitude and:

$$\delta_b \equiv 1 - \frac{b}{a^2}, \quad (2.50)$$

$$\delta_{d_3} \equiv 1 - \frac{d_3}{a}. \quad (2.51)$$

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

$$\sigma = a^4\sigma_{SM}\left(1 + A\delta_b + B\delta_{d_3} + C\delta_b\delta_{d_3} + D\delta_b^2 + E\delta_{d_3}^2\right), \quad (2.52)$$

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} [21]. 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; the detector effect will affect the distributions because of, for example, the loss of the reconstruction efficiency in the barrel/endcap overlap region.

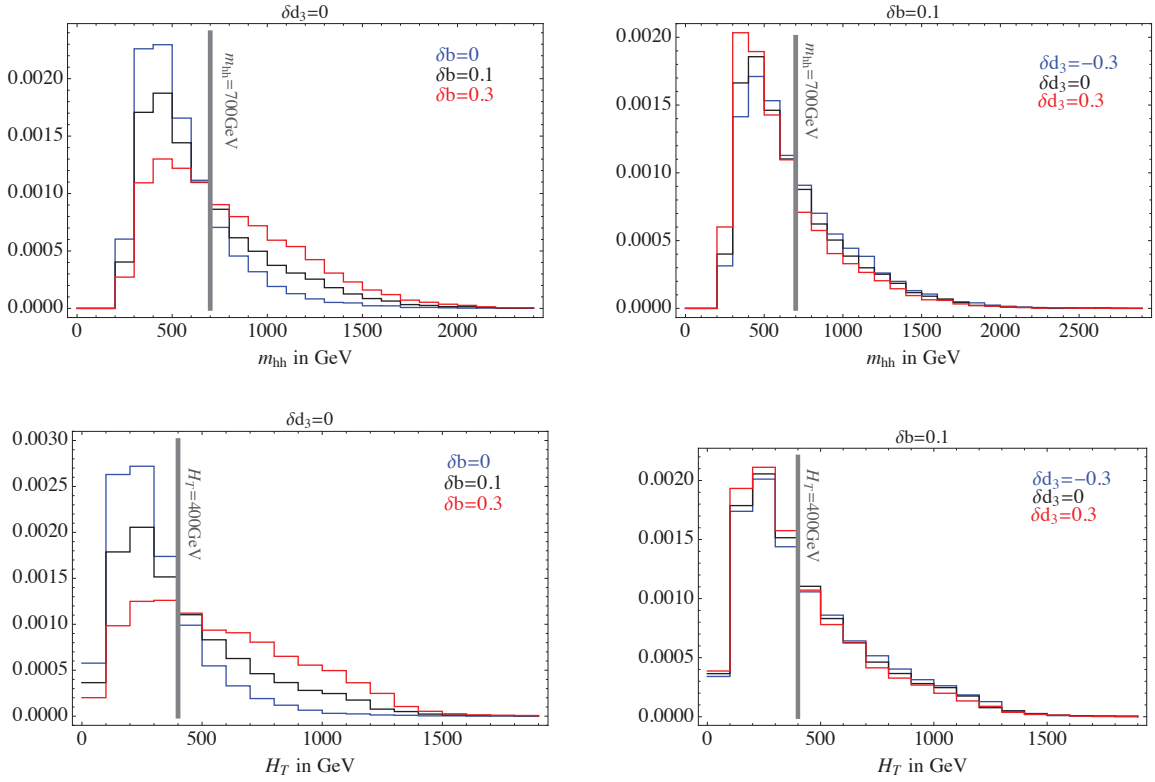


Figure 2.3: Normalised differential cross sections $d\sigma/dm_{hh}$ and $d\sigma/dH_T$ for $e^+e^- \to \nu\bar{\nu}hh$ process for CLIC at $\sqrt{s} = 3\,\text{TeV}$ after applying generator-level identification cuts, for several values of δ_b and δ_{d_3}. Figures are taken from [21].

In the expression of the cross section for the double Higgs production via $e^+e^- \to \nu\bar{\nu}hh$ in equation 2.52, 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 $e^+e^- \to \nu\bar{\nu}h$ channel is given by:

$$\sigma = \sigma_{SM}\Big(1 + A\Delta a + B\Delta a^2\Big), \tag{2.53}$$

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

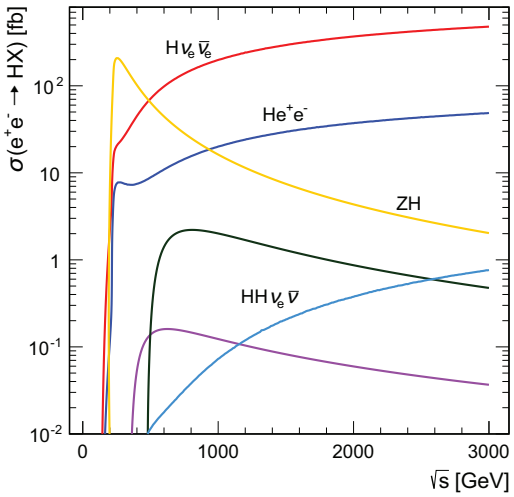


Figure 2.4: Cross sections as a function of centre-of-mass energy for Higgs production processes at an electron-positron collider for a Higgs mass of 126 GeV. The cross section values correspond to unpolarised beams and do not include the effect of beamstrahlung. The plot is taken from [23].

2.9 Tau pair polarisation correlations as a signature of Higgs boson

The advantage of the highly granular linear colliders can be demonstrate by studying the tau lepton decay products. The tau lepton has been studied extensively in the past at the Large Electron−Positron Collider (LEP) [24] and HERA [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 the tau lepton is massive. Therefore 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 ratios above 2% are listed in table 2.3.

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.

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

Table 2.3: Decay modes, final state particles, and branching ratios of the seven major τ^- decays, taken from [3].

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. Following the notation in reference [27], the $\tau^\pm \to \pi^\pm \nu_\tau$ decay mode is used as the example. The Higgs and Z boson decay to a tau pair where both tau leptons subsequently decay via $\tau^\pm \to \pi^\pm \nu_\tau$ can be represented as:

$$X \to \tau^+_\alpha \tau^-_\beta \to \pi^+\pi^- + \nu s, \tag{2.54}$$

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

$$z = \frac{E_{\pi^-}}{E_{\tau^-}}, \tag{2.55}$$

$$\overline{z} = \frac{E_{\pi^+}}{E_{\tau^+}}. \tag{2.56}$$

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

$$\frac{1}{\Gamma_\tau}\frac{d\Gamma}{dz} = Br\left(\tau^- \to \pi^-\nu_\tau\right) f\left(\tau^-_\alpha \to \pi^-; z\right), \tag{2.57}$$

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

$$f\left(\tau^-_\alpha \to \pi^-; z\right) = 1 + P_\alpha(2z - 1), \tag{2.58}$$

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

$$\frac{d^2N(X \to \tau^+\tau^- \to \pi^+\pi^- + \nu's)}{dz\, d\overline{z}} = \Big(Br\Big(\tau^- \to \pi^-\nu_\tau\Big)\Big)^2 \sum_{\alpha,\,\beta} C^X_{\alpha\beta} f\Big(\tau^-_\alpha \to \pi^-; z\Big) f\Big(\tau^+_\beta \to \pi^+; \overline{z}\Big), \tag{2.59}$$

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

$$C^{\mathrm{H}}_{LL} = C^{\mathrm{H}}_{RR} = \frac{1}{2}. \tag{2.60}$$

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

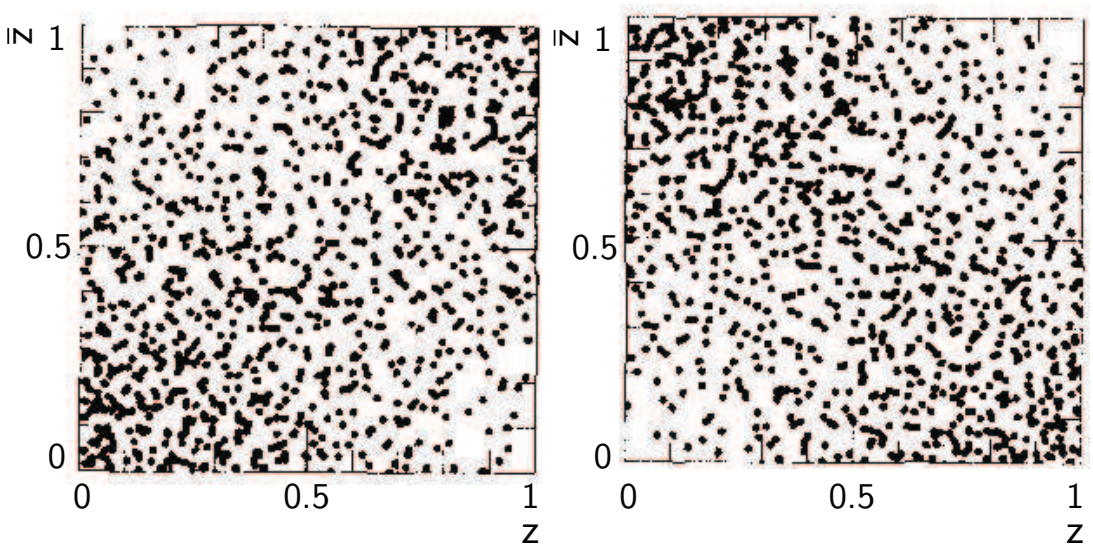
$$C^{\mathrm{Z}}_{LR} = \frac{1}{2}\Big(1 - P_\tau\Big),\ C^{\mathrm{Z}}_{RL} = \frac{1}{2}\Big(1 + P_\tau\Big), \tag{2.61}$$

where P_τ is the mean 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:

$$P_\tau = \frac{-2va}{v^2 + a^2}, \tag{2.62}$$

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



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

Figure 2.5: Two-dimensional distributions of $\bar{z} = E_{\pi^+}/E_{\tau^+}$ 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 $\tau^\pm \to \pi^\pm \nu_\tau$, 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 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 programmes at these colliders, 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 will have a centre-of-mass energy of 250 GeV. The second stage will 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. The ILC will be between 30 km and 50 km in length. The main parameters of the ILC machine are listed in table 3.1. Two detector concepts have been developed for the ILC: the International Large Detector (ILD) [29] and the Silicon Detector (SiD) [30]. Both ILD and SiD detectors are shown in figure 3.2.

	250 GeV	500 GeV
Collision rate	5 Hz	5 Hz
Electron linac rate	10 Hz	5 Hz
Number of bunches	1312	1312
Bunch population	2×10^{10}	2×10^{10}
Bunch separation	554 ns	554 ns
Pulse current	5.8 mA	5.8 mA
Main linac average gradient	14.7 MV m^{-1}	31.5 MV m^{-1}
Average total beam power	5.9 MW	10.5 MW
Estimated AC power	122 MW	163 MW

Table 3.1: ILC main parameters for 250 GeV and 500 GeV. The table is adapted from [1].

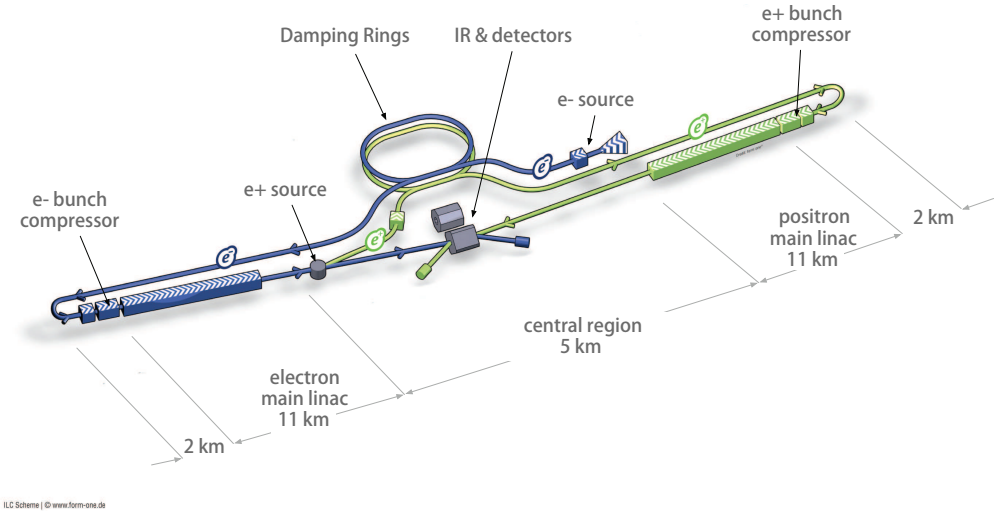
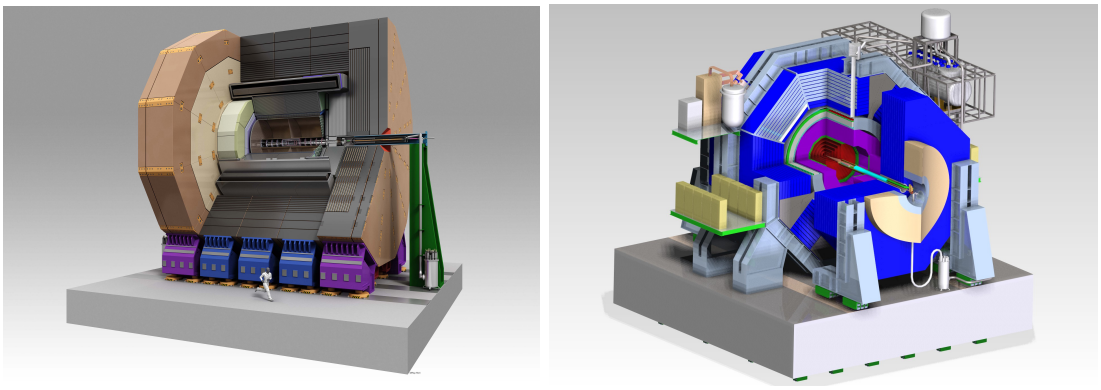


Figure 3.1: Schematic layout of the International Linear Collider, indicating all the major subsystems (not to scale), taken from [31].

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 main parameters of the CLIC machine are listed in table 3.2.

The two linear colliders use different technologies for accelerating electrons and positrons. The physics processes that can be studied are different due to different centre-



(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].

	500 GeV	3 TeV
Total site length	13.0 km	48.4 km
Loaded acceleration gradient	80 MV m^{-1}	100 MV m^{-1}
Main Linac RF frequency	12 GHz	12 GHz
Beam power / beam	4.9 MW	14 MW
Bunch separation	0.5 ns	0.5 ns
Bunch length	72 μm	44 μm
Beam pulse duration	177 ns	156 ns
Repetition rate	50 Hz	50 Hz

Table 3.2: CLIC main parameters for 500 GeV and 3 TeV. The table is adapted from [2].

of-mass energies that can be achieved by each one. Nevertheless the ILC and CLIC share some common features. Both colliders will be linear colliders as opposed 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 $\gamma\gamma \rightarrow$ hadrons background events per bunch on average, 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

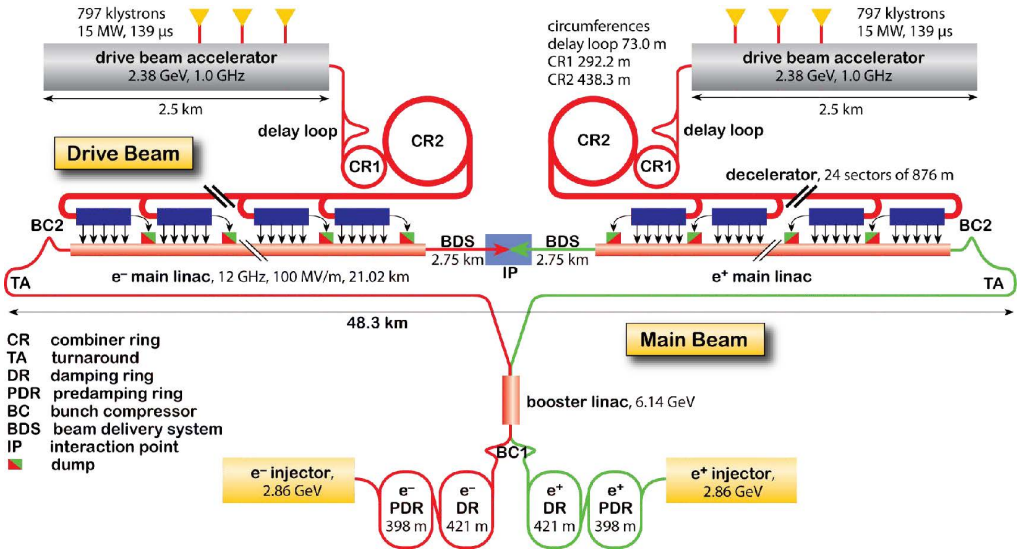


Figure 3.3: The layout of the Compact Linear Collider at a centre-of-mass of energy of 3 TeV, taken from [32].

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.

3.3 Physics at future linear colliders

An e^+e^- linear collider has advantages over a hadron collider such as the LHC. These advantages include:

1. 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.

2. Electroweak interactions in the e^+e^- collider will be democratic as the photon couples to all particles in and beyond the Standard Model equally [3–6]; 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. As the parton distributions fall sharply for a composite object like a proton [34], heavy particles such as protons have lower production rates than light ones.

3. In the LHC the calculation of cross sections depends on quantum chromodynamics and the proton structure function, which have larger systematic errors than QED predictions. At an e^+e^- collider the initial particles, e^+ and e^-, are point-like fermions interacting through electroweak forces only. Consequently, theoretical uncertainties are smaller.

4. The physics at the e^+e^- collider can be studied in detail. Without complicated underlying events, complete events can be reconstructed.

5. Polarised beams of electrons and positrons with known initial and final polarisation states could also 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 are the driving forces behind the detector design, share some common features. At a centre-of-mass energy of 250 GeV, the collider would operate as Higgs factories, allowing precise measurements of 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 colliders 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 [35].

3.4 Detector requirements

Many physics processes at future linear colliders 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 [36] allowed precise invariant mass reconstruction. At future linear colliders, reconstructing the invariant mass of multiple jets for final states with missing momentum will rely heavily on the intrinsic jet energy resolution of the detector.

One of the main objectives of future linear colliders is to be able to separate W and Z bosons by reconstructing their invariant masses via quark-jets using the hadronic decay channel. The idealised reconstructed W and Z boson mass 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. $\sigma_m/m \approx \Gamma_W/m_W \approx \Gamma_Z/m_Z$, a separation of 2.5σ in the

mass distributions implies a jet energy resolution of 3.5% [37] for a range of jet energies from 50 GeV to 1 TeV.

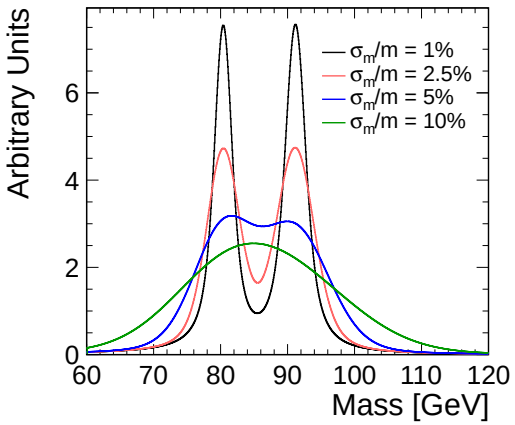


Figure 3.4: Ideal W/Z boson mass separation for different jet mass resolutions obtained using a Gaussian smearing of Breit−Wigner distribution, taken from [2].

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

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

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 [38, 39]. 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 3.3, assuming 30% of the jet energy (photon energy) is measured with $\sigma_E/E = 15\%/\sqrt{E(\text{GeV})}$, and 10% of the jet energy (hadron energy) is measured with $\sigma_E/E = 55\%/\sqrt{E(\text{GeV})}$ [31], a jet energy resolution of $\sigma_E/E = 19\%/\sqrt{E(\text{GeV})}$ can be obtained. This satisfies the jet energy resolution requirement for separating W and Z bosons via their hadronic decays, which requires a jet energy resolution of 3.5% or better. In reality, this level of performance is unattainable due to incorrect association of energy deposits to particles. Results from imperfect reconstruction rather than the intrinsic detector performance limit the performance of particle flow calorimetry [37] at jet energies beyond tens of GeV.

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

Table 3.3: 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 [37].

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 concepts using the PandoraPFA [37, 40] 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. Both are 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 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. From the interaction point (IP) outwards the design includes: 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 3.4. 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 [41] used for the ILD technical design report [31].

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 3.4: 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].

3.6.1 Vertex detector

The pixel vertex detector (VTX) needs to be close to the interaction point to reconstruct secondary vertices. Since the TPC is the main tracking detector, the VTX mainly

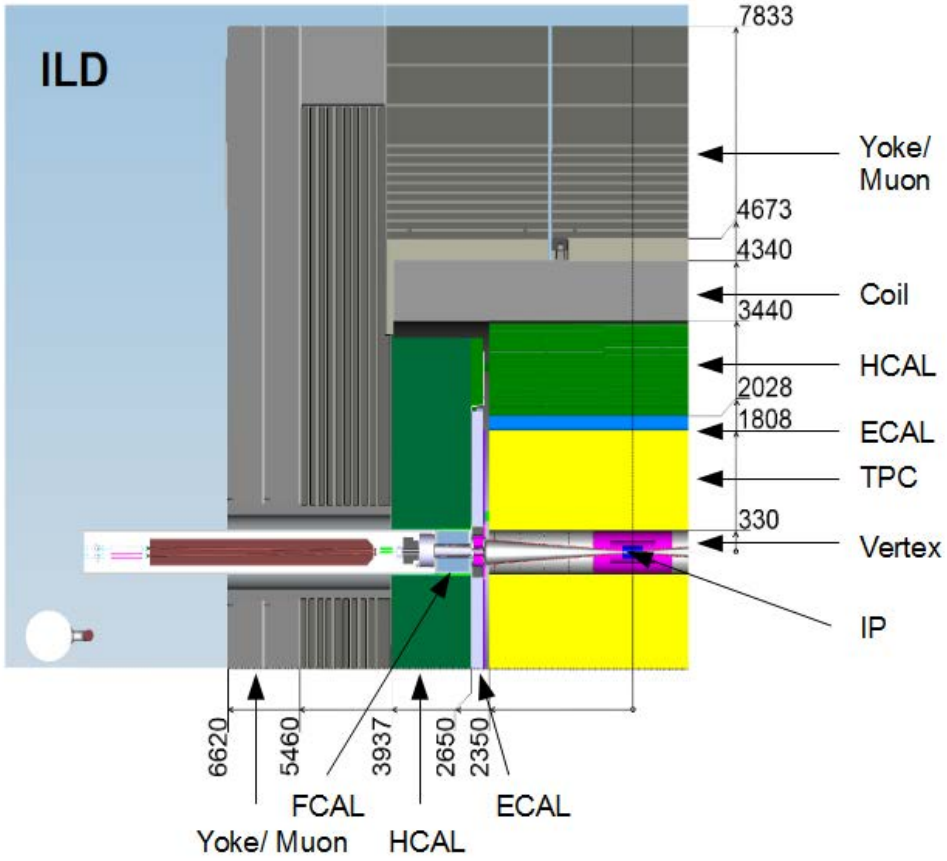


Figure 3.5: The longitudinal cross section of top quadrant of the ILD, taken from [31]. From the interaction point (IP) outwards, the design includes: a tracking system comprising a large time projection chamber (TPC) augmented with silicon tungsten layers; highly granular electromagnetic calorimeters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

measures the impact parameter of tracks. Figure 3.6 shows the structure of the VTX detector. The structure is of three almost 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 3.5. The radii covered by the detector range from 16 mm to 60 mm.

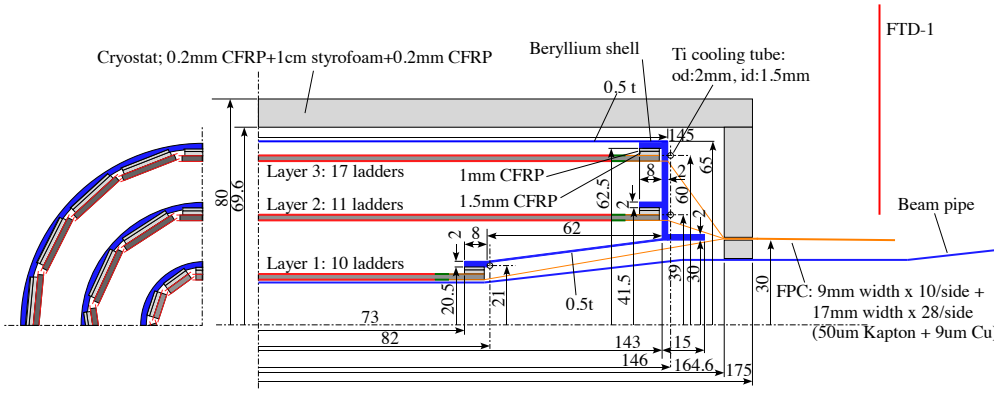


Figure 3.6: Structure of the ILD vertex detector, taken from [31].

	R	$\lvert z\rvert$	$\lvert\cos(\theta)\rvert$
Layer 1	16 mm	62.5 mm	0.97
Layer 2	18 mm	62.5 mm	0.96
Layer 3	37 mm	125 mm	0.96
Layer 4	39 mm	125 mm	0.95
Layer 5	58 mm	125 mm	0.91
Layer 6	60 mm	125 mm	0.90

Table 3.5: Key parameters of vertex detector in the ILC. The table is adapted from [31].

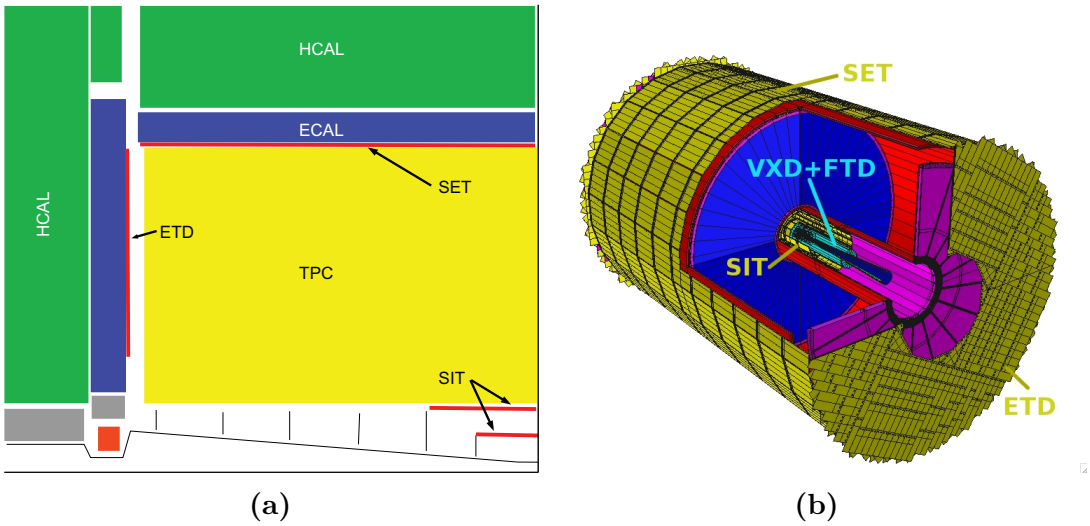
3.6.2 Tracking detectors

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 endcap 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 3.6.

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 3.6: 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

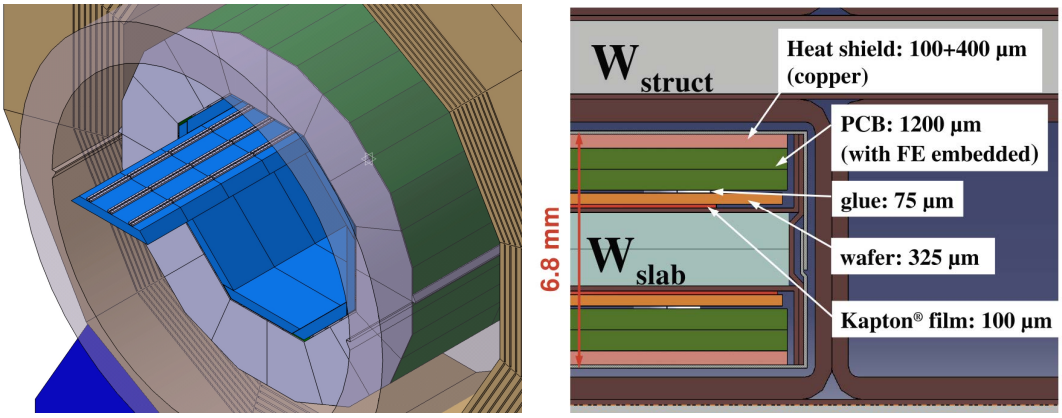
The silicon−tungsten sampling electromagnetic calorimeters in the ILD consist of an octagonal barrel and two endcap systems. The fine granularity ECAL is located inside the HCAL. Both ECAL and HCAL are inside the superconducting solenoid. Figure 3.8a shows the position of the electromagnetic calorimeter 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, since approximately 50% of hadronic showers start in the ECAL.

Test beam data and simulation studies [42–44] show that a sampling calorimeter with a longitudinal segmentation below one radiation length and the transverse segmentation 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, while 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 inner 20 layers use 2.1 mm thick absorber plates and the outer 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 [31].



(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].

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. The layout of a technological design, the "EUDET prototype" [45] is shown in figure 3.9.

Stainless steel is chosen for the absorber material for mechanical and calorimetric reasons. Steel allows for 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. Longitudinally the HCAL and ECAL, provide about six interaction lengths, which is sufficient to contain the hadronic showers.

The scintillator tiles provide both energy and position measurements. The transverse segmentation, dictated by previous optimisation studies [37], is $3 \times 3\,\mathrm{cm}^2$. This level of segmentation is sufficient to meet the requirement of the particle flow paradigm [31].

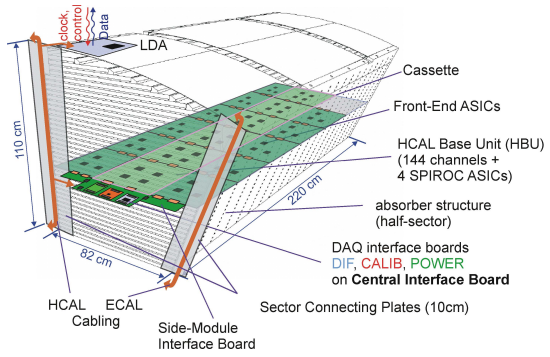


Figure 3.9: The schematic view of a CALICE analogue HCAL technological prototype module, taken from [31].

3.6.5 Solenoid, yoke, and muon system

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

The iron yoke returns the magnetic flux. The yoke is designed to ensure safety: the magnetic field at 15 m radial distance from the detector is fewer than 50 Gauss [46].

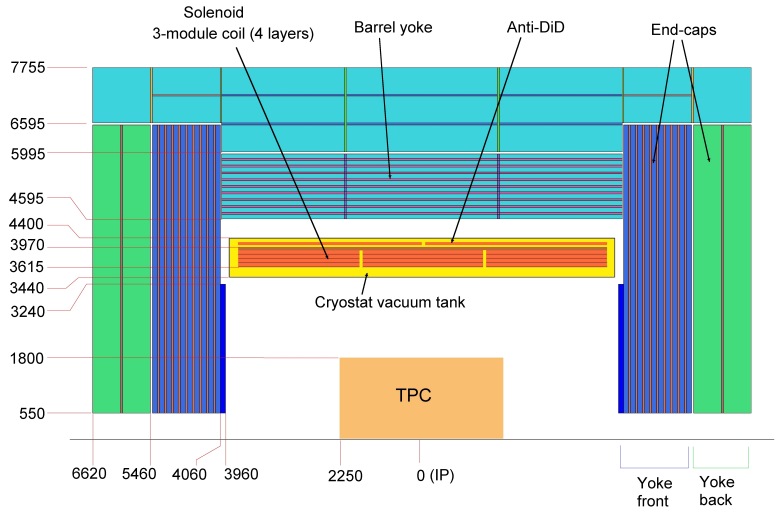


Figure 3.10: The ILD magnet cross section. Dimensions are in mm. Figure is taken from [31].

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

The first layer of the muon detector, which also acts 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 [47].

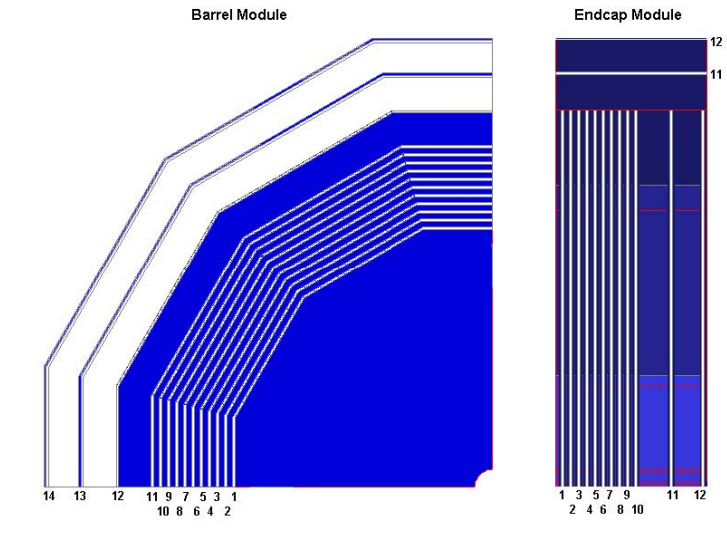


Figure 3.11: Sensitive layers of the ILD muon system, taken from [31].

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 event to measure the luminosity to a precision of 10^{-3} at a centre-of-mass energy of 500 GeV [48]. The beam 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 provide enough information for high-energy electron tagging [49], which aids event reconstruction at a high centre-of-mass energy. Table 3.7 lists the key parameters of the LumiCAL and the BeamCAL in the ILD.

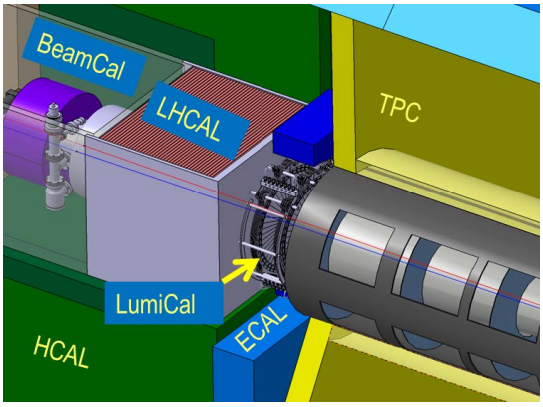


Figure 3.12: The calorimeters in the forward region of the ILD, taken from [31]. The LumiCAL, the BeamCAL, and the LHCAL are the luminosity calorimeter, the beam calorimeter, and the forward hadronic calorimeter, respectively.

3.7 Detector optimisation

Detector optimisation studies were performed to select the optimal parameters of the ILD sub-detectors [31]. Here, the optimisation studies for the ECAL and the HCAL are presented driven by PFA.

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

3.7.1 Electromagnetic calorimeter optimisation

The typical metric used for PFA optimisation is the jet energy resolution - defined as the root-mean-square divided by the mean for the smallest width of distribution that contains 90% of entries using $e^+e^- \to Z'Z'$ events 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.

Figure 3.13 shows the jet energy resolution for a single jet as a function of the number of longitudinal layers 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 where the number of layers is fewer than 20, although the impact is smaller for high energy jets. Therefore, 30 longitudinal layers is chosen for the ECAL.

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 $\sigma_E/E < 3.8\%$ for the jet energies relevant at $\sqrt{s} = 1\,\mathrm{TeV}$, with $5 \times 5\,\mathrm{mm}^2$ cell size being preferable.

3.7.2 Hadronic calorimeter optimisation

The jet energy resolutions as a function of HCAL scintillator square cell sizes for four different jet energies are shown in figure 3.15. There is no substantial gain in the jet

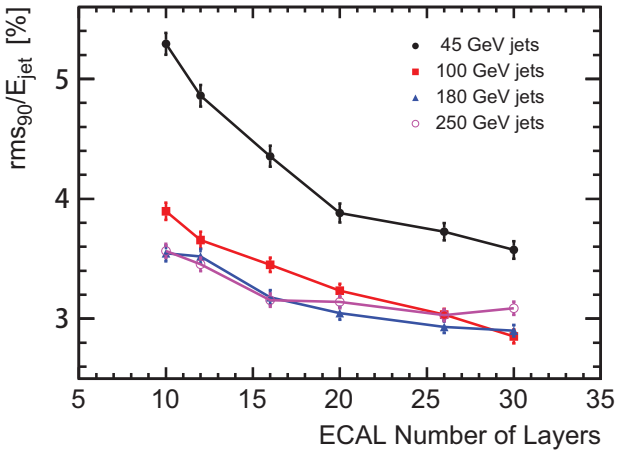


Figure 3.13: The single jet energy resolution as a function of the number of longitudinal ECAL layers, adapted from [31].

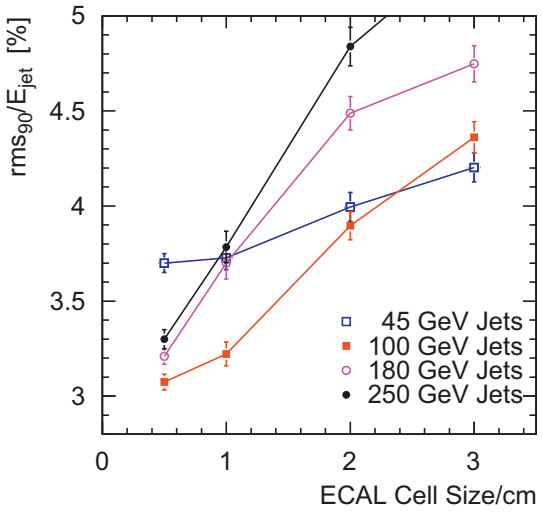


Figure 3.14: The single jet energy resolution as a function of the ECAL transverse cell sizes, adapted from [37].

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.

3.8 CLIC_ILD detector concepts

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

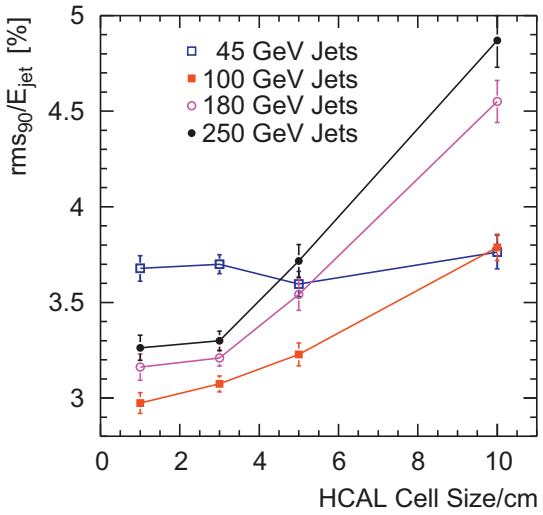


Figure 3.15: The single jet energy resolution as a function of the hadronic calorimeter scintillator cell sizes, adapted from [31].

CLIC_ILD detector. A comparison of key parameters of the ILD and the CLIC_ILD detector concepts is shown in table 3.4.

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.

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 $12° \lesssim \theta \lesssim 168°$

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. A denser material, in this case tungsten, is selected as the absorber material in the HCAL barrel to sustain the same inner bore radius as the ILD detector solenoid.

The magnetic field is increased to 4 T for a better jet energy resolution [37] at a higher centre-of-mass energy. Due to the different magnetic field strength the iron yoke thickness is increased to 230 cm.

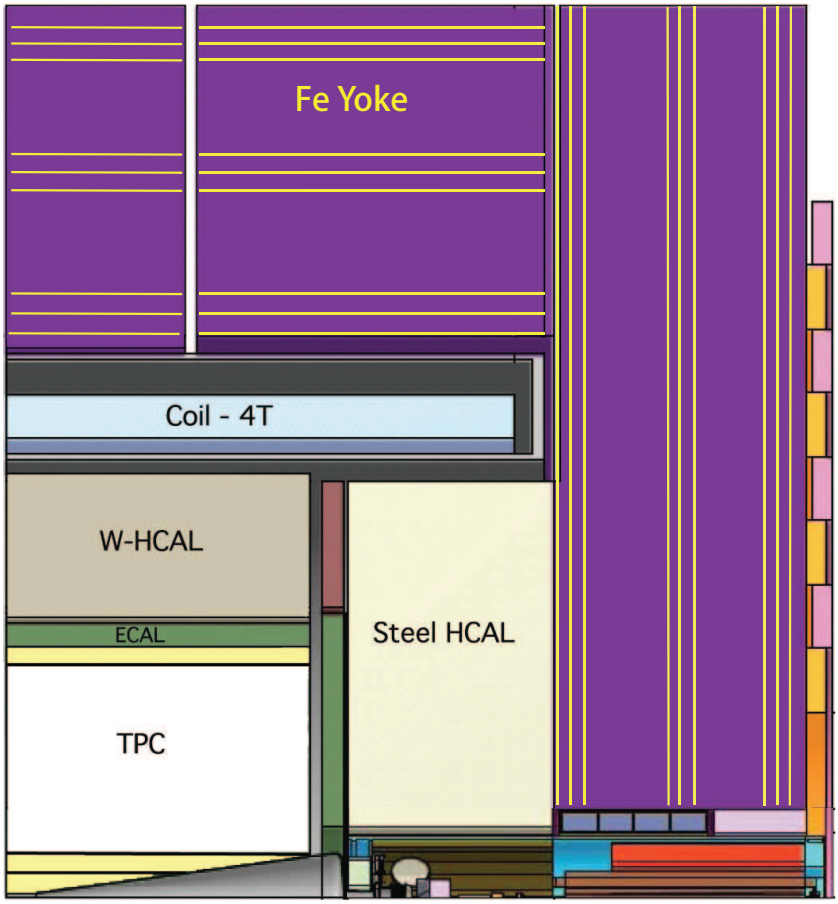


Figure 3.16: The longitudinal cross section of top quadrant of the CLIC_ILD detector concept, taken from [2]. From interaction point (IP) outwards, the design includes: a tracking system comprising a large time projection chamber (TPC) augmented with silicon tungsten layers; highly granular electromagnetic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

The CLIC_ILD detector model 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 of the LumiCAL and the BeamCAL at the ILD and the CLIC_ILD is shown in table 3.7.

Chapter 4

Event Generation, Simulation, Reconstruction, and Analysis

'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 is discussed. Reconstruction of particle flow objects with PandoraPFA is presented, which is the framework used for the photon reconstruction algorithms developed in chapter 5. The Boosted Decision Tree multivariate analysis technique, which is used in a number of places in this thesis, is described in detail.

4.1 Event generation

Most of the Monte Carlo (MC) samples used in this thesis were generated using Whizard generator software [50, 51]. Initial State Radiation (ISR) is simulated in Whizard with the ISR photons collinear with the beam direction. Events were generated with head-on e^+e^- collisions. Pythia [52] was used to describe parton showering, hadronisation and fragmentation. The fragmentation parameters of Pythia were tuned to OPAL data [53] from the Large Electron-Positron Collider (LEP). The Final State Radiation (FSR) is treated by Pythia with its default parameters. Tauola [54] was used to describe the tau lepton decay with correct spin correlations of the tau decay products.

4.1.1 CLIC luminosity spectrum

The small CLIC beam size at the interaction point, which implies a high bunch charge density, results in electrons and positrons radiating strongly in the electromagnetic field of the other beam. This effect is known as Beamstrahlung. Consequently, the centre-of-mass energies of the actual e^+e^- collisions have a long low energy tail towards lower values than the nominal centre-of-mass energy. The luminosity spectrum for CLIC operating at $\sqrt{s} = 3\,\text{TeV}$ is shown in figure 4.1, generated with GuineaPig [55]. Only 35% of the effective luminosity falls within 1% of the nominal centre-of-mass energy.

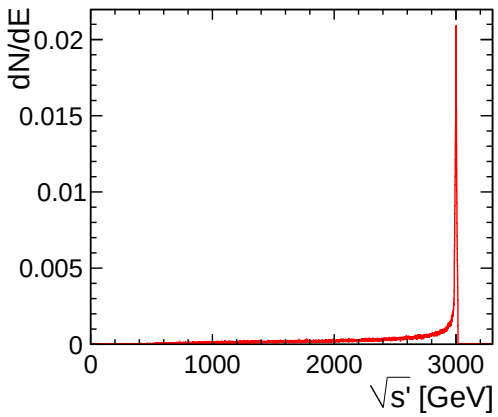


Figure 4.1: The luminosity spectrum for CLIC operating at $\sqrt{s} = 3\,\text{TeV}$, taken from [2].

Due to the presence of a large amount of Beamstrahlung photons, the photon−photon and electron/positron−photon background must be accounted for. The instantaneous luminosities for the electron−positron, electron/positron−photon, and photon−photon interactions at a $\sqrt{s} = 3\,\text{TeV}$ CLIC machine are listed in table 4.1, obtained from GuineaPig.

Instantaneous luminosity / $10^{34}\text{cm}^{-2}\text{s}^{-1}$	$\sqrt{s} = 3\,\text{TeV}$
e^+e^-	6.7
$e^+\gamma$	5.3
$e^-\gamma$	5.3
$\gamma\gamma$	4.6

Table 4.1: Instantaneous luminosity for the electron−positron, electron/positron−photon, and photon−photon interactions at a $\sqrt{s} = 3\,\text{TeV}$ CLIC machine. Values are taken from [56].

It is known that the GuineaPig prediction overestimates the luminosities for interactions that involve Beamstrahlung photons [56]. Hence, luminosity corrections are applied to the interactions that involves Beamstrahlung photons according to table 4.2.

Luminosity correction	$\sqrt{s} = 1.4\,\text{TeV}$	$\sqrt{s} = 3\,\text{TeV}$
$L(e^+\gamma)$ / $L(e^+e^-)$	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: Correction to integrated luminosities of the positron−photon, electron−photon, and photon−photon interactions where photons are from the Beamstrahlung process for CLIC at $\sqrt{s} = 1.4\,\text{TeV}$ and 3 TeV. The corrections, normalised to the electron−positron luminosity, are taken from [56].

4.2 Event simulation

Studies presented in this thesis are based on fully simulated event samples, reconstructed in ILD or CLIC_ILD detector models. Geant4 [57] is used to simulate the interactions of particles through the detector material. The ILD and CLIC_ILD detector geometries are implemented in the Mokka package [41]. The QGSP_BERT physics list from Geant4 is used to simulate the detailed development of hadronic showers in the detector. The beam crossing angle (20 mrad for CLIC and 14 mrad for the ILC) is introduced by applying a corresponding Lorentz boost to all generated particles in the events prior to detector simulation.

4.2.1 CLIC beam induced background

At CLIC, pile-up from beam induced background needs to be considered. The two most significant types of background at CLIC are $\gamma\gamma \to$ hadrons and incoherent e^+e^- pairs [2]. The $\gamma\gamma \to$ hadrons background is produced when the interaction of real and virtual photons from the colliding beams leads to two-photon interactions, resulting in hadronic final states [58, 59]. 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 field of the other beam [60].

Simulation of $\gamma\gamma \to$ hadrons uses the photon spectrum from GuineaPig and a parametrisation of the total cross section of the $\gamma\gamma \to$ hadrons process [61]:

$$\sigma_{\gamma\gamma}\left(s_{\gamma\gamma}\right) = 211\,\text{nb}\left(\frac{s_{\gamma\gamma}}{\text{GeV}^2}\right)^{0.0808} + 215\,\text{nb}\left(\frac{s_{\gamma\gamma}}{\text{GeV}^2}\right)^{-0.4525}, \tag{4.1}$$

where $s_{\gamma\gamma}$ is the Mandelstam s variable for the two photons. On average, there are 3.2 $\gamma\gamma \to$ hadrons events per bunch crossing within the detector acceptance at $\sqrt{s} = 3\,\text{TeV}$ with a $\gamma\gamma$ centre-of-mass energy greater than 2 GeV [62]. Pythia is used to simulate the hard interaction and the hadronisation of these $\gamma\gamma \to$ hadrons events. Table 4.3 shows the average energy deposited from $\gamma\gamma \to$ hadrons and the incoherent e^+e^- pairs in different parts of the CLIC_ILD detector. The energies in the calorimeter are integrated over the 300 ns from the start of the bunch train, which corresponds to 60 bunch crossings. The incoherent e^+e^- pairs are the dominant background in the HCAL endcaps due to the interactions of the large incoherent e^+e^- pairs in the BeamCAL, resulting in low-energy neutrons depositing energies in the HCAL endcaps. Except in the HCAL endcaps, the $\gamma\gamma \to$ hadrons is the dominant background in all calorimeters.

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: Average energy deposited from $\gamma\gamma \to$ hadrons and the incoherent e^+e^- pairs in different parts of the CLIC_ILD subdetectors. Numbers correspond to the background for an entire CLIC bunch train. The reconstructed calorimeter energies are integrated over 300 ns from the start of the bunch train. The table is adapted from [2].

For the study presented in chapter 8, only the $\gamma\gamma \to$ hadrons background is included in the simulation. The hits from simulated $\gamma\gamma \to$ hadrons events from 60 bunch crossings are superimposed on simulated e^+e^-, $e^\pm\gamma$, and $\gamma\gamma$ collisions before the event reconstruction. The included background corresponds to an integration time window of the CLIC_ILD detector of -5 ns to $+25$ ns around the generated physics event. Bunch trains are separated by 0.5 ns to mimic the CLIC train structure [2]. For each bunch crossing the

number of $\gamma\gamma \to$ hadrons events included is drawn from a Poisson distribution with a mean of 3.2.

4.3 Event reconstruction

The linear collider reconstruction software runs in the MARLIN framework [63]. The event reconstruction consists of three main steps: digitisation of simulated tracks and calorimeter hits, reconstruction of tracks in the tracking system [64], and reconstruction of particle flow objects (PFOs) with PandoraPFA [37, 40].

Several different MARLIN processors are used to reconstruct tracks: CLUPATRA [64] is used to reconstruct tracks in the TPC; FORWARDTRACKING [64] is used to reconstruct tracks in the FTD; and SILICONTRACKING [64] is used to reconstruct tracks in other silicon tracking detectors. A final MARLIN tracking processor, FULLLDCTRACKING [64], is used to combine tracks segments produced from individual processors.

4.3.1 PandoraPFA

PandoraPFA was developed for the ILD detector concept [37] at the ILC and at CLIC [2,40]. It has also been used for the SiD detector concept [30] and the studies for the CMS endcap calorimeter upgrade [65]. The latest improvement to PandoraPFA, including a better internal memory management, is summarised in [66].

PandoraPFA adopts a multi-algorithm approach to the particle flow object reconstruction. There are over sixty e^+e^- linear collider specific reconstruction algorithms. Each algorithm aims to address a particular topological issue in the reconstruction. There are nine main steps in the PandoraPFA reconstruction, listed below:

- Track processing: identifies special topologies, V^0s and kinks, of tracks and prepares them for subsequent reconstruction;
- Calorimeter hit processing: calculates the properties of a calorimeter hit, such as its position and its energy response from the calorimeter digitiser. The calorimeter hits are used in subsequent reconstruction;
- Particle identification algorithms find calorimeter hits associated with neutral particles, such as photons. Chapter 5 describes photon reconstruction algorithms in detail;

- An initial clustering step group calorimeter hits into clusters. This step is discussed further in section 4.3.1, since it is used in photon reconstruction algorithms described in Chapter 5;
- Topological cluster association merges clusters based on clear topological signatures. 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.
- The track−cluster matching step associates clusters to tracks obtained from the tracking detectors;
- The re-clustering step improves the compatibility of the cluster energy and the associated track momentum on a statistical basis. This step is important for events with a dense jet environment and jets above 50 GeV;
- The fragment removal step focuses on merging clusters that are likely to be fragments of other particles. Algorithms for photon fragment merging are described in chapter 5;
- Particle Flow Object creation is the last step of the PandoraPFA reconstruction. It creates the output objects, Particle Flow Objects (PFOs). The reconstructed particles, PFOs, are the basis of all subsequent analyses.

Initial clustering

In PandoraPFA, cone-based clustering algorithms are used to group calorimeter hits into clusters, illustrated in figure 4.2. The cone-based clustering algorithm is used because the direction of the particle flow is largely unchanged from the original particle. The seed for the cone clustering can be a projection of a track onto the front of the ECAL, and the initial cone direction is the direction of the track projection. Alternatively, the seed can be a calorimeter hit and the initial cone direction is the direction from the IP to the calorimeter hit. A cone with a specified opening angle is then 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 the cone are made by considering calorimeter hits in previous layers and in the same layer. If a calorimeter hit is not associated with any cone the hit is used to seed a new cluster.

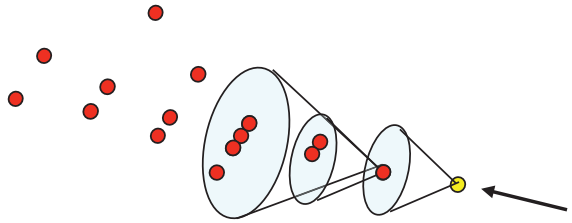


Figure 4.2: Illustration of the cone-based clustering algorithm used in PandoraPFA, taken from [67]

4.3.2 CLIC beam induced background suppression

The analysis in chapter 8 is performed for the CLIC_ILD detector, where the CLIC beam induce background is considered. The output of PandoraPFA is a list of reconstructed particles. Two packages have been developed to suppress the background from pile up of $\gamma\gamma \to$ hadrons: TrackSelector and PfoSelector [40].

The TrackSelector [40] package removes poor quality and fake tracks that are likely to be from the beam induced background. It examines the number of track hits in individual tracking subdetectors and imposes track-quality cuts. 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 using the helical fit of the track trajectory differs more than 50 ns from using a straight line fit, the track will be rejected.

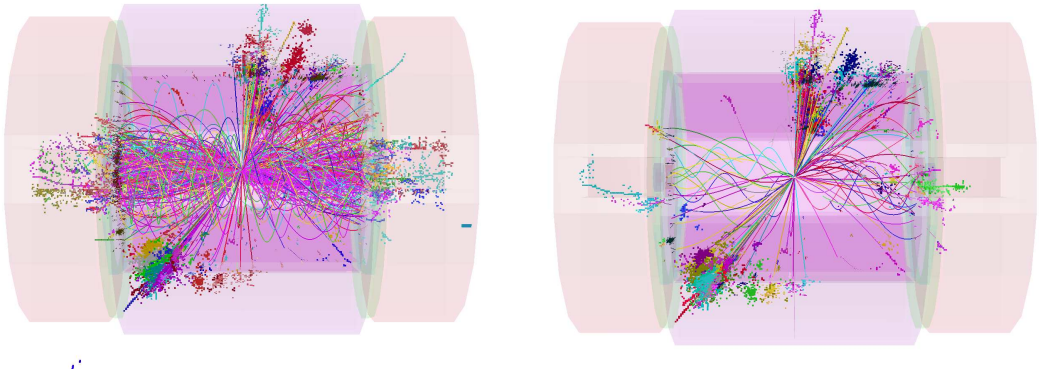
The PfoSelector [40] package discards reconstructed particles that are likely to be from the beam induced background from the event reconstruction based on the transverse momentum (p_T) and time information of the reconstructed particles. The reconstructed particles from $\gamma\gamma \to$ hadrons often have low p_T and are distributed in time across the 10 ns 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 bunch crossing that contains the event. The time of the bunch crossing containing the event is determined by the high-level software trigger.

PfoSelector uses different p_T and timing cuts for the central part and the forward part of the detector. PfoSelector also uses different p_T and timing cuts for different types of particles: photons, neutral PFOs, and charged PFOs.

Three configurations of these cuts were 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.

Figure 4.3 shows the effect of the background suppression with the tight PFO selection. Figure 4.3a shows reconstructed particles in a simulated $e^+e^- \to HH \to t\bar{b}b\bar{t}$ event in the CLIC_ILD detector model assuming an integration time window of 10 ns (100 ns in HCAL barrel), with 60 bunch crossings of $\gamma\gamma \to$ hadrons background overlaid. 60 bunch crossings correspond to an integration time window of 30 ns, which is sufficient to account for main effect of the background. The effect of applying tight PFO selection cuts is shown in figure 4.3b. The energy deposited in the detector by the background is reduced from 1.2 TeV to the level of 100 GeV.



(a) **(b)**

Figure 4.3: Reconstructed particles in a simulated $e^+e^- \to HH \to t\bar{b}b\bar{t}$ event, integrated over a time window of 10 ns (100 ns in HCAL barrel) in the CLIC_ILD detector model, with 60 bunch crossings of $\gamma\gamma \to$ hadrons background overlaid in a). The effect of applying tight PFO section cuts is shown in b). 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 [40].

4.4 Analysis software

4.4.1 Monte Carlo truth linker

For the purpose of algorithm development and event selection optimisation, it is important to be able to associate reconstructed particles to the Monte Carlo particles. The MC truth linker processor provides the link between an MC particle and a reconstructed calorimeter hit. From the link, the MC particle contributing the most to a reconstructed particle can be determined based on highest sum of the energies of calorimeter hits with

the link to the same MC particle. Similarly, the MC particle contributing the most to a group of reconstructed particles (a jet) can be determined.

4.4.2 Jet algorithms

Jets result from the hadronisation process from high energy quarks or gluons. A jet is typically a visually obvious structure in an event display. The momentum and the direction of a jet are largely the same as 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 [68], and descriptions on later developments can be found in reviews [69–71].

There are two large families of jet finding algorithms: cone based algorithms and sequential combination algorithms. 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 is updated after each combination. This procedure is repeated until stopping criteria are satisfied. Various jet algorithms typically differ in the definitions of distance metrics and stopping criteria.

The jet algorithm implementation used in this thesis is the FastJet C++ software package [72, 73]. The notations in the subsequent discussion follow the convention in [72].

Longitudinally invariant k_t algorithm

The longitudinally invariant k_t algorithm [74, 75] 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

$$d_{ij} = d_{ji} = \min\left(p_{\mathrm{T}i}^2,\, p_{\mathrm{T}j}^2\right)\frac{\Delta R_{ij}^2}{R^2}, \tag{4.2}$$

$$d_{iB} = p_{\mathrm{T}i}^2, \tag{4.3}$$

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

$\Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$. $y_i = \frac{1}{2}\ln\frac{E_i+p_{z\,i}}{E_i-p_{z\,i}}$ and ϕ_i are the rapidity and azimuthal angle of particle i respectively. The free parameter R controls the jet radius.

The distance metric and the beam distance are calculated for all pairs of particles. If the minimum value of distance metrics and beam distances of all pairs of particles is a distance metric, indicated as 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 of distance metrics and beam distances of all pairs of particles is a beam distance, indicated as d_{iB}, particle i becomes an output jet and is 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 of distance metrics and beam distances of all pairs of particles is a beam distance, indicated as d_{iB}, particle i forms part of the beam jet. The beam jet is discarded at the end of the jet clustering. The second difference is that when distance metrics and beam distances of all pairs of particles are all 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.

Durham algorithm

The Durham algorithm [76], also known as e^+e^- k_t algorithm, is commonly used in e^+e^- collider experiments. It has one pair-wise distance metric:

$$d_{ij} = 2\ \min\left(E_i^2, E_j^2\right)(1 - \cos(\theta_{ij})), \tag{4.4}$$

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 in exclusive mode, which means that the clustering will stop when all distance metrics are above a threshold, d_{cut}.

Jet algorithms for CLIC

Although CLIC is an 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 environment. Figure 4.4 shows the reconstructed Z boson mass distribution for $ZZ \to q\bar{q}q'\bar{q}'$ at a

500 GeV CLIC [77] using the Durham jet algorithm and the longitudinally invariant k_t jet algorithm. The $\gamma\gamma \rightarrow$ hadrons background, corresponding to 300 bunch crossings, was overlaid on the event where each bunch crossing contains approximately 0.3 $\gamma\gamma \rightarrow$ hadrons events. The longitudinally invariant k_t algorithm gives a better reconstructed Z boson mass distribution. Other studies [2, 78] have also shown that jet algorithms for the pp colliders, such as the longitudinally invariant k_t algorithm, give better reconstructed mass and energies resolutions at CLIC. Therefore, the longitudinally invariant k_t algorithm is usually used in analyses at CLIC.

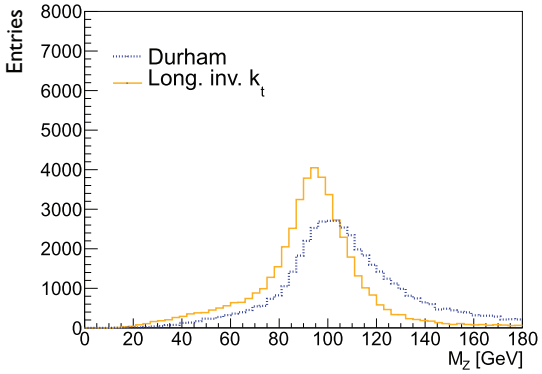


Figure 4.4: The reconstructed Z boson mass distribution for $ZZ \rightarrow q\bar{q}q'\bar{q}'$ at a 500 GeV CLIC, using the Durham jet algorithm and the longitudinally invariant k_t jet algorithm. The $\gamma\gamma \rightarrow$ hadrons background corresponding to 300 bunch crossings was overlaid on the event, where each bunch crossing contains approximately 0.3 $\gamma\gamma \rightarrow$ hadrons events. The figure is adapted from [77].

The y parameter

The y parameter is a measure of the number of jets in an event. It describes the transition 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, and below which the jet algorithm returns 3 jets. Numerically the y parameter is often much smaller than one and it is usually quoted in terms of the negative logarithm of the number.

4.5 Multivariate analysis

Multivariate analysis (MVA) has become increasingly important in high energy physics. The implementation of the machine learning based MVA used in this thesis is provided by TMVA package [79, 80]. MVA can be used for classification or regression. Classification classifies an event into one of several classes. Regression of an event gives an output in a

continuous numerical range. In a typical physics analysis, MVA is often used to classify one type of event from other types.

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 discriminant variables as inputs, which have different distributions for the signal and the background. To use the MVA classification, the trained model will be applied to the testing data. The classification response is to classify a testing event into the signal class or the background class.

This two-class classification scheme can be easily extended to multiple classes, as implemented in TMVA with the MULTICLASS class. For example, The MULTICLASS approach is used in the tau decay mode classification in section 6.5 and in the flavour tagging classifier in section 8.5.

4.5.1 Optimisation and overfitting

Two important concepts with MVAs are optimisation and overfitting of a model. The optimisation of a model refers to selecting the optimal free parameters of the model. One could build a complex model which fits the training samples extremely well, but the model may not be optimal for a testing sample. A simple model is less prone to statistical fluctuations of samples. However, the model might be too simple to achieve the optimal model. The former case is known as overfitting or overtraining. The latter case is called underfitting or undertraining.

Overfitting occurs 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 4.5 shows the signal selection efficiency as a function of the model complexity using an example from the double Higgs analysis of chapter 8 and the Boosted Decision Tree model. The efficiency of the signal selection is defined as the fraction of the signal selected when the fraction of background 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 of fewer than seven to avoid overfitting.

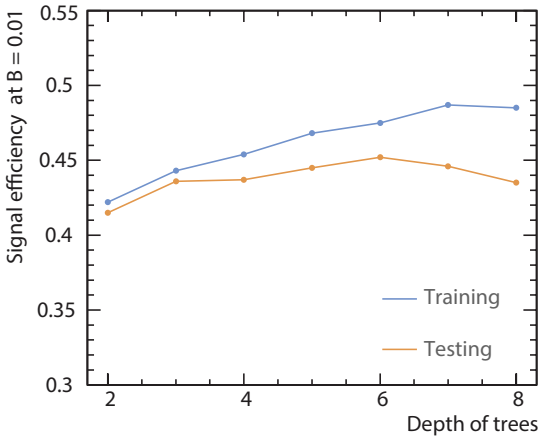


Figure 4.5: Example of the signal selection efficiency as a function of the model complexity. The example is chosen from the double Higgs analysis at $\sqrt{s} = 3\,\text{TeV}$, using the Boosted Decision Tree model. The efficiency of the signal selection is defined as the fraction of the signal selected when the fraction of background 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 used 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 neural 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 [81].

To rigourously identify the best model, individual optimisations of models are required, which is computationally very expensive. However, as researchers in the machine learning field suggest, the boosted decision tree is probably the best out-of-the-box machine learning model [81]. A neural network model could potentially perform better than the boosted decision tree model but it requires more tuning and is less intuitive to interpret. For these reasons the boost decision tree model (BDT) to conduct physics analyses in this thesis. Before describing the BDT model in detail some simpler models are discussed.

4.5.3 Rectangular Cut model

The rectangular cut method optimises cuts to maximise pre-defined metrics. The metric could be the signal efficiency that corresponds to 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 events passing the rectangular cuts.

4.5.4 Projective Likelihood model

The projective likelihood model with probability density estimators (PDE) is used in PandoraPFA for the photon ID test due to its simplicity and low requirement on computing resources. The PandoraPFA implementation of the projective likelihood model is discussed in section 5.3.3.

The likelihood classifier calculates the probability density for each discriminant variable, for the signal class and the background class. The overall signal and background likelihood are defined as products of the individual probability densities of each variable for the respective signal class and background class. The likelihood ratio, R, can then be defined as the signal likelihood divided by the sum of the signal likelihood and the background likelihood.

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 (splitting point) uses a cut on a single discriminant 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 sub-samples: signal-like and background-like sub-samples. This splitting process (tree growing) starts at the root node. For each sub-sample, the splitting process stops after certain criteria are met. The stopping criteria 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 signal than background events in an end node it is referred to as a signal-like end node. The opposite is referred to as a background-like end node.

The training of the decision tree refers to finding the optimal cut at each 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 application of a trained decision tree is performed by traversing the tree from the root node to the end node. The event is classified as signal or background, depending on whether it falls in a signal-like or background-like end node.

Figure 4.6 illustrates a simple example of a trained decision tree. PhD student is the signal class and undergraduate student is the background class. The top diamond box in figure 4.6, "Leave party before 1am", is the root node. All diamond boxes are splitting nodes. Rectangular 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 used to find the optimal cut at the splinting node is the Gini index.

The attributes of signal and background classes are listed table 4.4. In this example, there are ten PhD students who leave parties before 1 am and know where a free pizza is located. In contrast, five undergraduate students leave parties after 1 am and know where a free pizza is located. Three undergraduate students leave parties before 1 am and do not know where a free pizza is located. Two undergraduate students leave parties after 1 am and do not know where a free pizza is located.

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 4.6.

Details of finding the optimal cut at the root node are outlined to demonstrate the first step of the training of the model. There are two possible cuts for the root node, "Leave party 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. The Gini index gives

$$2p(1-p) \backsimeq 0.36. \tag{4.5}$$

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 a free pizza is located. The Gini index gives

$$2p(1-p) \backsimeq 0.44. \tag{4.6}$$

Therefore, by choosing the cut that minimises the Gini Index, the optimal cut for the root node is "Leave party before 1am".

The simple tree in figure 4.6 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.

Improving 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 (due to overfitting). Many methods have been developed to overcome the instability of the decision tree model. 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, is a method that selects a random subset of the training sample, and uses the subset in the training stage.
- Random Forest: when a tree is grown, a randomly selected subset of discriminant variables are used to grow the tree.

4.5.6 Boosted Decision Tree model

A Boosted Decision Tree (BDT) contains many decision trees where the boosting is used to grow trees. There are two common boosting methods: adaptive boosting and gradient boosting. Adaptive boosting, first introduced in [82], is discussed in further detail as it is simpler to understand than the gradient boosting. The adaptive boosting algorithm, adapted from [81], is outlined below:

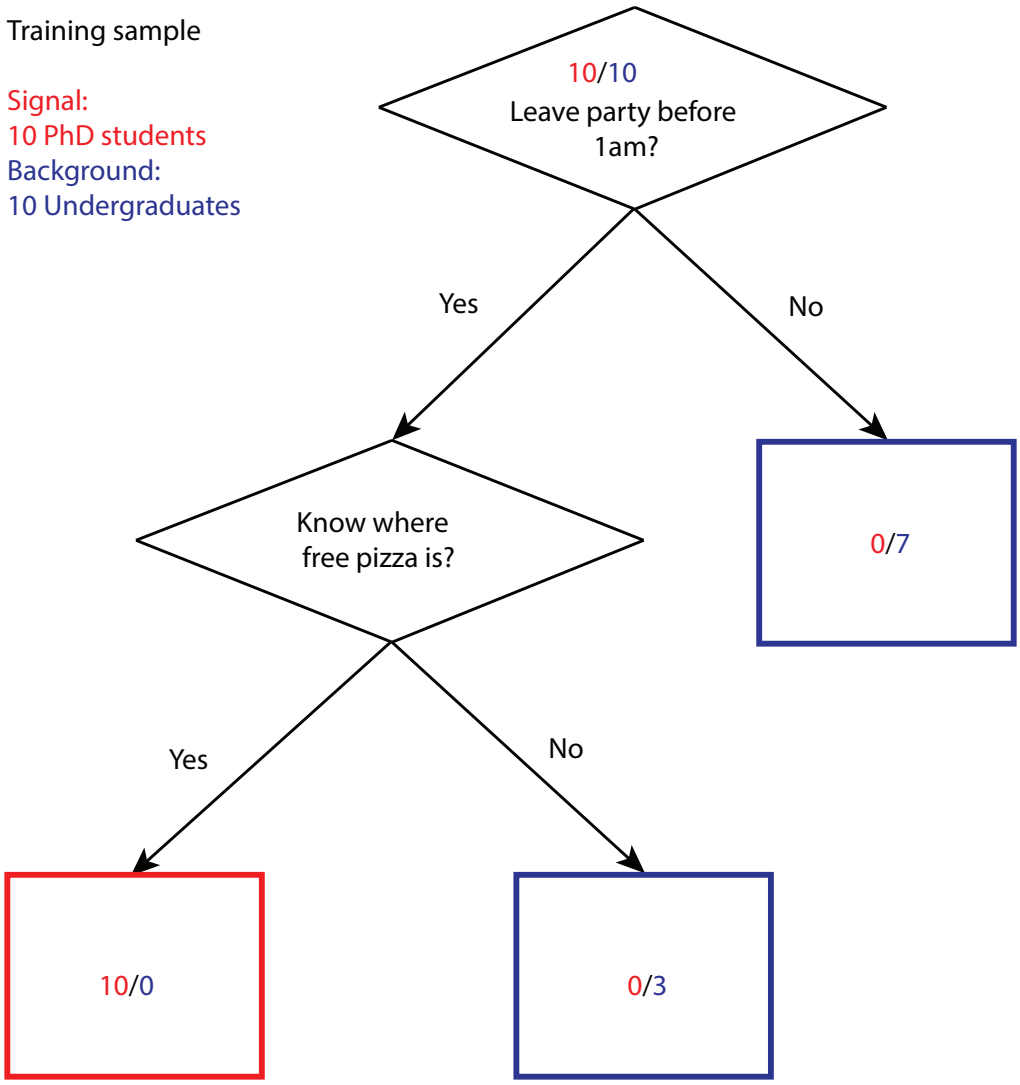


Figure 4.6: Example of a decision tree. Numbers in each node represent the number of PhD students (red) and the number of undergraduate students (blue). Diamond boxes represent splitting nodes. Rectangular boxes represent end nodes. Blue boxes are background-like end nodes. The red box is a signal-like end-node.

- At the initialisation stage the 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 the 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 the m^{th} tree weight, $\alpha_m = \log\left(\frac{1-err_m}{err_m}\right)$
  - Update the 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:

$$G(x) = \begin{cases} -1, & \text{if} \sum_{m=1}^{M} \alpha_m G_m(x) < 0, \\ 1, & \text{otherwise.} \end{cases} \tag{4.7}$$

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

Adaptive boosting dramatically improves the performance of a weak classifier. A weak classifier is a classifier which gives a predictive performance only sightly better than a random guess. A decision tree with one or two layers 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 at selecting signals.

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 lead to the overfitting of the model.

The number of trees is another important parameter. Previous studies on BDTs show that using many small trees yields the best result [81]. It has been shown that a large number of trees does not lead to overfitting [81]. 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 criterion for tree growing, affects the size of the tree but is less important than the depth of the tree parameter.

The boosting algorithm has two variants in the 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 on BDT show that a small learning rate ($\sim$0.1) with many trees works better than a large learning rate with fewer trees [81].

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 [81].

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

The number of bins per variable is a necessary parameter to make tree growing efficient, since it is faster to compute the optimal cut at splitting nodes for discretely binned variables than continuous variables. This parameter, however, has little impact on the model performance. Nevertheless because variables are binned, these variables should be pre-processed before feeding into the training model. For example, a variable should be limited to a range to avoid the extreme values that distort the shape of the variable distribution. If the original distribution of a variable is highly skewed, the variable should be transformed to obtain a more uniform distribution.

For the end node the output can either be 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 range of [0,1].

The bagging fraction determines the fraction of randomly selected events used in each boosting iteration. By choosing a small value, events between each boosting iteration are less correlated. Hence the overall model performance improves.

The DoPreSelection flag in the BDT of Tmva allows the classifier to identify phase spaces where there are only background events and apply cuts to discard them.

4.5.7 Multiple classes

The above discussion assumes exactly two classes; the signal class and the background class. The classification can be extended to multiple classes. There are two ways to train a classifier for multiple classes. The "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, the "one versus one" scheme trains class A against class B; class B against class C; and class C against class A. "One versus all"

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

The TMVA implementation of the multiple classes classifier, 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. The individual response of a trained classifier for an event can be treated as the likelihood. In the application 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 that 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 discriminant variables for each individual class need to enter the training stage simultaneously, resulting in a large number of discriminant 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

A good single photon energy resolution and the ability to reconstruct two spatially close photons are necessary to reconstruct particles using decay processes involving photons, such as $\pi^0 \to \gamma\gamma$ decays. Furthermore, the ability to correctly reconstruct photons in a dense jet environment improves the charged particle reconstruction by removing the calorimeter hits belonged to the photons and simplifying the pattern recognition problem for the charged particle reconstruction.

This chapter starts with an overview of the electromagnetic shower produced by photons passing through a thick absorber. It then discusses photon reconstruction algorithms within the PandoraPFA framework followed by a description of the performance of these algorithms. Part of this chapter has been published in the proceedings of 2015 International Workshop on Future Linear Colliders [83]. The photon reconstruction algorithms presented in this chapter have benefited a number of physics analyses. The most recent example is the $H \to \gamma\gamma$ simulation study at CLIC [84].

5.1 Electromagnetic showers

Electromagnetic (EM) showers develop through processes of the pair production and bremsstrahlung when a high energy photon or electron passes though a thick absorber. Many low-energy photons and electrons are generated producing shower-like structures in the detector. Two suitable length scales to describe the EM shower growth are the radiation length and the Molière radius [85, 86].

The radiation length of a material describes the EM longitudinal shower profile defined as the mean distance travelled by an electron for its energy to be reduced by a factor of $1/e$ via bremsstrahlung. It is also defined as the 7/9 of the mean free path for pair production by a high energy photon [87].

Figure 5.1 shows the simulated longitudinal electromagnetic shower profiles as a function of the longitudinal shower depth for electrons and photons. The mean EM longitudinal shower profile can be described by the following function [88]:

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

where t is the number of radiation lengths; the parameter E_0 is the initial energy of the photon/electron; the parameter b takes the value of 0.5 which is sufficient for the purpose of photon reconstruction [3]; and the parameter a is given by [37]:

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

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 [89]. The alternative definition of the critical energy is the energy at which the energy loss by ionisation per radiation length is the same as the particle energy [90]. This parametrisation of the EM longitudinal shower profile should only be used to describe an average behaviour of the EM shower; fluctuations in individual EM shower profiles are significant.

The EM transverse shower profile can be described as a narrow core, widening as the shower develops. 90% of the shower energy is contained in a cylinder with a radius of one Molière radius. About 99% of the shower energy is contained inside of 3.5 Molière radii [85, 86].

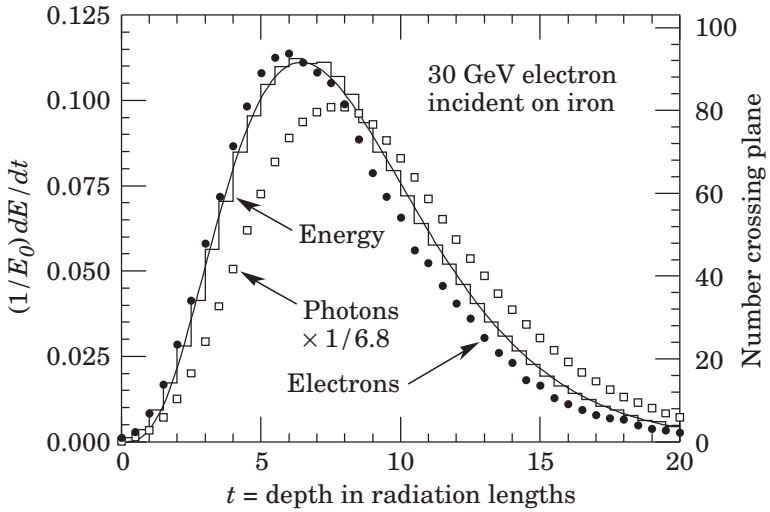


Figure 5.1: An EGS4 [91, 92] simulation of a 30 GeV electron-induced electromagnetic shower in iron. The histogram shows the fractional energy deposition as a function of radiation length. The curve is a gamma-function fit to the distribution. Circles and squares are the numbers of electrons and photons respectively with total energy greater than 1.5 MeV crossing planes with scale on right. Plot taken from [3].

5.2 Overview of photon reconstruction in PandoraPFA

Five algorithms are developed to tackle different issues in photon reconstruction in PandoraPFA:

- Photon Reconstruction algorithm reconstructs photons from calorimeter hits in the ECAL, including forming a photon candidate and applying a photon identify test, with special treatments for photons close to charged particles.
- Two photon fragment removal algorithms remove fragments in the ECAL. Fragments refers to multiple reconstructed particles corresponding to the same MC particle.
- One algorithm removes fragments in the HCAL.
- A photon splitting algorithm separates merged photons.

Places of the photon algorithms used in the PandoraPFA are shown in figure 5.2. The five photon algorithms are highlighted in red.

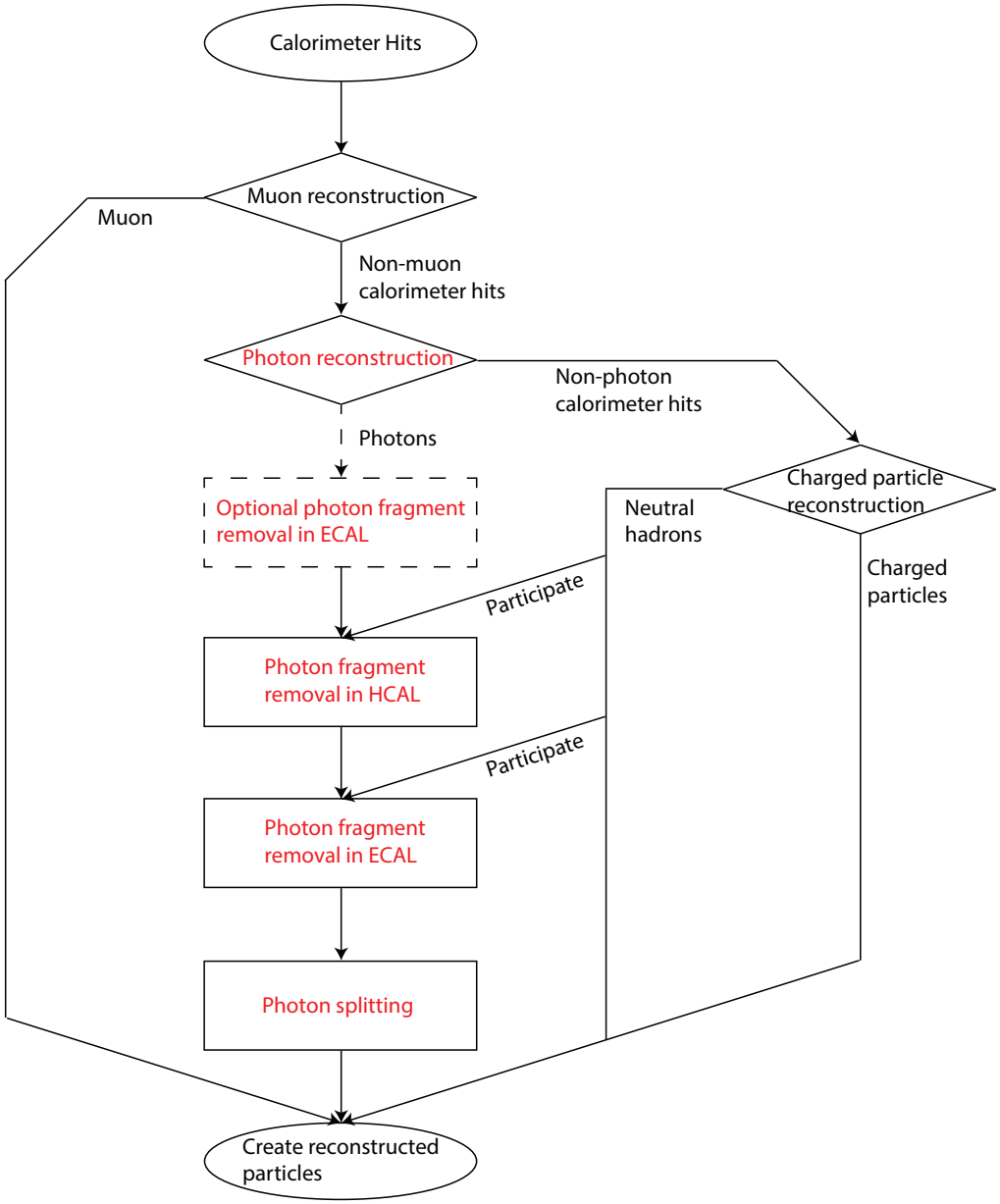


Figure 5.2: Places of the photon algorithms used in the PandoraPFA. Five photon algorithms are highlighted in red.

5.3 Photon Reconstruction algorithm

The PHOTON RECONSTRUCTION algorithm runs at an early stage of the overall reconstruction, before the charged particle reconstruction. The main steps of the PHOTON RECONSTRUCTION algorithm, shown in figure 5.3, are: forming photon clusters; finding photon candidates; and a photon identity test.

The PHOTON RECONSTRUCTION algorithm runs after the muon reconstruction algorithm as shown in the schematic diagram of the algorithms in PandoraPFA in figure

5.2. Inputs of the PHOTON RECONSTRUCTION algorithm are calorimeter hits in the ECAL that are not associated with reconstructed muons.

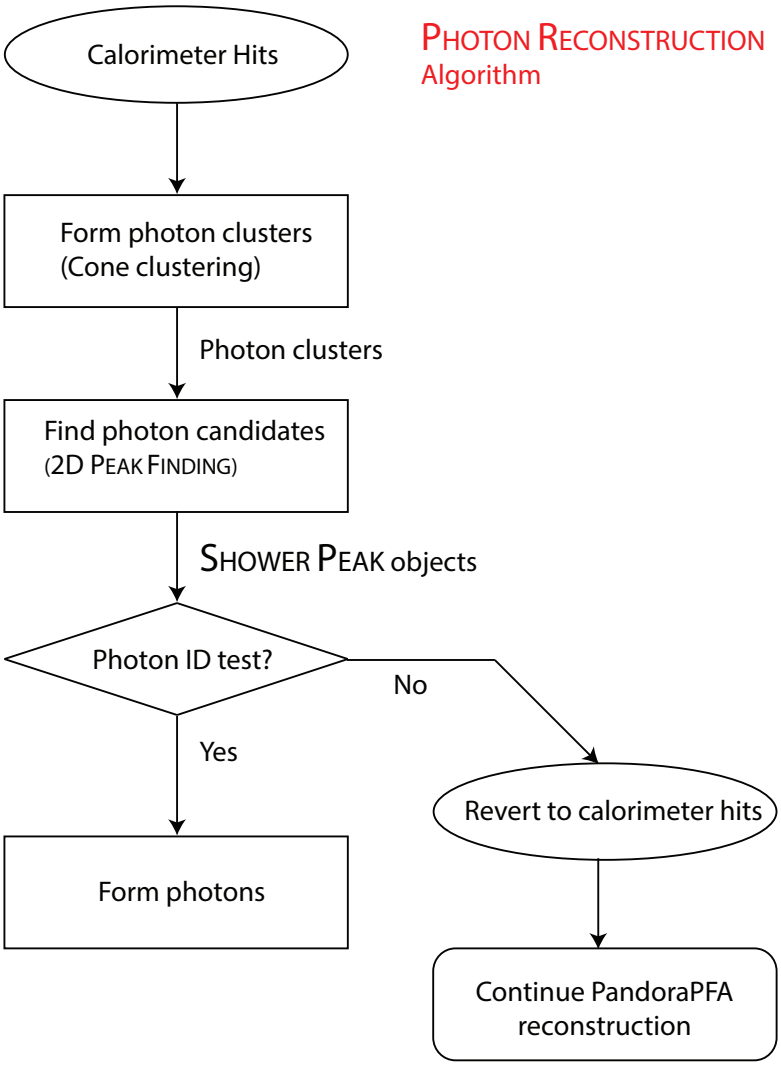


Figure 5.3: Main steps of the PHOTON RECONSTRUCTION algorithm: forming photon clusters; finding photon candidates; and photon identity test.

5.3.1 Forming photon clusters

Clusters are created from calorimeter hits in the ECAL using the cone clustering algorithm. As the target for reconstruction is the neutral photon, the cone clustering algorithm uses high-energy calorimeter hits in the ECAL as initial seeds in the order of descending energies instead of using track projections as initial seeds. Parameters for large cones are used to form clusters such that it is unlikely that one photon is split into two clusters but one cluster may contain calorimeter hits from multiple photons.

5.3.2 Finding photon candidates and 2D Peak Finding algorithm

If a cluster contains calorimeter hits from several photons, the algorithm aims to split the three-dimensional cluster into several smaller clusters (photon candidates). Ideally each photon candidate should contain calorimeter hits from one photon only.

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. 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 peak-like structure in the plane. Figure 5.4 shows two EM shower cores from a single cluster.

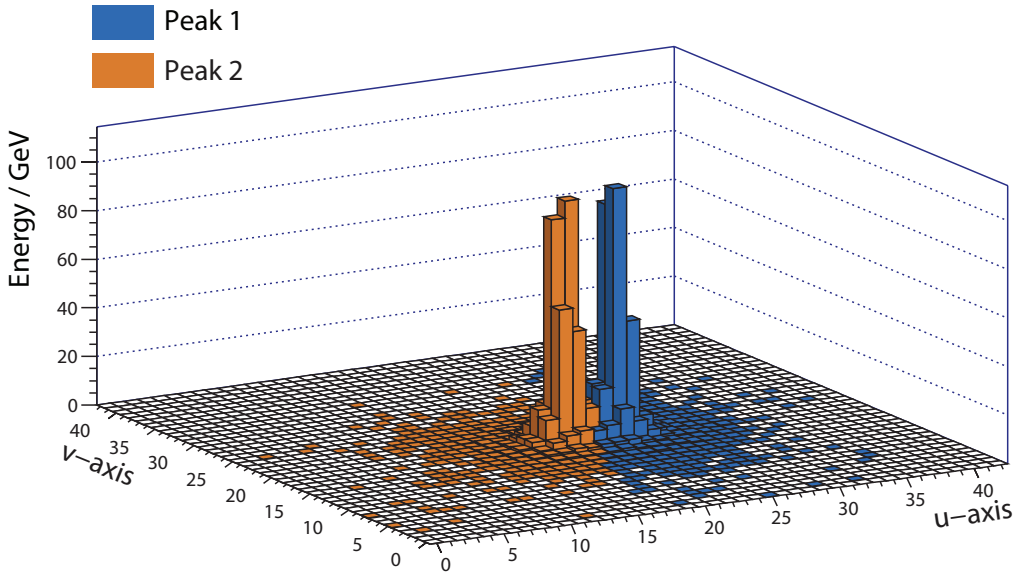


Figure 5.4: Two 500 GeV photons (yellow and blue) within a cluster, just resolved in a transverse plane orthogonal to the direction of the cluster. The axes U and V are orthogonal axes in units of the ECAL cell lengths. The height of a bin in the histogram is the sum of the calorimeter hit energy associated with the bin.

Identifying photon candidates inside a cluster is equivalent 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 non-peak bins to identified peaks.

There are two variants of the 2D Peak Finding algorithm: the neutral cluster variant and the charged cluster variant. The main steps of the neutral cluster variant are

shown in figure 5.5: creating the 2D energy deposition; peak reconstruction; and forming Shower Peak objects.

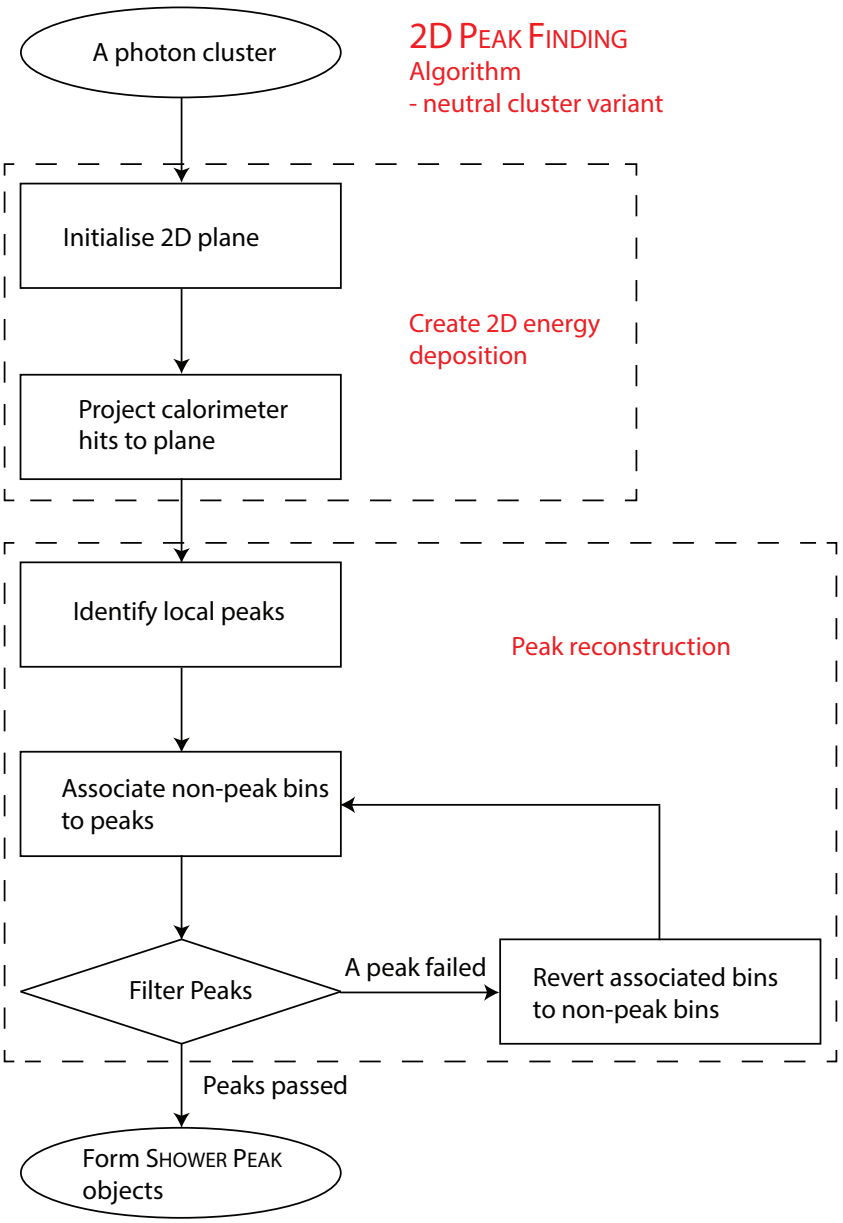


Figure 5.5: Main steps of the neutral cluster variant of the 2D Peak Finding algorithm: creating the 2D energy deposition; peak reconstruction; and forming Shower Peak objects.

Creating the two-dimensional energy deposition

A two-dimensional (2D) plane is used to host the projection of the calorimeter hits 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.

The direction of the cluster is the direction of the IP to the centroid of the cluster. The first axis vector, $\vec{u}$, is defined as:

$$\vec{u} = \text{norm}\Big(\langle\vec{a}\rangle_y,\ -\langle\vec{a}\rangle_x,\ 0\Big), \tag{5.3}$$

where norm is the normalisation operator and $\langle\vec{a}\rangle_x$ is the x component of the centroid position of cluster a assuming the IP is at origin. The second axis vector, $\vec{v}$, is defined as:

$$\vec{v} = \text{norm}\Big(\vec{u} \times \langle\vec{a}\rangle\Big). \tag{5.4}$$

The calorimeter hits associated with the cluster are projected onto the two-dimensional plane. The distance between the calorimeter hit position and the cluster centroid position is converted into a distance vector. The distance vector, $\vec{s_i}$, of a calorimeter hit i, is defined as:

$$\vec{s_i} = \frac{\vec{a_i} - \langle\vec{a}\rangle}{d_{cell}}, \tag{5.5}$$

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 plane is calculated from the scalar products of the distance vector, $\vec{s_i}$, with the axes vectors: $\vec{u}$ and $\vec{v}$. The calorimeter hits in the two-dimensional plane are binned in a two-dimensional histogram. 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. One bin size along either axis on the 2D histogram corresponds to one ECAL square cell length.

Peak reconstruction

Local peaks are identified in the 2D histogram. A local peak is defined as a bin where its height is above all eight neighbouring bins. Figure 5.6 shows an example of a 2D histogram with two local peak bins (orange and blue) identified. Red bins are non-peak bins.

Having identified all local peaks, non-peak bins are associated 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.

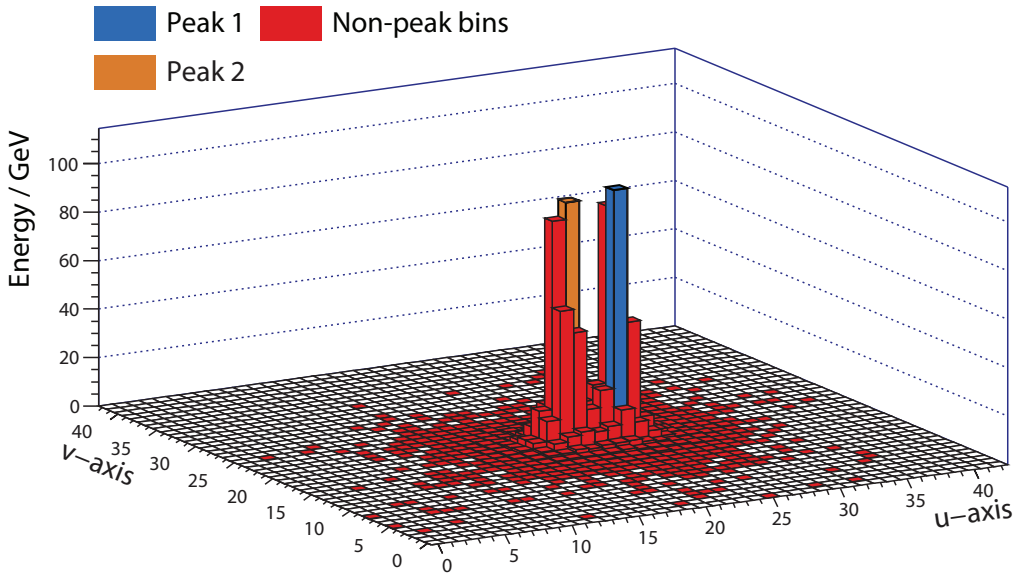


Figure 5.6: Two peak bins indicated with orange and blue colours are identified. Red bins are non-peak bins.

A non-peak bin is associated with the peak bin that gives the smallest value of the metric:

$$\frac{d_i}{\sqrt{E_i}} \tag{5.6}$$

where d_i is the Euclidean distance between a non-peak bin and a peak bin i in the 2D histogram, and E_i is the height (energy) of the peak bin i. For each non-peak bin, the metric is considered for all peak bins to find the peak bin that produces the smallest metric. For the 2D histogram in figure 5.6 with two peaks identified, the result after associating non-peak bins to peak bins is shown in figure 5.4.

In the 2D histogram, major peaks with many associated non-peak bins most likely correspond to physical photons, while minor peaks with a few associated non-peak bins are more likely from fluctuations in the energy deposition of the EM shower. The performance of the 2D Peak Finding algorithm is thus improved by discarding small peaks. After all non-peak bins are associated with peak bins, peaks with fewer than three non-peak bins associated are discarded. These discarded non-peak bins are re-associated with other peak bins. This process is iterated until all peak bins have at least three bins associated.

After filtering peaks, Shower Peak objects are created. 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.

Charged cluster variant

In a dense jet environment, if a photon nearby a charged hadron is well reconstructed, the charged particle reconstruction will be improved. Charged cluster variant aims to carefully identify photon candidates next to charged hadrons, by using track information and features of EM showers. An EM shower typically starts in the first few layers of the ECAL with direction of the EM shower largely unchanged when the shower develops.

Figure 5.7 shows the main steps in the full 2D Peak Finding algorithm, including the treatment of clusters close to tracks. The first step of the algorithm, "Close to track", determines if a cluster is close to a track. If the distance between a cluster and the closest track projection onto the front of the ECAL is fewer than 3 mm, the charged cluster variant of the 2D Peak Finding algorithm is applied to the cluster.

The "Create and iterate over sub-clusters" stage performs the following. The ECAL is sliced longitudinally to create fiducial volumes. For example, the default three slices will result in three fiducial volumes in the ECAL. Each fiducial volume covers the space from the front of the ECAL to a third, to two thirds, and to the back of the ECAL. Three sub-clusters are created from the calorimeter hits of the cluster that are contained in each fiducial volume. In the example of the ILD detector model, the first sub-cluster is formed with the calorimeter hits of the cluster in the first 10 layers of the ECAL. The second and the third sub-cluster are formed with the calorimeter hits of the cluster in the first 20 layers the ECAL and the entire ECAL respectively.

After creating sub-clusters, the neutral cluster variant of the 2D Peak Finding algorithm is applied to each sub-cluster to find peaks. Shower Peak objects are created from peaks, identified with the 2D Peak Finding algorithm.

The Shower Peak objects created from each sub-cluster undergo the "Shower Peak filter" step. Peaks in the first sub-cluster are preserved. Peaks in the second sub-cluster are preserved if the peak bin position is the same as a preserved peak bin position in the first sub-cluster and a shift in the peak bin position by no more than one neighbouring bin is allowed. Similarly, peaks in the third sub-cluster are preserved if the peak bin position is the same as a preserved peak bin position in the second sub-cluster, allowing a shift in the peak bin position by no more than one neighbouring bin. Only preserved peaks in the third sub-cluster are used to form the final Shower Peak objects.

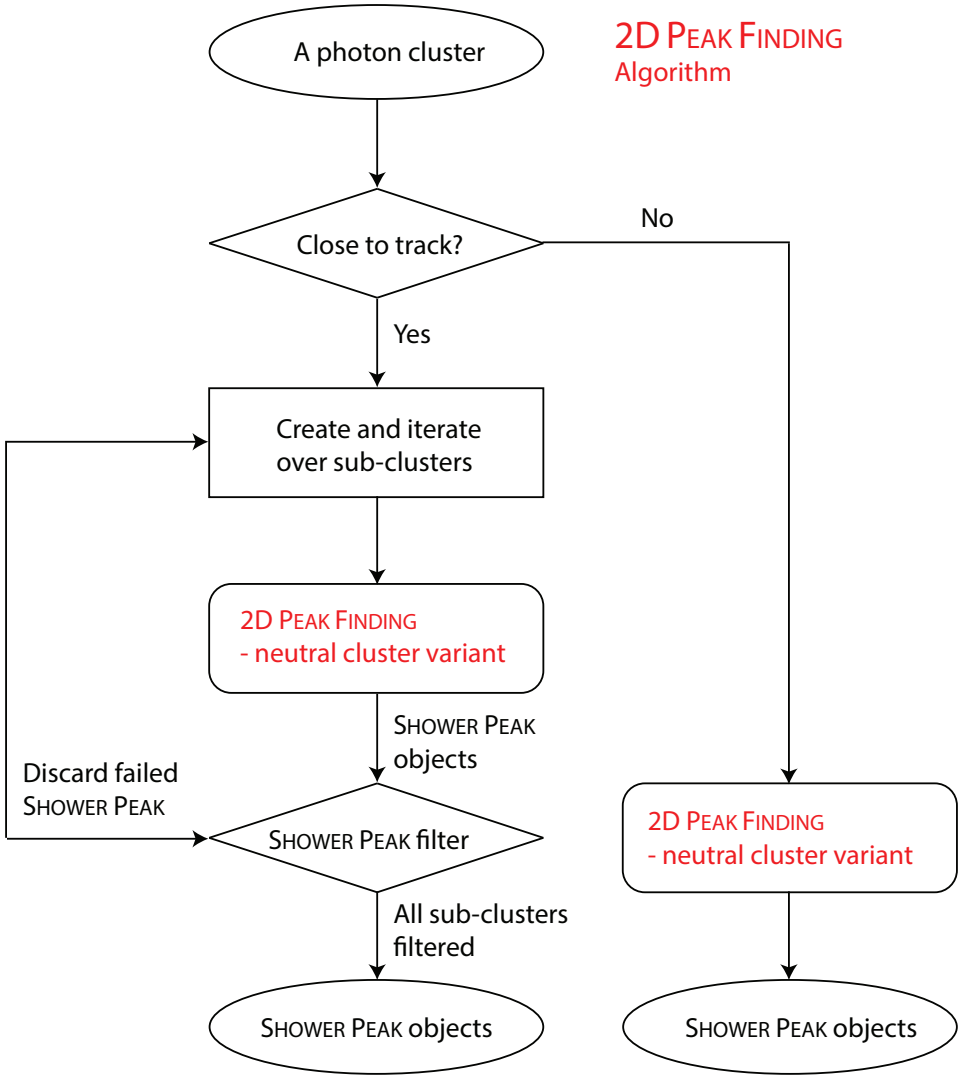


Figure 5.7: Main steps of the 2D Peak Finding algorithm, including the charged cluster variant: identifying whether the cluster is close to a track; creating and iterating over sub-clusters; applying 2D Peak Finding algorithm neutral cluster variant to sub-clusters; filtering Shower Peak objects in sub-clusters; and creating final Shower Peak objects.

Furthermore, if a peak bin is within one neighbouring bin of a track projection bin, the peak is discarded. The track projection bin is the bin where the track projection onto the front of the ECAL projects onto the 2D histogram.

Figure 5.8 illustrates an example of three sub-clusters created during the charged variant of the 2D Peak Finding algorithm. Peaks with associated bins and track projection bins are labelled. Figure 5.8a shows the first sub-cluster, created using calorimeter hits in the first 10 layers of the ECAL. One peak is identified. Figure 5.8b shows the second sub-cluster, created using calorimeter hits in the first 20 layers of the

ECAL. Two peaks are identified. Only the blue peak is in the same position of the blue peak in the first sub-cluster. Hence, the blue peak in the second cluster is preserved. Figure 5.8c shows the third sub-cluster created using calorimeter hits in the entire ECAL. Three peaks are identified. Only blue peak is in the same position of the blue peak in the second sub-cluster. Hence, only blue peak in the third sub-cluster is preserved. The preserved blue peak and associated bins in the third sub-cluster are then used to create one Shower Peak object.

Inclusive mode

The time complexity of iterating a 2D 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, due to the finite size of the histogram only calorimeter hits projected onto the histogram would be considered by 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 the purpose of 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.

5.3.3 Photon Identity test

This step applies the photon identity test on the Shower Peak object. The photon identity test uses a multi-dimensional likelihood classifier.

Variables used in likelihood classifier

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 particle. Variables used in the classifier are listed in table 5.1. All plots in this section are produced from simulated $e^+e^- \to Z'Z'$ events where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$ at a centre-of-mass energy of 500 GeV.

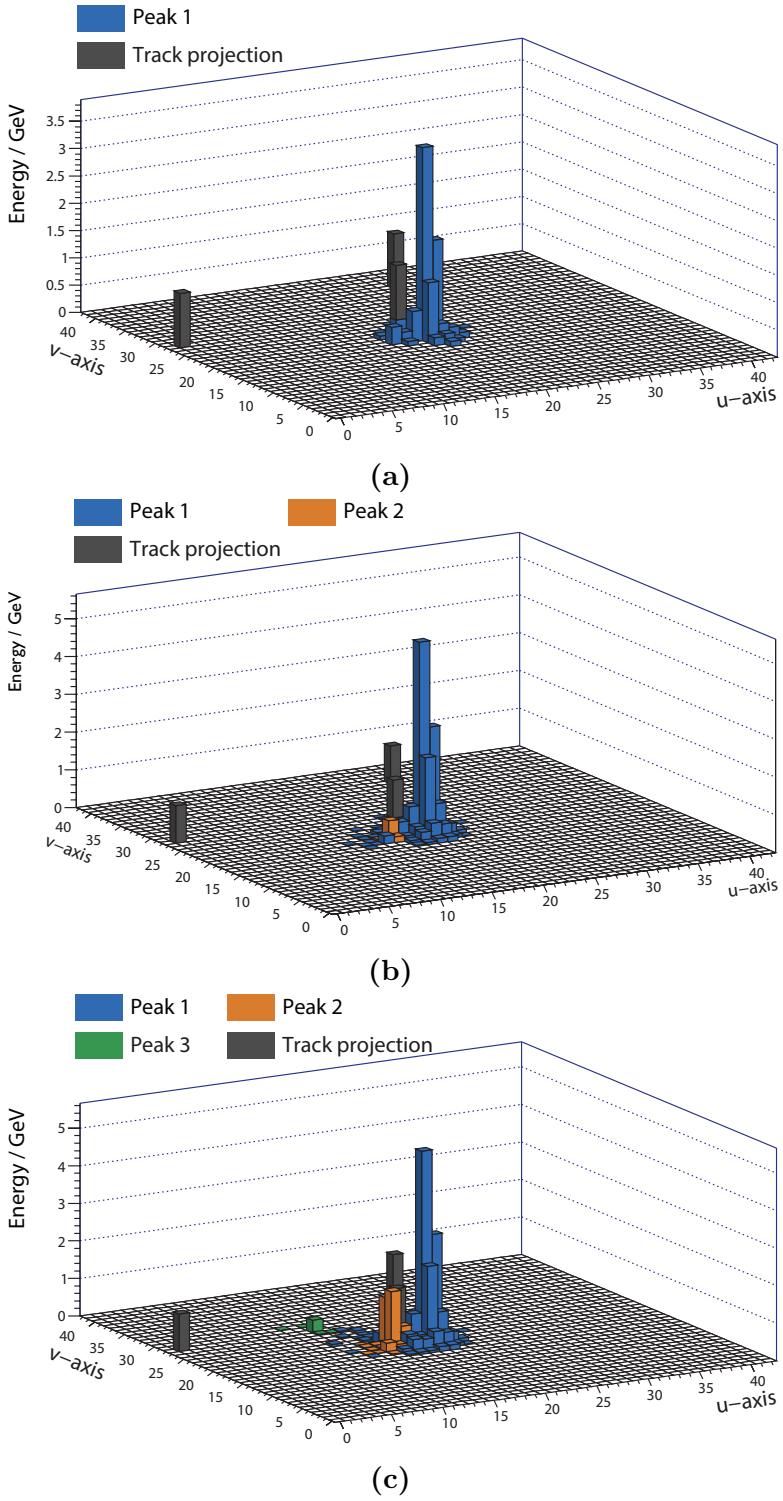


Figure 5.8: An illustration of sub-clusters created during the charged cluster variant of the 2D Peak Finding algorithm. Sub-clusters are created using calorimeter hits in: a) first 10 layers of the ECAL, b) first 20 layers of the ECAL, and c) the entire ECAL. Peaks with associated bins and track projection bins are labelled.

Two variables are obtained from the EM longitudinal shower profile: t_0 is the start layer in the ECAL of the fitted EM longitudinal shower profile shown in figure 5.9a; and δl is fractional difference of the observed EM longitudinal shower profile to the expected EM longitudinal shower profile described in equation 5.1:

$$\delta l = \frac{1}{E_0} \sum_i \left| \Delta E^i_{obs} - \Delta E^i_{EM} \right|, \tag{5.7}$$

where E_0 is the energy of the EM shower; ΔE^i_{EM} is the energy of the expected EM longitudinal shower profile in bin i; ΔE^i_{obs} is the energy of the observed EM longitudinal shower profile in bin i; the index i is summed over the ECAL layers as the EM longitudinal shower profile is binned according to the ECAL layers; and the quantity δl is minimised as a function of ΔE^i_{EM} which is a function of t_0 via equation 5.1. The δl distributions for photons and non-photon particles are shown in figure 5.9b. 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 observed EM longitudinal shower profile should be similar to an expected EM longitudinal shower profile.

Three variables are obtained from the EM transverse 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 5.9c; the variable $\langle w_{UV} \rangle$ 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 variable, $\tilde{E}_{cluster}$, is the ratio of the energy of the Shower Peak object to the cluster energy, a measure of the dominance of the 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. The distributions of d for photons and non-photon particles are shown in figure 5.9d.

Categories	Variables
EM longitudinal shower profile	δl, t_0
EM transverse shower profile	$\langle w \rangle$, $\langle w_{UV} \rangle$, $\tilde{E}_{cluster}$
Distance to track	d

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

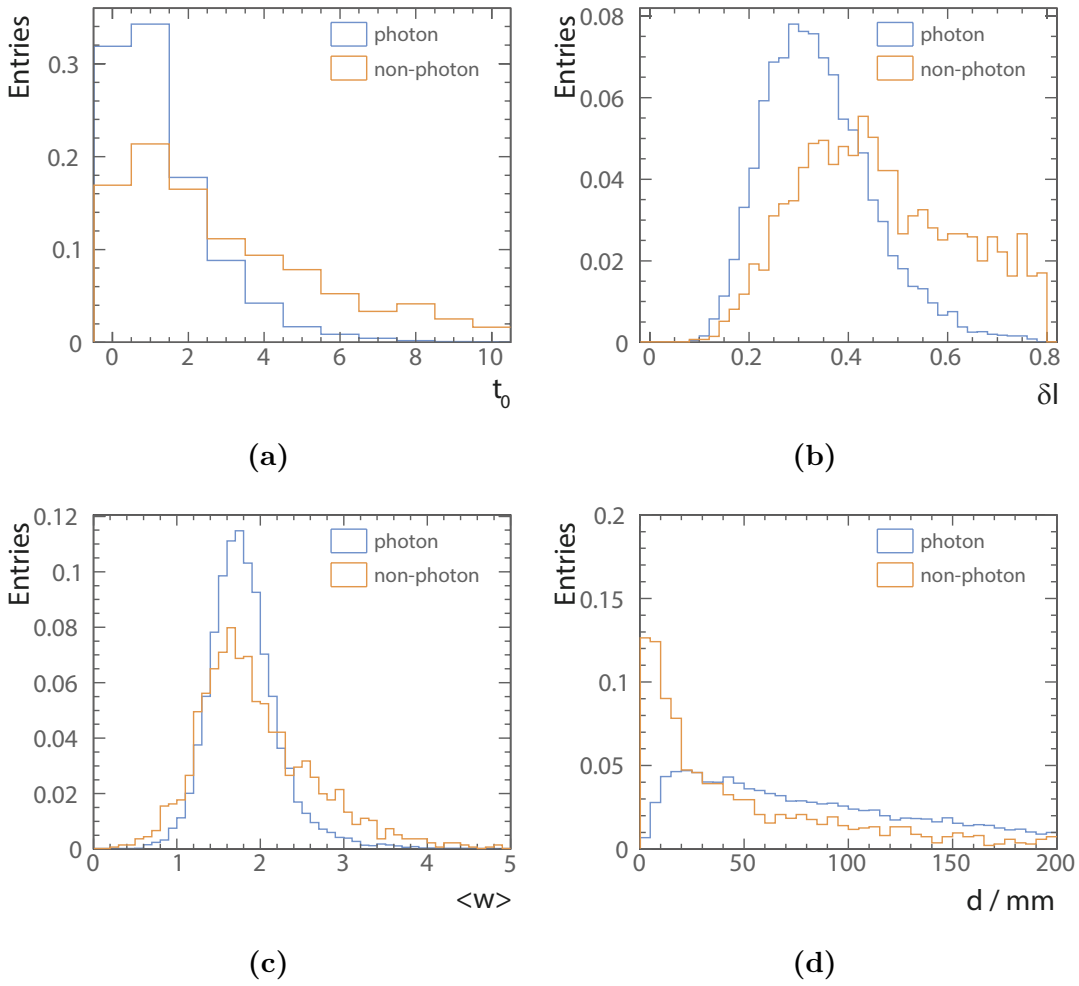


Figure 5.9: Distributions of: a) the start layer from the longitudinal shower profile (t_0); b) the fractional difference of the observed shower profile to the expected EM shower profile (δl); c) the energy weighted root-mean-square distance of all bins in a Shower Peak to its peak bin ($\langle w \rangle$); and d) the distance between the photon candidate and the closest track projection onto the front of the ECAL (d). The area under each curve is normalised to unity.

Projective Likelihood classifier

Projective likelihood classifier is used for the photon identity test due to its low requirement on computing resources comparing to a Boosted Decision Tree classifier or a Neural Network classifier.

The probability distributions for each variable for photons and non-photon particles are obtained in the training stage. The distributions of these variables are normalised to unity, stored in binned histograms. The classifier is improved by realising that the variable distributions depend on photon energy. Thus, the variables distributions are

stored separately for different photon energy ranges. Eight photon energy ranges are used by binning the distribution of photon energies at 0.2, 0.5, 1, 1.5, 2.5, 5, 10, 20 GeV.

The training stage of the classifier uses simulated $e^+e^- \to Z'Z'$ events where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$, at a centre-of-mass energy of 500 GeV. Events at the 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 Shower Peak object with the energy in the energy bin α, the classifier output is given by

$$\mathrm{PID}_\alpha = \frac{N \prod_i^6 P_i}{N \prod_i^6 P_i + N' \prod_i^6 P'_i} \tag{5.8}$$

where P_i and P'_i are the values in the i^{th} variable probability distributions of the respective photon and non-photon particles in the energy bin α; the variables N and N' are the number of respective photons and non-photon particles in the energy bin α in the training samples.

A Shower Peak object passes the photon identity test if

$$\begin{cases} \mathrm{PID} > 0.6, & \text{if } 0.2 < E < 0.5\,\mathrm{GeV}, \\ \mathrm{PID} > 0.4, & \text{if } E \geqslant 0.5\,\mathrm{GeV}, \end{cases} \tag{5.9}$$

where E is the energy of the Shower Peak object. Two values of the cuts on PID are motivated by the fact that 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. Figure 5.10 shows the distributions of PID for photons and non-photons with energies between 0.2 and 0.5 GeV. Hence for Shower Peak objects with energies between 0.2 and 0.5 GeV, PID > 0.6 is required instead of PID > 0.4.

If a Shower Peak object passes the photon identity test, the Shower Peak object is tagged as a photon. If a Shower Peak object fails the photon identity test, the Shower Peak object is discarded. Calorimeter hits associated with the discarded Shower Peak object are freed up and are passed onto the next stage of the reconstruction.

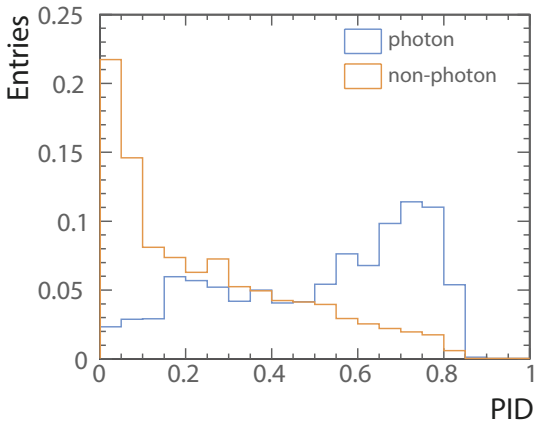


Figure 5.10: The distributions of PID for photons and non-photons with energies between 0.2 and 0.5 GeV. The area under curve is normalised to unity.

5.4 Photon fragment removal in the ECAL

Sometimes not all hits from a photon are reconstructed and identified as one photon cluster. Hits form small clusters, known as fragments. Figure 5.11 shows an example of creation of a photon fragment. A fragment typically does not have the electromagnetic shower structure, and has a much lower energy than a main photon.

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

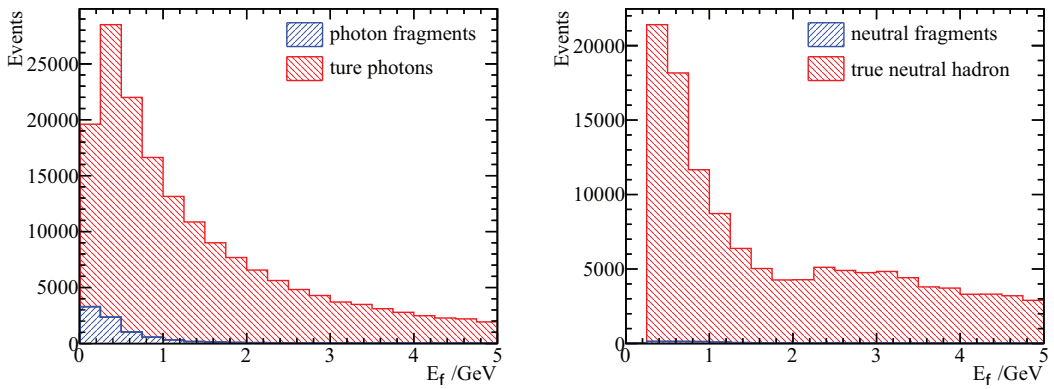
Figure 5.11: An event display of a) a typical 10 GeV photon, reconstructed into b) the main photon cluster, and c) a photon fragment.

A photon and a fragment form a photon−fragment pair. Depending on whether the fragment is reconstructed as a photon or a neutral hadron, the photon−fragment pairs are further classified into photon−photon-fragment pairs and photon−neutral-fragment pairs. The neutral fragment refers to the fragment reconstructed as a neutral hadron.

Figure 5.12 shows the energies of the second most energetic reconstructed photon in the photon−photon-fragment pairs, the true photon−photon pairs, photon−neutral-fragment pairs, and true photon−neutral-hadron pairs. Most photon and neutral hadron fragments have energies below than 1 GeV. Hence the photon−fragment pairs are subsequently

divided into low-energy and high-energy pairs, depending on whether the fragment energy, E_f, is above 1 GeV.

Plots in this section are obtained with 10000 simulated $e^+e^- \to Z'Z'$ events where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$ at $\sqrt{s} = 500\,\text{GeV}$ reconstructed with the PandoraPFA version 1.



(a) **(b)**

Figure 5.12: The energies of the second most energetic reconstructed photon in a) the photon−photon-fragment pairs and true photon−photon pairs, and in b) the photon−neutral-fragment pairs and true photon−neutral-hadron pairs.

There are two variants of the photon fragment removal algorithms: one immediately after the PHOTON RECONSTRUCTION algorithm, and the other one after the charged particle reconstruction, shown in the schematic diagram of the algorithms in PandoraPFA in figure 5.2. Since two algorithms share similar logics for fragment removal, the algorithm used after the charged particle reconstruction will be discussed in detail here.

The aim for the photon fragment removal algorithm is to merge fragments to main photons based on sets of cuts. Table 5.2 lists cuts for merging photon−photon-fragment pairs and photon−neutral-fragment pairs for both low-energy and high-energy fragments. Using the cuts for photon−photon-fragment pairs with low-energy fragments as an example, each set of logics for merging fragments is discussed. There are five sets of cuts. A photon−fragment pair passing any one set of cuts will be merged.

1. The transverse EM shower comparison cut merges fragments 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 energy in the fragment, $\frac{E_{p2}}{E_f} < 0.5$. And the most energetic peak should

have more energy than the main photon, $E_{p1} > E_m$, where E_m is the energy of the main photon; E_f is the energy of the fragment; and E_{p1} and E_{p2} are the respective energies of the two most energetic EM showers identified by the 2D Peak Finding algorithm using the photon−fragment pair as input, ordered by descending energy. Lastly the fragment should be close to the main photon, $d < 30\,\text{mm}$, where d is the average energy weighted intra-layer distance between the photon and the fragment:

$$d = \frac{\sum_i^{layers} d_l^i \; E_f^i}{\sum_i^{layers} E_f^i} \tag{5.10}$$

where index i indicates the i^{th} layer of the ECAL; the parameter d_l^i is the minimum distance between calorimeter hits of the photon and the fragment in the i^{th} layer; and E_f^i is the total energy of calorimeter hits of the fragment in the i^{th} layer of the ECAL. Figure 5.13a and figure 5.13b show the average energy weighted intra-layer distance, d, for photon−photon-fragment pairs and true photon−photon pairs, for low-energy and high-energy fragments respectively. Photon−fragment pairs typically have a small distance separation between the photon and the fragment.

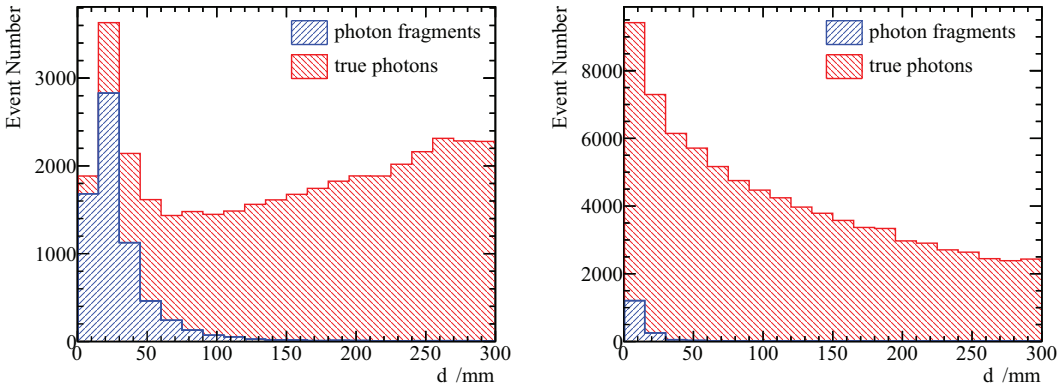
2. The close proximity cut merges fragments when the fragment has a low energy and is spatially close to the main photon: $d < 20\,\text{mm}$ and E_f <0.2 GeV.

3. The third set of cuts is used for the case when fragments are spatially close to the main photon and have very few number of associated calorimeter hits. Either the photon−fragment pair satisfies: $d < 30\,\text{mm}$; $d_c < 50\,\text{mm}$; and $N_{calo} < 40$, or the photon−fragment pair satisfies: $d < 30\,\text{mm}$, and $N_{calo} < 50$, where N_{calo} is the number of the calorimeter hits in the fragment. The multiple cuts allow the merging of a fragment with a fewer calorimeter hits with a slightly larger distance separation to the main photon, or the merging of a fragment with a slightly larger number of calorimeter hits with a smaller distance separation to the main photon.

4. The fourth set of cuts merges low-energy fragments in the endcap region of the detector. Fragments are merged if: $d_c < 60\,\text{mm}$; $|\cos(\theta_Z)| > 0.7$; $E_f < 0.6\,\text{GeV}$; and $N_{calo} < 40$. Here $|\cos(\theta_Z)|$ is the absolute value of the cosine of the polar angle of the main photon with respect to the beam direction, and d_c is the distance between centroids of the photon and the fragment. Figure 5.14a and figure 5.14b shows the distance between centroids, d_c, for photon−neutral-fragment pairs and the true photon−neutral-hadron pairs, for low-energy and high-energy fragments

respectively. Photon−fragment pairs typically have a small distance separation between the photon and the fragment.

5. The last set of cuts is that the merged fragment should be relatively low energetic. The ratio of the fragment energy to the main photon energy should be less than 0.01. The distance between the pair should satisfies $d < 40\,\text{mm}$ and $d_h < 20\,\text{mm}$, where d_h is the minimum distance between calorimeter hits of the photon and the fragment.



(a) **(b)**

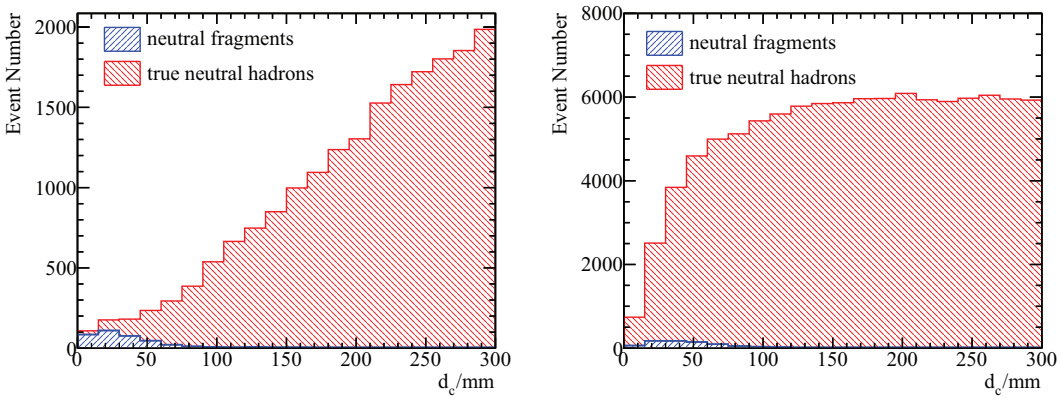
Figure 5.13: Distributions of average energy weighted intra-layer distances, d, for: a) photon−photon-fragment and the true photon−photon pairs for low-energy fragments; and b) photon−photon-fragment and the true photon−photon pairs for high-energy fragments.

All possible photon−fragment pairs are considered. 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 considered, this is a costly operation. The speed of the algorithm is improved by only considering pairs with $d < 80$ mm.

5.4.1 Photon fragment removal algorithm after the Photon Reconstruction algorithm

The photon fragment removal algorithm immediately after the PHOTON RECONSTRUCTION algorithm shares similar logics as the stated above. The cuts for merging fragments are listed in table 5.3.



(a) (b)

Figure 5.14: Distributions of distances between centroids, d_c, for: a) photon−neutral-fragment and the true photon−neutral-hadron pairs for low-energy fragments; and b) photon−neutral-fragment nd the true photon−neutral-hadron pairs for high-energy fragments.

5.5 Photon fragment removal algorithm in the HCAL

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 (red) and a neutral hadron fragment in the HCAL (blue) is shown in figure 5.15. This section presents an algorithm to merge fragments in the HCAL to the main photon.

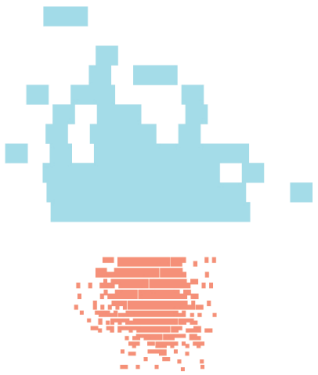


Figure 5.15: An event display of a typical 500 GeV photon, reconstructed into a main photon in the ECAL (red) and a neutral hadron fragment in the HCAL (blue).

$E_f \leqslant 1\,\text{GeV}$	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-fragment pairs for both low-energy and high-energy fragments, after charged hadron reconstruction.

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. It considers all pairs of reconstructed photons and neutral hadrons. Photon−fragment pairs passing all cuts will be merged. There are six sets of cuts:

1. The adjacent in layers cut demands that the photon cluster deposits energies in the last outer layer of the ECAL and the fragment deposits energies in the first inner layer of the HCAL.

$E_f \leqslant 1\,\text{GeV}$	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-fragment pairs for both low-energy and high-energy fragments, immediately after photon reconstruction.

2. The energy comparison cut requires that the fragment have a low energy relative to the main photon. The ratio, $\frac{E_f}{E_m}$, has to be less than 0.1 for merging. The variables E_m and E_f are the energy of the main photon and the energy of the fragment respectively. Figure 5.16 shows the distributions of the energy fractions, $\frac{E_f}{E_m}$, after passing the adjacent in layers cut, for photon fragments in jet samples (blue), non-fragments in jet samples (orange), and photon fragments in single-photon samples (green). Jet samples are $e^+e^- \to Z'Z'$ events where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$ at a centre-of-mass energy of 500 GeV reconstructed with the PandoraPFA version 1. Single-photon samples are single 500 GeV photon events reconstructed with the PandoraPFA version 1. The cut $\frac{E_f}{E_m} < 0.1$ contains most of the fragments.

3. The distance comparison cuts requires that fragment in the HCAL is spatially close to the main photon measured by three distance 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. The fitted direction is the direction where the most calorimeter hits align to. Three distances should be small for merging. The cuts demand: $d_c^l \leqslant 173\,\text{mm}$; $d_{fit}^l \leqslant 100\,\text{mm}$; and $d_{fit} \leqslant 100\,\text{mm}$. Figure 5.17 shows the distributions of ${d_c^l}^2$ after passing the adjacent in layers cut and the energy comparison cut, for photon fragments in jet samples (blue), non-fragments in jet samples (orange), and photon fragments in single-photon samples (green). The cut at $d_c^l \leqslant 173\,\text{mm}$ (${d_c^l}^2 \leqslant 3000\,\text{mm}^2$) covers most of the fragments.

4. The projection comparison cut states that the fitted direction of the fragment should be similar to the fitted 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. The shower width comparison cut requires that the shower width of the fragment and the shower width of the main photon are similar. Variable w_m^l is the root-mean-square distance of the calorimeter hits of the main photon in last outer layer in the ECAL to the centroid of the calorimeter hits in the same layer. Variable w_f^l is the root-mean-square distance of the calorimeter hits of the fragment in the first inner layer in the HCAL to the centroid of the calorimeter hits in the same layer. The ratio $\frac{w_f^l}{w_m^l}$ needs to be in the range from 0.3 to 5 to pass the cut. The generous upper bound is because the HCAL cell size is much larger than the cell size of the ECAL.

6. The last cut, the cone comparison cut, demands that 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 fitted cone of a photon is the cone with the smallest opening angle that contains all calorimeter hits of the photon. The variable, $\frac{N_{cone}}{N_f}$, 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.

If multiple photon−fragment pairs pass the cuts with the same fragment, the pair with highest $\frac{N_{cone}}{N_f}$ will be merged.

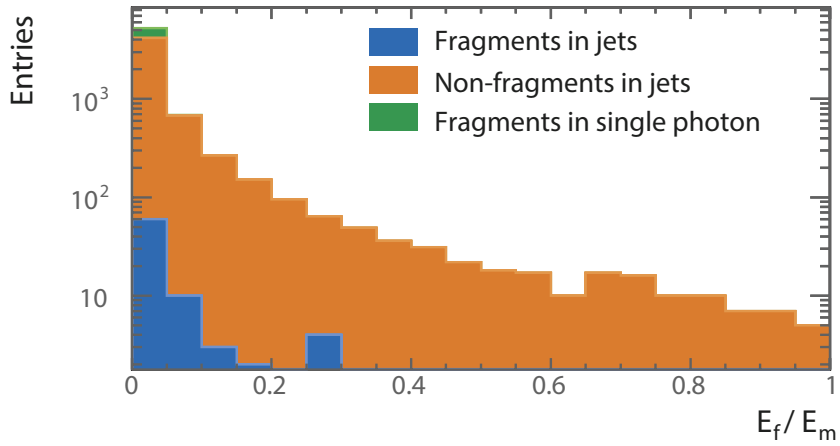


Figure 5.16: The stacked distributions of the energy fractions ($\frac{E_f}{E_m}$) after passing the adjacent in layers cuts, for photon fragments in jet samples (blue), non-fragments in jet samples (orange), and photon fragments in single-photon samples (green).

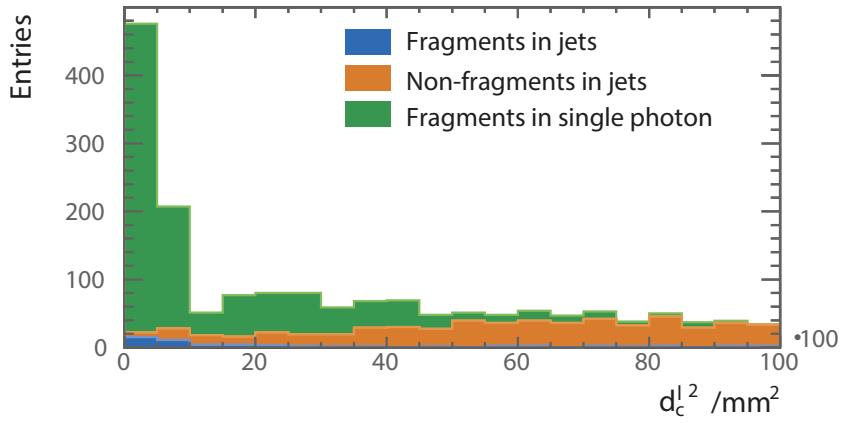


Figure 5.17: The stacked distributions of $d_c^{l\,2}$ after passing the adjacent in layers cuts and the energy comparison cuts, for photon fragments in jet samples (blue), non-fragments in jet samples (orange), and photon fragments in single-photon samples (green).

Photon fragment recovery	Cuts
Adjacent in layers	yes
Energy comparison	$\frac{E_f}{E_m} \leqslant 0.1$
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	$\frac{N_{cone}}{N_f} \geqslant 0.5$

Table 5.4: The cuts for merging photon fragment in the HCAL to the main photon in the ECAL.

5.6 Photon splitting algorithm

During the event reconstruction, it is possible that photons are accidentally merged if they are spatially close. Hence photon splitting algorithm addresses this issue and tries to split merged photons.

If a photon has the topology of multiple spatially closed photons, the parent photon will be split into several daughter photons. Extra care are taken if the parent photon is close to a track projection onto the front of the ECAL. Table 5.5 lists the cuts used in the algorithm.

The algorithm works as follows. If an energetic photon is identified, the 2D Peak Finding algorithm will be used to identify EM showers in the parent photon. If energy of the parent photon is bigger than E_{c1}, and the energy of the 2^{nd} energetic EM shower is bigger than E_{c2}, the parent photon will be split into daughter photons according to the number of EM showers identified by the 2D Peak Finding algorithm.

The values of E_{c1} and E_{c2} depend on whether the parent photon is close to a track projection onto the front of the ECAL. The algorithm demands higher values of E_{c1} and E_{c2}, 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 parent photon centroid position. If there is no nearby tracks to the parent photon, E_{c1} is set to 10 GeV and E_{c2} is set to 1 GeV. If there is one nearby track, E_{c1} is set to 10 GeV and E_{c2} is set to 5 GeV. If there is more than one nearby track, 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 parent photon, should be fewer than five as one reconstructed photon is unlikely to be merged from more than four photons.

Photon splitting	Cuts
Cuts	$E > E_{c1}$, $E_{p2} > E_{c2}$, $N_p < 5$
E_{c1} and E_{c2} values	
0 track nearby	$E_{c1} = 10\,\text{GeV}$, $E_{c2} = 1\,\text{GeV}$
1 track nearby	$E_{c1} = 10\,\text{GeV}$, $E_{c2} = 5\,\text{GeV}$
> 1 tracks nearby	$E_{c1} = 20\,\text{GeV}$, $E_{c2} = 10\,\text{GeV}$

Table 5.5: Cuts used in the photon splitting algorithm.

5.7 Photon reconstruction performance

Three different versions of the PandoraPFA are used to demonstrate the improvement of the photon reconstruction performance:

1. with no stand-alone photon reconstruction algorithms;
2. with a stand-alone photon reconstruction algorithm from PandoraPFA version 1; and
3. with full photon algorithms described above, incorporated in PandoraPFA version 3;

Without photon reconstruction algorithms, PandoraPFA applies a simple photon identity test at the end of the reconstruction. In PandoraPFA version 1, there is a rudimentary photon reconstruction algorithm. PandoraPFA version 3 contains all the photon algorithms developed in this chapter.

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

5.7.1 Improvement over no stand-alone photon algorithms

The improvement in the photon reconstruction is demonstrated using MC samples with two photons. The two-photon samples were generated with an uniform distribution in the solid angle of the first photon, and an uniform distribution in the opening angle between the photon pair. Events are discarded if there is a photon converting to electron pairs in the tracking detector or a photon escapes the detector undetected. The events are further restricted to the photon depositing energies in barrel and endcap regions only to avoid the barrel/endcap overlap region. Events were reconstructed using the nominal ILD detector model.

Figure 5.18 shows the average number of reconstructed photons as a function of true distance separation between two photons, using two-photon samples with photon energies of 500 GeV and 50 GeV, reconstructed with and without photon algorithms. For the reconstruction without the photon algorithms, the number of photon fluctuates

between 1 and 1.5 for a distance separation of 0 to 30 mm between two photons. For the reconstruction with the photon algorithms, two photons start to be resolved at a distance separation of 10 mm between two photons, and fully resolved at a distance separation of 20 mm. The average number of reconstructed photon is 2 at a distance separation of 20 mm.

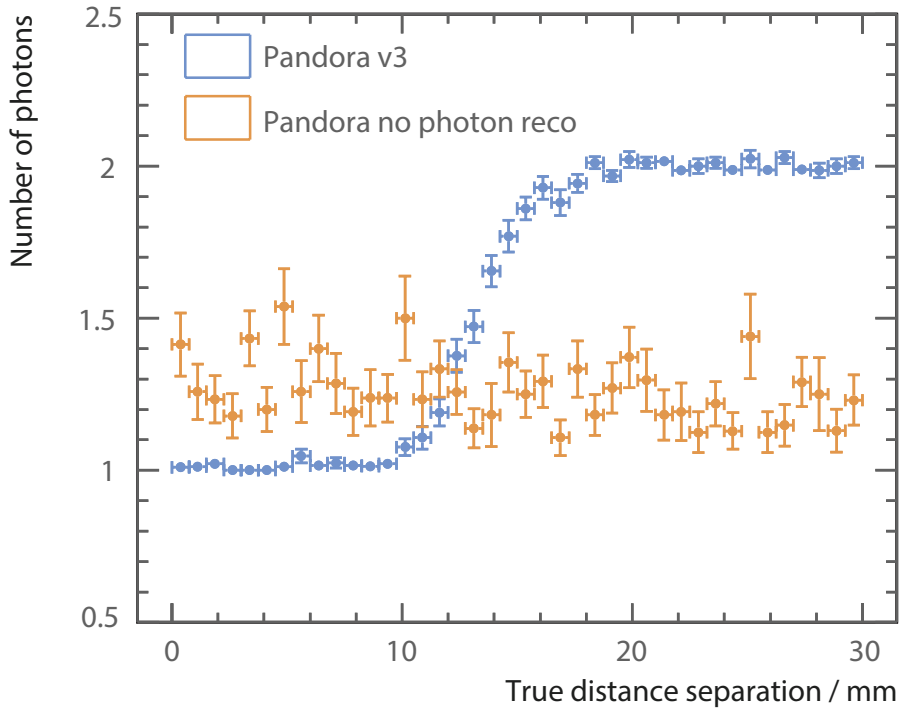


Figure 5.18: Average number of reconstructed photons using two-photon samples with photon energies of 500 GeV and 50 GeV, without (orange) and with (blue) photon algorithms, as a function of the true distance separation between two photons.

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 $e^+e^- \to Z'Z'$ events 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 jet energy distribution. The total jet energies are sampled at centre-of-mass energies of 91, 200, 360 and 500 GeV.

As shown in figure 5.19, jet energy resolutions are much better at $\sqrt{s} = 360$ GeV and 500 GeV for the reconstruction 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 $\sqrt{s} = 91$ GeV and 200 GeV,

jet energy resolutions are worse for the reconstruction with photon algorithms, because photon algorithms are developed and optimised with jet environments at a high centre-of-mass energy of 500 GeV.

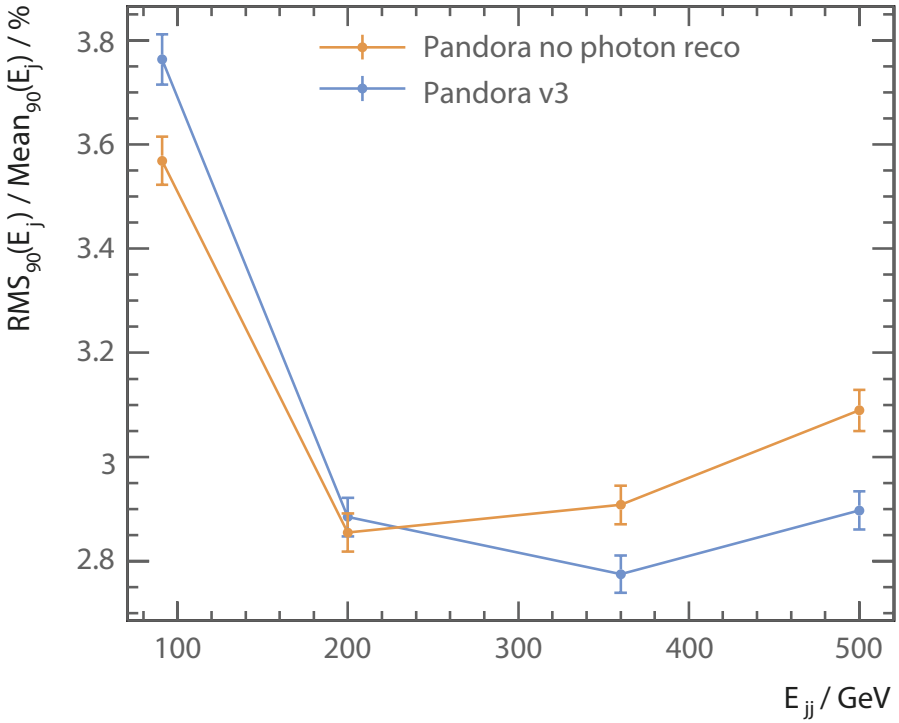


Figure 5.19: Jet energy resolutions as a function of the total jet energy using $e^+e^- \to Z'Z'$ events where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$ at barrel region. The orange and bottom points represent the reconstruction without and with photon algorithms, respectively.

The impact of photon algorithms on the jet energy resolution was studied using the same jet samples with the perfect photon reconstruction, which identifies photons by associating calorimeter hits using the truth information. The photon confusion term which are defined as the quadrature differences of the jet energy resolutions between a non-cheated reconstruction and a perfect photon reconstruction is a measure of the failure in the photon reconstruction. Table 5.6 lists the photon confusion terms as a function of the centre-of-mass energies for the jet sample. The photon confusion terms, except at $\sqrt{s} = 91$ GeV, have been reduced to 0.9% for the reconstruction with photon algorithms.

5.7.2 Improvement over PandoraPFA version 1

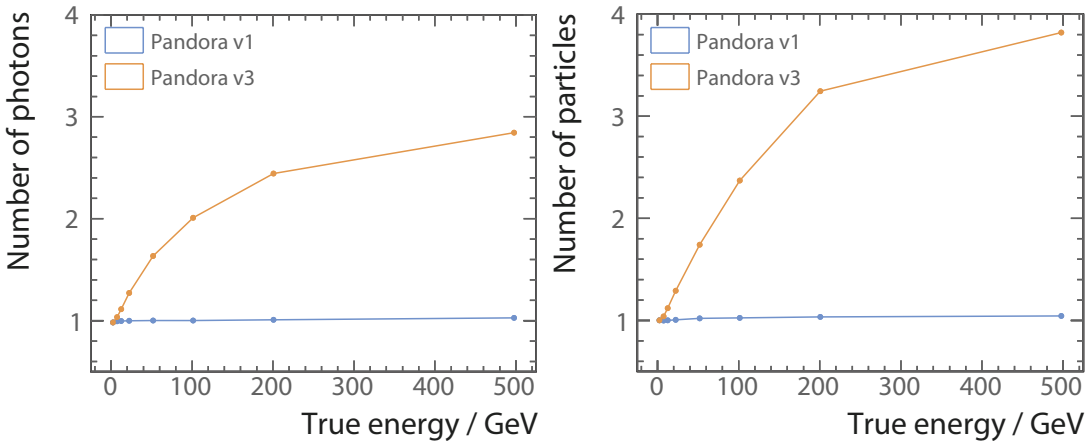
This section reviews the photon reconstruction improvement from PandoraPFA version 1 to version 3 using single-photon, two-photon, and jet samples.

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 full photon algorithms	1.4%	0.9%	0.9%	0.9%

Table 5.6: Photon confusion terms as a function of total jet energies in the $e^+e^- \to Z'Z'$ events where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$ at barrel region for reconstruction with and without photon algorithms.

The single-photon MC samples were generated with an uniform distribution in the solid angle of the photon. Other samples were generated and simulated in the same way as previously. The same pre-selection as previously was applied to the single-photon and two-photon samples.

Figure 5.20a shows the average number of reconstructed photons as a function of the true photon energies, using single-photon samples. Figure 5.20b shows the average number of reconstructed particles as a function of the true photon energies. Drastic decrease in the number of fragments can be seen in both plots for photons up to 500 GeV.



(a) **(b)**

Figure 5.20: Average numbers of: a) photons; and b) particles, as a function of their true energies using single-photon samples. For both figures, the top orange and bottom blue points are reconstructed with PandoraPFA version 1 and version 3, respectively.

Figure 5.21 shows the numbers of reconstructed photons and particles as a function of the true distance separation between the two photons using a two-photon sample with photon energies of 500 GeV and 50 GeV. The average numbers of photons and particles for reconstruction with PandoraPFA version 3 are both below 2.05 at a distance separation of 30 mm, which is significantly lower than the numbers for reconstruction

with PandoraPFA version 1. For reconstruction with PandoraPFA version 3, two photons start to be resolved at a distance separation of 10 mm and fully resolved at a distance separation of 20 mm.

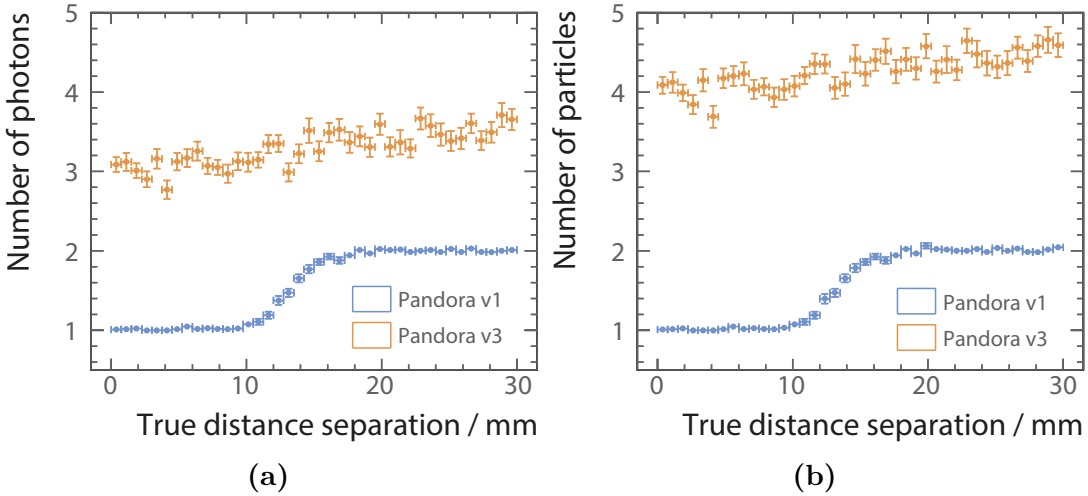


Figure 5.21: Average numbers of: a) photons; and b) particles, as a function of the true distance separation between two photons, using two-photon samples with photon energies of 500 GeV and 50 GeV. For both figures, the top orange and bottom blue points represent the reconstruction with PandoraPFA version 1 and version 3, respectively.

Another metric to reflect the improvement in photon reconstruction is the fraction of the fragment energy to the total energy in an event. In a two-photon sample, the fragment energy is defined as the total energy of particles excluding the two most energetic photons. Shown in figure 5.22, using two-photon sample with photon energies of 500 GeV and 50 GeV, a reduction in fragment energy can be seen in PandoraPFA version 3. For the photon reconstruction in PandoraPFA version 3, the average fragment energy fraction is below 0.1% up to a distance separation of 30 mm, while around 5% energy would be in fragments for the reconstruction with PandoraPFA version 1.

The reduction in the fragments leads to a small improvement in the jet energy resolutions at a high jet energy. Using the same jet sample as in the previous section, the jet energy resolutions are better at total jet energies of 360 and 500 GeV with the photon reconstruction in PandoraPFA version 3, as shown in figure 5.23.

5.7.3 Performance of individual photon algorithms

Two-photon events with photon energies of 500 GeV and 500 GeV are used to show the incremental improvement of the performance of individual photon algorithms. Figure 5.24 shows the average number of reconstructed particles as a function of true distance

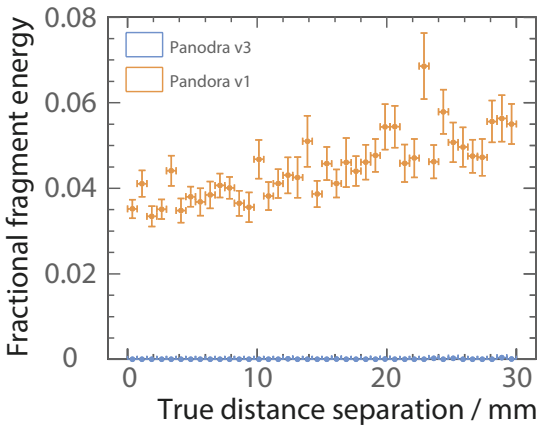


Figure 5.22: Average fraction of fragments energy to the total energy in the event, as a function of the true distance separation between two photons, using a two-photon sample with photon energies of 500 GeV and 50 GeV. The top orange and bottom blue points represent the reconstruction with PandoraPFA version 1 and version 3 respectively.

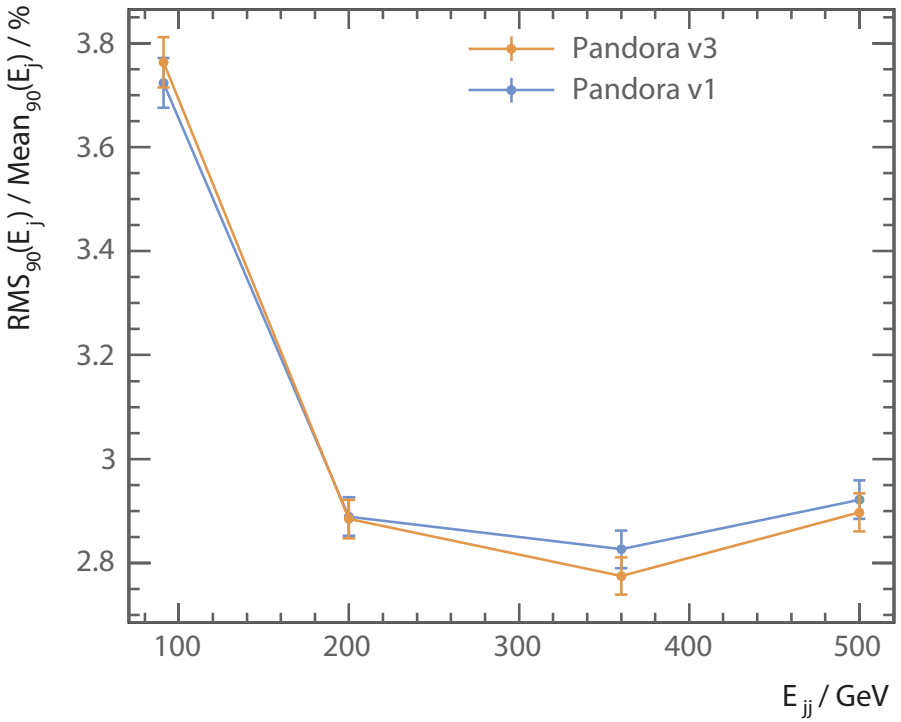


Figure 5.23: Jet energy resolutions as a function of the total jet energy using $e^+e^- \to Z'Z'$ events where $Z' \to u\bar{u}/d\bar{d}/s\bar{s}$, at barrel region. The top orange and bottom blue points represent the reconstruction with PandoraPFA version 1 and version 3.

separation between two photons, 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 fragments is reduced significantly when it is compared with photon reconstruction in PandoraPFA version 1 (red). With the additional fragment removal algorithm in the HCAL (green), the number of fragments is reduced further. At a distance separation of 40 mm, there is on average less than 0.05 fragment per photon pair for the reconstruction with fragment removal algorithms in the ECAL and the HCAL (green).

The introduction of the photon reconstruction and photon splitting algorithm (blue) makes the photon pair resolve at a much shorter distance separation between two photons. Photon pairs start to be resolved at a distance separation of 5 mm and fully resolved at a distance separation of 15 mm when reconstructed with full photon algorithms in PandoraPFA version 3.

5.7.4 Photon reconstruction performance with PandoraPFA version 3

Figure 5.25 shows the average single photon reconstruction efficiency as a function of the true photon energies, using single-photon samples. In a single-photon sample, an event can have an efficiency of 1 or 0 depending on whether there is a reconstructed photon corresponding to the true photon. The average single photon reconstruction efficiency is above 98% for photons with energies above 2GeVand above 99.5% for photons with energies above 100 GeV. The low efficiency in the first bin in figure 5.25a for photon energies in the range from 0 to 0.25 GeV is because photon reconstruction algorithms do not attempt to reconstruct photons with energies below 0.2 GeV.

Figure 5.26 shows the average numbers of reconstructed photons and particles as a function of the true distance separation between two photons using a two-photon sample with photon energies of 500 GeV and 500 GeV. A good match between the number of photons and the number of particles is achieved. The average numbers of photons and particles are both fewer than 2.05 for a distance separation beyond 20 mm, less that 1 fragment produced per 20 events.

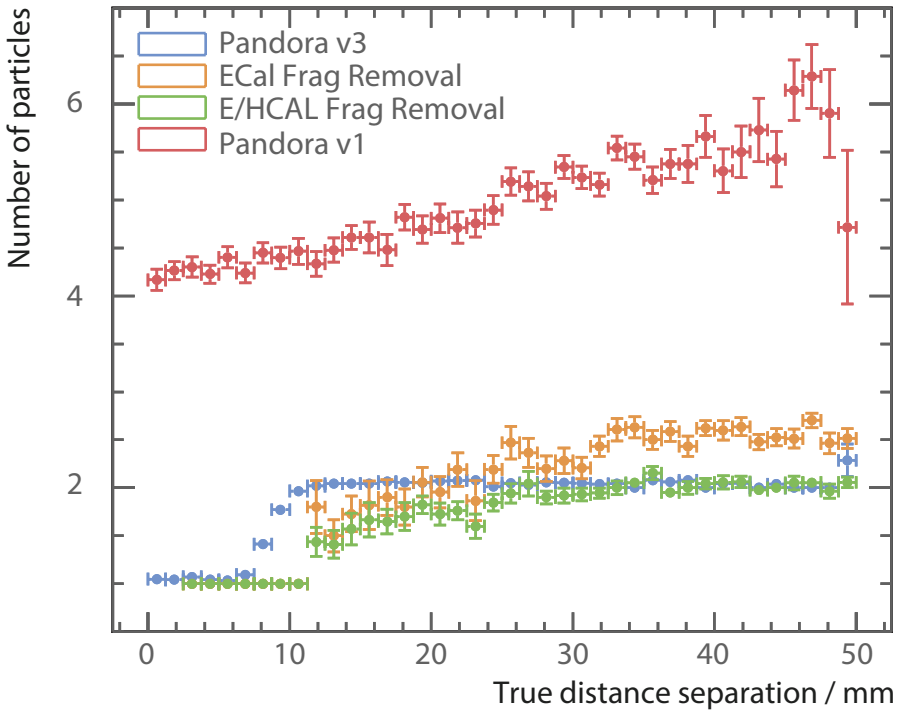
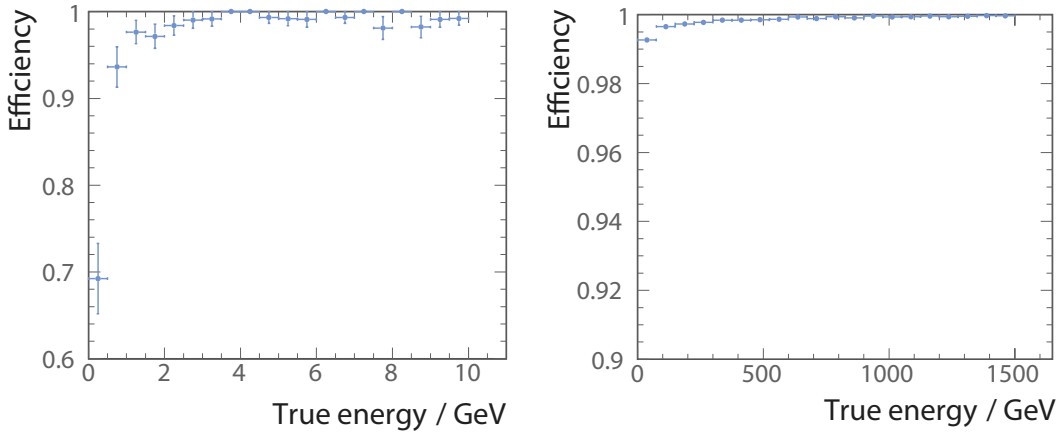


Figure 5.24: Average numbers of photons, as a function of the true distance separation between two photons, using a two-photon sample with photon energies of 500 GeV and 500 GeV. The blue, orange, green, and red points represent the reconstruction with PandoraPFA version 3, the reconstruction with fragment 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.



(a) **(b)**

Figure 5.25: Single photon reconstruction efficiency as a function of true photon energies, using single-photon samples for: a) the low photon energy regime; and b) the high photon energy regime.

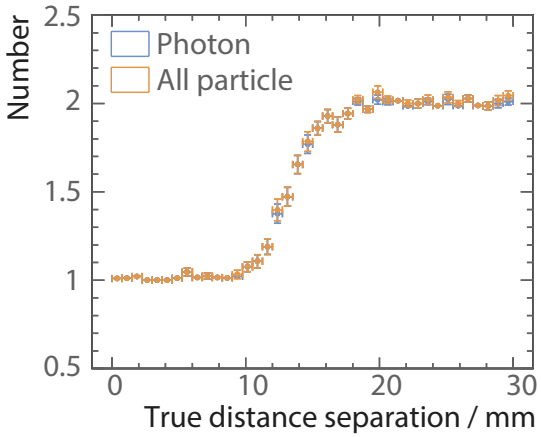


Figure 5.26: Average numbers of reconstructed photon (blue) and particle (orange), as a function of the true distance separation between two photons, using two photons of 500 GeV and 50 GeV per event samples.

Figure 5.27 shows the average numbers of photon reconstructed using two-photon samples as a function of the true distance separation between two photons 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 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 a distance separation of 6 mm, which is about one ECAL cell length in the simulated nominal ILD detector. In contrast, photon pairs with different energies, for example 500 GeV−50 GeV and 100 GeV−10 GeV pairs, start to be resolved at a distance separation of 10 mm, 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 5.27, at a distance separation of 20 mm, 500 GeV−500 GeV photon pairs are fully resolved, whereas approximately only 60% of 10 GeV−10 GeV photon pairs are resolved.

5.8 Summary

Using the ILD detector model, the single photon reconstruction efficiency is above 98% for photons with energies above 2GeVand above 99.5% for photons with energies above 100 GeV.

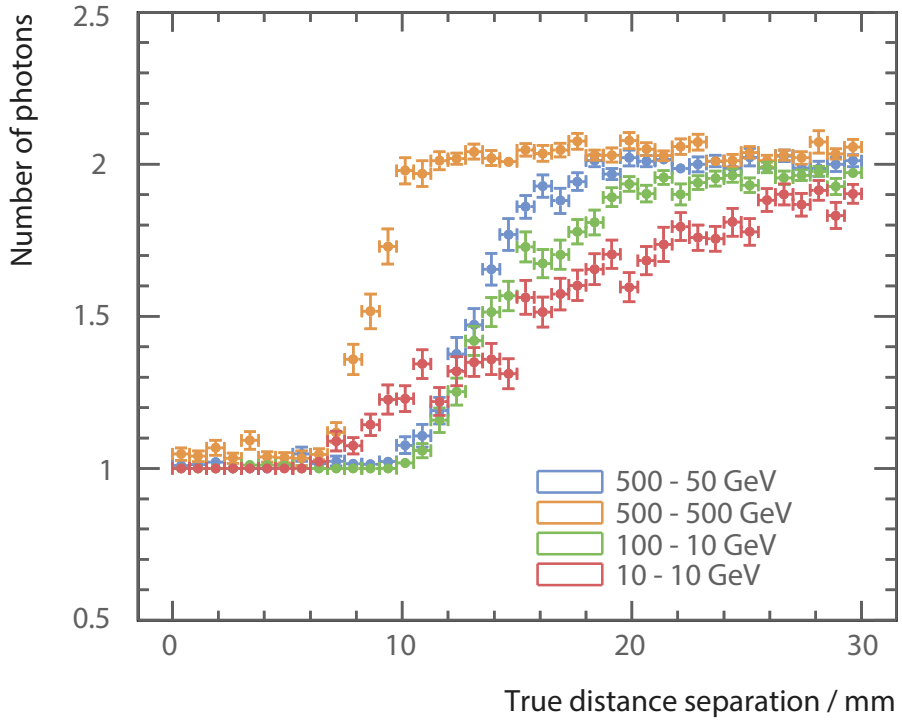


Figure 5.27: Average numbers of reconstructed photons for four different photon pairs: 500 GeV−50 GeV (blue), 500 GeV−500 GeV (orange), 100 GeV−10 GeV (green), and 10 GeV−10 GeV (red), as a function of the true distance separation between two photons.

The number of photon fragments produced have been greatly reduced. Using a two-photon sample with photon energies of 500 GeV and 50 GeV, the average numbers of photons and particles beyond a distance separation of 20 mm are both less than 2.05, where the true value is 2.

The minimal distance separation of resolved photon pairs is reduced to 6 mm for two photons with the same energy and 10 mm for two photons with different energies.

The jet energy resolution has been improved for high centre-of-mass energies. The photon confusion terms, except at $\sqrt{s} = 91$ GeV, have been reduced to 0.9%.

Chapter 6

Tau Lepton Decay Mode 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, 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 the $H \to \tau^+\tau^-$ decay process [93].

Since the tau lepton has a mean decay lifetime of 290 fs [26], only tau decay products will be detected in the calorimeters and tracking detectors of the ILD detector. Therefore, the performance of the calorimetric and tracking systems determines the ability to reconstruct tau lepton decay products and to classify different tau decay modes.

The main challenge in classifying tau lepton hadronic decay modes is the reconstruction and separation of spatially close photons. For tau leptons with energies above tens of GeV, visible decay products are highly boosted. Consequently electromagnetic showers from photons from π^0 decays often overlap in the ECAL. Reconstructing these photons as separate entities requires a good photon reconstruction. Hence, the photon reconstruction algorithms described in chapter 5 are used in this study. This chapter presents a study of the classification of tau decay modes in a highly granular linear collider detector.

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 WHIZARD [50]. TAUOLA [54] was used to describe the tau lepton decays with correct spin correlations of the tau decay products. The study was focused on separating tau decay modes. Hence, beam effects were not included, 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 were reconstructed with iLCSoft version v01-17-07 [94] and PandoraPFA version 3 [66] using the photon reconstruction algorithms described 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 $\pi^-\pi^0\nu_\tau$. The bottom half of the event shows a tau lepton decaying into $\pi^+\pi^-\pi^-\pi^0\nu_\tau$.

6.2.1 Tau decay modes

To study the main decay modes of the tau lepton, decay modes with branching ratios above 2% are classified. The classified seven decay modes cover 92.58 % of the tau decay branching fraction [3]. The seven tau decay modes, their branching ratios, and detectable final states are listed in table 6.1.

In the $\tau^- \to \rho(\pi^-\pi^0)\nu_\tau$ decays, 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 and a_1 mesons are 775.11 ± 0.34 MeV and 1230 ± 40 MeV respectively [3].

6.2.2 Tau selection

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

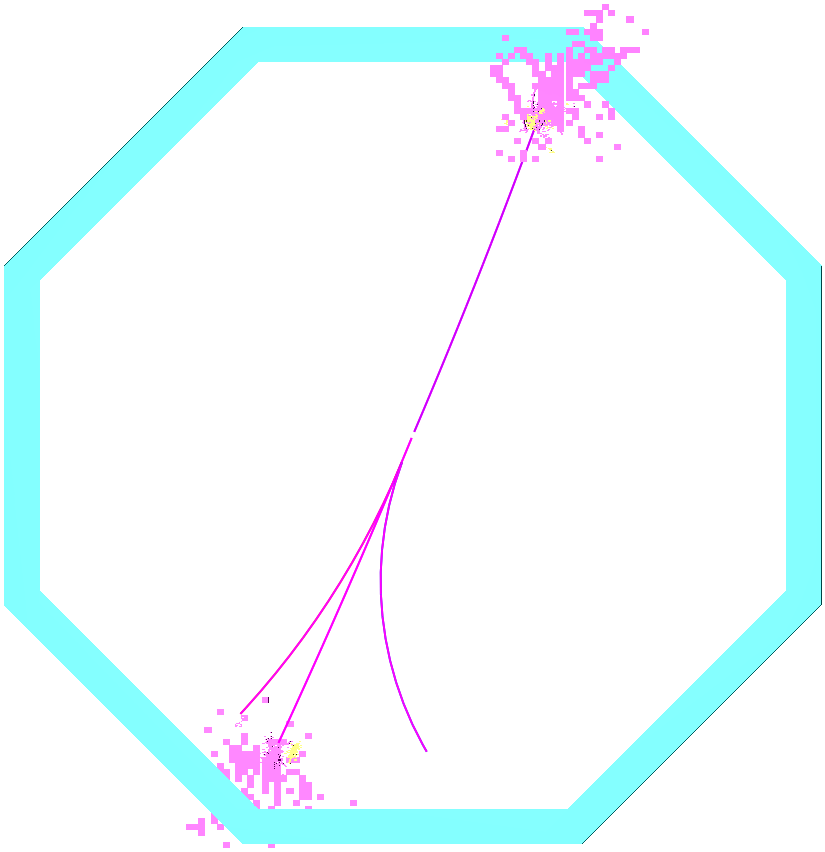


Figure 6.1: An event display of a simulated $e^+e^- \to \tau^+\tau^-$ event using the ILD detector model. The top half of the event shows a tau lepton decaying into $\pi^-\pi^0\nu_\tau$. The bottom half of the event shows a tau lepton decaying into $\pi^+\pi^-\pi^-\pi^0\nu_\tau$. Purple lines represent $\pi^\pm$ tracks in the tracking detectors. Purple squares represent calorimeter hits of the $\pi^\pm$ hadronic showers in the ECAL and the HCAL. Yellow squares 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.

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 the seven major tau decay modes with the largest branching ratios. Values are taken from [3].

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 particles are aligned to. The principle thrust axis vector, $\hat{t}$, is determined by maximising the thrust [95], 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 product are in one set and particles with a negative sign of the scalar product are in another set.

6.3 Pre-selection

For the purpose of this study, three pre-selection cuts, based on the MC information of the particles, are used. The cuts and the fractions of tau decays passing the cuts are listed in table 6.2.

Since this study is focused on photon reconstruction in the ECAL to classify tau decay modes, the tau decays with one or more photons converting to electron pairs in the tracking detector are not considered.

The focus of the study is on high energy tau decays. Thus tau decays with the total visible energy (i.e. not accounting for neutrinos) of tau decay products, E_{vis}^{MC}, less than 5 GeV are not considered.

Lastly, tau decays are discarded when the tau decay products are in the region between barrel and endcap parts of the calorimeters as there is a degradation in the particle reconstruction efficiency in this region. Tau decays with the generated polar angle of the tau lepton in the region $0.6 < \left|\theta_\tau^{MC}\right| < 0.9$ rad are not considered.

Table 6.2 shows fractions of tau decays passing successive cuts for different tau decay final states. As expected, the cut on photon conversions 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 effect on the leptonic decay modes with two neutrinos in the final states. The cut on the tau polar angle affects different tau decay modes equally.

Final state	No photon conversion	$E_{vis}^{MC} > 5\,\text{GeV}$	$\left\lvert\theta_{\tau}^{MC}\right\rvert$
$e^{-}\overline{\nu}_{e}\nu_{\tau}$	100.0%	84.7%	66.2%
$\mu^{-}\overline{\nu}_{\mu}\nu_{\tau}$	100.0%	85.2%	66.7%
$\pi^{-}\nu_{\tau}$	100.0%	88.3%	60.9%
$\pi^{-}\pi^{0}\nu_{\tau}$	77.1%	76.9%	61.9%
$\pi^{-}\pi^{0}\pi^{0}\nu_{\tau}$	61.3%	61.2%	50.5%
$\pi^{+}\pi^{-}\pi^{-}\nu_{\tau}$	100.0%	100.0%	78.0%
$\pi^{+}\pi^{-}\pi^{-}\pi^{0}\nu_{\tau}$	77.0%	77.0%	61.8%

Table 6.2: Fractions of tau decays passing successive pre-selection cuts for different tau decay final states.

6.4 MVA variables

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

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	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 different types of final state particles. There are five particle number 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^-}). Here muon and electron ID is provided by the PandoraPFA output.

Figure 6.2a shows distributions of numbers of reconstructed charged particles for different tau decay modes. Over 98% of $\tau^- \to \pi^- \nu_\tau$ decays have exactly one reconstructed charged particle, and approximately 95% of $a_1(\pi^+\pi^-\pi^-)$ decays give exactly three reconstructed charged particles. Figure 6.2b and figure 6.2c show distributions of numbers of reconstructed muons and electrons respectively for different tau decay modes. Here 99% of $\tau^- \to \mu^- \overline{\nu}_\mu \nu_\tau$ decays produce exactly one reconstructed muon and 99% of $\tau^- \to e^- \overline{\nu}_e \nu_\tau$ decays have one reconstructed electron. Figure 6.2d shows distributions of numbers of reconstructed photons for different tau decay modes, which distinguishes hadronic tau decay 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 $a_1(\pi^-\pi^0\pi^0)$ 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 reconstructed charged pions (m_{π^-}).

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

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 ($\tilde{E}_e$); the normalised total energy of photons ($\tilde{E}_\gamma$); and the normalised total energy of charged pions ($\tilde{E}_{\pi^-}$). All variables are normalised relative 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

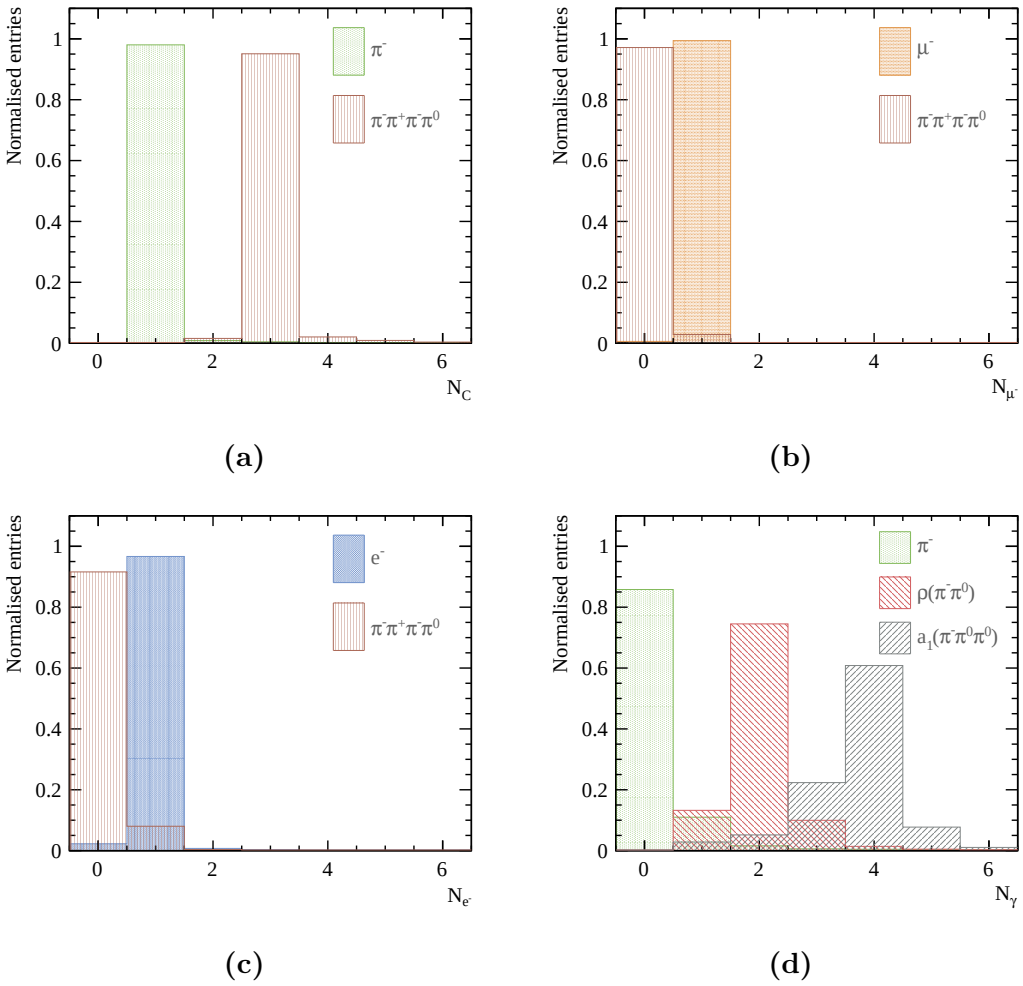
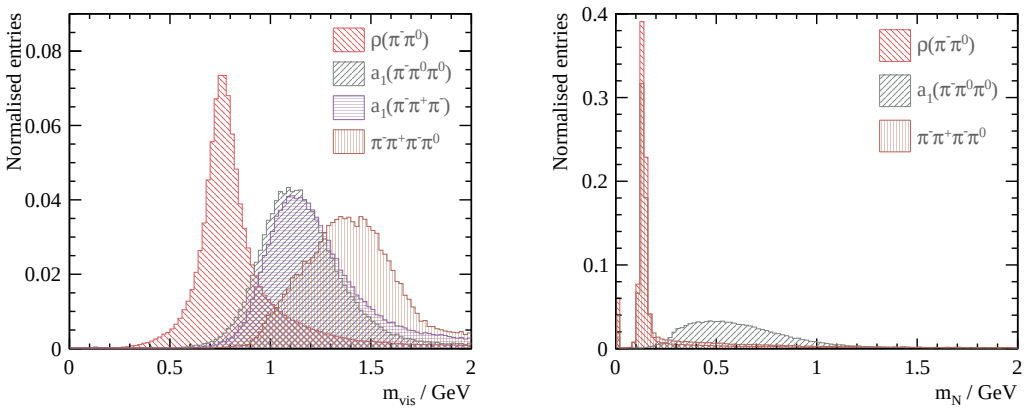


Figure 6.2: Distributions of the number of reconstructed particles of different types: a) charged particles (N_C); b) muons (N_μ); c) electrons (N_e); and d) photons (N_γ). The particle ID information comes from the output of the PandoraPFA reconstruction. The area under the curve for each decay mode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

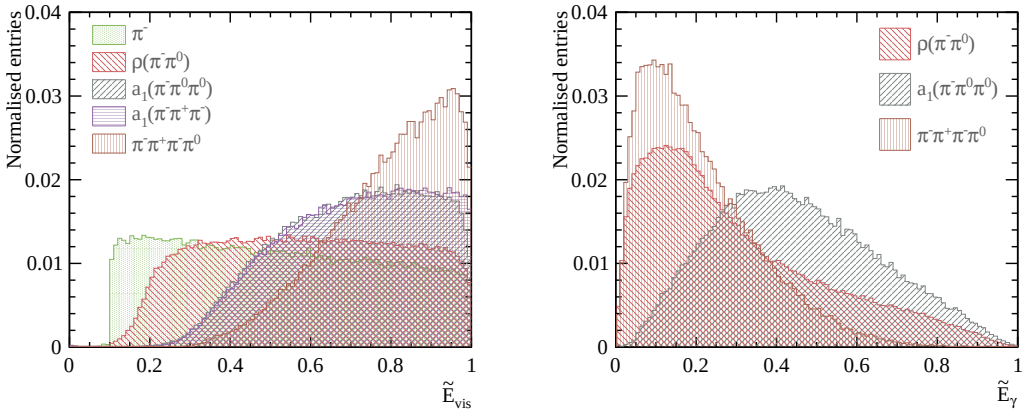
and E_τ is the energy of the associated tau lepton. E_τ is obtained from the generated value, i.e. 50 GeV.

Figure 6.4a and figure 6.4b show distributions of normalised energies of all reconstructed particles and photons respectively for different tau decay modes. Differences in the distributions for different tau decay modes can be seen. The cut-off at 0.1 for the $\tilde{E}_{vis}$ distribution is due to the pre-selection of $E_{vis}^{MC} > 5\,\mathrm{GeV}$, which approximately corresponds to $\tilde{E}_{vis} > 0.1$.



(a) **(b)**

Figure 6.3: Distributions of the invariant mass of a) all particles (m_{vis}); and b) all 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.



(a) **(b)**

Figure 6.4: Distributions of the normalised energies of: a) all reconstructed particles ($\tilde{E}_{vis}$); and b) all reconstructed 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.

6.4.4 Calorimetric energy variables

Two calorimetric energy variables 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 improve the identification of electron and muon decay modes. Figure 6.5 show distributions of E_C^{ECAL}/E_C and E^{ECAL}/E for different decay modes. An electron typically deposits over 95% of its energy in the ECAL, and a muon typically deposits 5% to 25% of its energy in the ECAL. The difference between E_C^{ECAL}/E_C and E^{ECAL}/E is that photons and neutral hadrons, which deposit most of their energies in the ECAL and in the HCAL respectively, are not included in the calculation of E_C^{ECAL}/E_C.

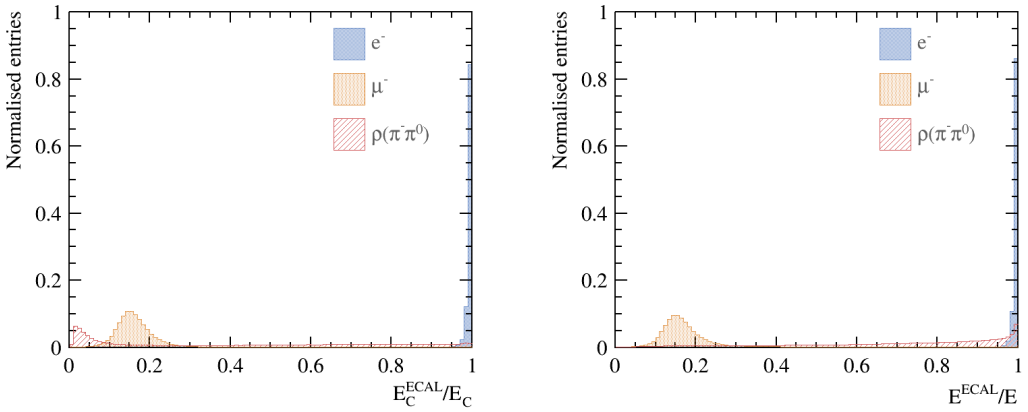


Figure 6.5: Distributions of the fractions of the energy deposited in the ECAL divided by the energy deposited in the ECAL and HCAL: a) where only calorimetric deposits associated with charged particles are considered (E_C^{ECAL}/E_C); and b) for all reconstructed particles (E^{ECAL}/E). The particle ID information comes from the output of the PandoraPFA reconstruction. The area under the curve for each decay mode is normalised to unity.

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

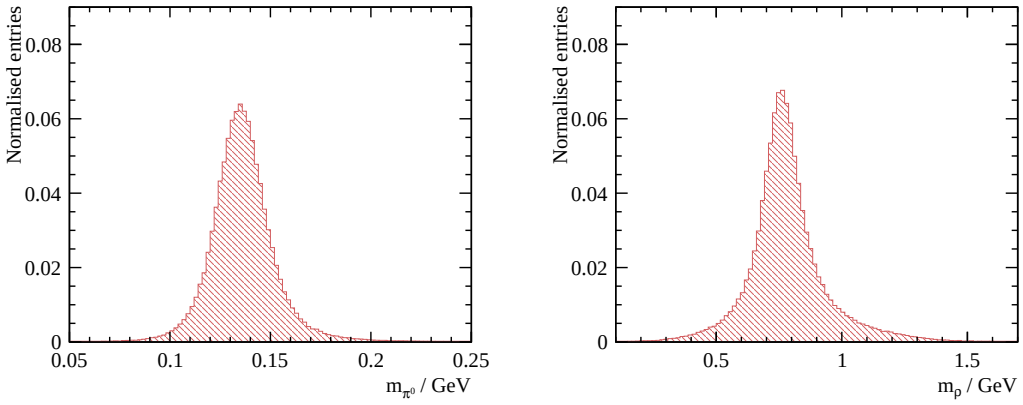
By utilising the photon identification potential of the highly granular ECAL, the identification of the $\rho(\pi^-\pi^0)$ and $a_1(\pi^-\pi^0\pi^0)$ decay modes is enhanced by reconstructing the ρ and a_1 invariant masses. For decays with at least one charged pion and one photon, the reconstruction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

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^-; m_ρ and m_{π^0} are the respective true masses of ρ and π^0; and σ_ρ and σ_{π^0} are the assumed mass resolutions. Figure 6.6 shows the reconstructed invariant mass distributions for π^0 and ρ in the $\rho(\pi^-\pi^0)$ decay mode obtained by selecting reconstructed particles using the MC truth information. A mass resolution of 20% is an good approximation for the invariant masses of π^0 and ρ and it is used for σ_ρ and σ_{π^0}.

The particle ID of charged pions and photons comes from the output of the PandoraPFA reconstruction. The χ^2 function works naturally if there are two reconstructed photons in a decay. If there are more than two photons in a decay, all combinations of two photons are considered and the combination with the smallest value of χ^2 is chosen. If there is only one photon in a decay, the second term in the equation 6.2 is dropped and m_{tot} is the total invariant mass of one photon and one π^-.



(a) **(b)**

Figure 6.6: Reconstructed invariant mass distributions of the 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 area under the curve is normalised to unity.

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

$$\chi^2 = \left(\frac{m_{tot} - m_{a_1}}{\sigma_{a_1}}\right)^2 + \left(\frac{m_{\gamma_1\gamma_2} - m_{\pi^0}}{\sigma_{\pi^0}}\right)^2 + \left(\frac{m_{\gamma_3\gamma_4} - m_{\pi^0}}{\sigma_{\pi^0}}\right)^2, \tag{6.3}$$

where the ρ mass has been replaced 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 χ^2 function. 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.

If there are at least four photons in a decay, all combinations of four photons are considered and the combination with the smallest value of χ^2 is chosen. If there are three photons in a decay, the last term in the equation 6.3 is dropped and the χ^2 function becomes:

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

where m_{tot} is the invariant mass of the charged pion and three photons and $m_{\gamma_1\gamma_2}$ is the invariant mass of two photons. Combinations of photons are iterated.

If there are only two photons in a decay, either the reconstruction failed to reconstruct one photon pair or the reconstruction fails to reconstruct both photon pairs. Hence two χ^2 functions are used and the one with the smallest value is chosen. The first function is

$$\chi^2 = \left(\frac{m_{tot} - m_{a_1}}{\sigma_{a_1}}\right)^2 + \left(\frac{m_{\gamma_1\gamma_2} - m_{\pi^0}}{\sigma_{\pi^0}}\right)^2, \tag{6.5}$$

where m_{tot} is the invariant mass of the charged pion and two photons and $m_{\gamma_1\gamma_2}$ is the invariant mass of two photons. The second function is

$$\chi^2 = 2\left(\frac{m_{tot} - m_{a_1}}{\sigma_{a_1}}\right)^2, \tag{6.6}$$

where m_{tot} is the invariant mass of the charged pion and two photons. The factor of 2 in equation 6.6 is for the direct comparison in the values of χ^2 with equation 6.5.

If there is only one photon in a decay, m_{tot} is the invariant mass of the charged pion and the photon.

From the ρ invariant mass reconstruction of the $\rho(\pi^-\pi^0)$ decay hypothesis test, two variables are obtained and 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). If there is only one photon, $m_{\pi^0}^{(\rho)}$ is set to 0.

From the a_1 invariant mass resonance reconstruction, three variables are obtained and used in the MVA classification: the a_1 mass ($m_{a_1}^{reco} \equiv m_{tot}$ in equation 6.3), the first π^0 mass ($m_{\pi^0}^{(a_1)} \equiv m_{\gamma_1\gamma_2}$ in equation 6.3), and the second π^0 mass ($m_{\pi^0}^{*(a_1)} \equiv m_{\gamma_3\gamma_4}$ in equation 6.3). If there are three reconstructed photons, $m_{\pi^0}^{*(a_1)}$ is set to 0. If there are two reconstructed photons, depending on whether equation 6.5 or equation 6.6 gives the smallest χ^2 value, either $m_{\pi^0}^{*(a_1)}$ is set to 0 or both $m_{\pi^0}^{(a_1)}$ and $m_{\pi^0}^{*(a_1)}$ are 0. If there is only one reconstructed photon, both $m_{\pi^0}^{(a_1)}$ and $m_{\pi^0}^{*(a_1)}$ are 0.

Figure 6.7 shows the distributions of $m_{a_1}^{reco}$ under $a_1(\pi^-\pi^0\pi^0)$ decay mode hypothesis test for three 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.

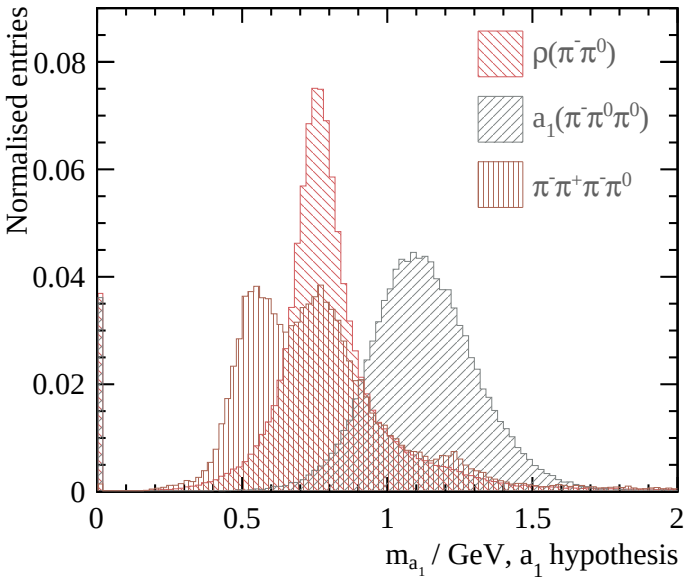


Figure 6.7: Reconstructed invariant mass distributions for a_1 ($m_{a_1}^{reco}$), reconstructed under the $a_1(\pi^-\pi^0\pi^0)$ decay mode hypothesis for three different tau decay modes. The area under the curve is normalised to unity.

6.4.6 Separating electrons from charged pions

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

An electron develops a characteristic EM shower in the ECAL, while 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 5.3.3.

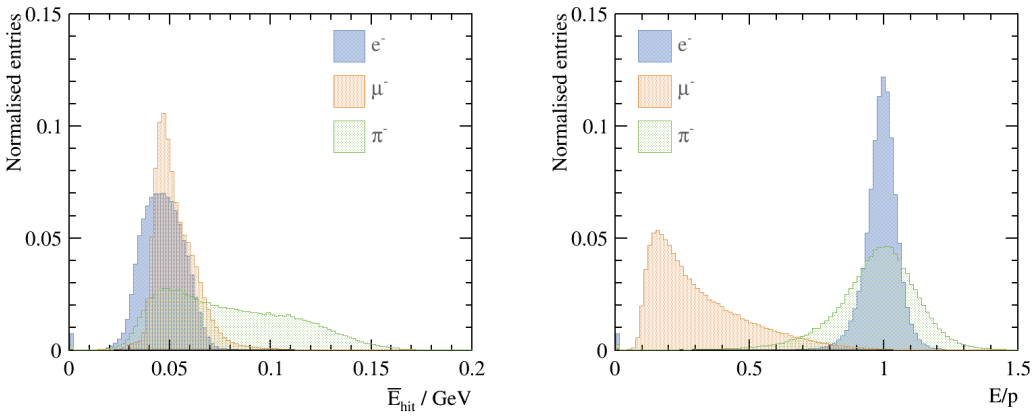
The calorimeter hit information is also used to differentiate an EM shower from a hadronic shower. Two variables used in the MVA classification are: the average energy of a calorimeter hit ($\bar{E}_{hit}$), which is the total energy deposited in the ECAL and HCAL divided by the number of the ECAL and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

Finally the track is used to provide the consistency check of the track momentum with the total energy in the ECAL and HCAL for charged particles. The variable used in the MVA classification is the energy in the ECAL and HCAL divided by the track momentum (E/p).

Figure 6.8 show distributions of the average energy of a calorimeter hit ($\bar{E}_{hit}$), and the energy in the ECAL and HCAL divided by the track momentum (E/p). Differences between the $e^-\overline{\nu}_e\nu_\tau$, $\mu^-\overline{\nu}_\mu\nu_\tau$, and $\pi^-\nu_\tau$ decay modes can be seen.

6.5 MVA classification

The MultiClass class of the Tmva package [80] 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 events, 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 of the classifier are listed in table 6.4, where an explanation of the parameters can be found in section 4.5.6.



(a) **(b)**

Figure 6.8: Distributions of: a) the average energy of a calorimeter hit ($\bar{E}_{hit}$); and b) the energy in the ECAL and HCAL divided by the track momentum (E/p). The area under the curve is normalised to unity.

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 of the Boosted Decision Tree with Gradient boost MultiClass classifier used for the tau decay mode classification.

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 tau decay modes 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 (bold numbers in table 6.5) 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 decay mode i and the N_i^{MC} is the total number of true tau decays for decay mode i.

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. 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 correct 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 the additional variables (section 6.4.6) dedicated to the separation between e^- and π^-.

For the separation of tau hadronic decay modes, 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 $\tau^- \to a_1(\pi^-\pi^0\pi^0)\nu_\tau$ and $\tau^- \to \rho(\pi^-\pi^0)\nu_\tau$ 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 can lead to the misclassification between the $\tau^- \to \rho(\pi^-\pi^0)\nu_\tau$ and $\tau^- \to \pi^-\nu_\tau$ decays. The misclassification between one-prong decays, as well as between three-prong decays, is highlighted in table 6.5.

A high correct classification rate is achieved for all seven classified tau decay modes. The leptonic decay modes have correct classification rates over 99.5%. For the hadronic tau decay modes, classification rates of 89.7% or above are achieved.

6.7 Electromagnetic calorimeter optimisation

The performance of 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, 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} using $e^+e^- \to \tau^+\tau^-$ samples at a centre-of-mass energy of 100 GeV. ClosestHitDistance is varied amongst values of 5, 10, 20, 30, 40, and 50 mm. The overall tau hadronic decay correct classification rate, ε_{had}, is the weighted average correct classification efficiency, defined as:

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

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 sizes. As expected, for larger cell sizes the distance metric for merging photons becomes larger.

Figure 6.9 shows ε_{had} as a function of ECAL cell sizes for four different centre-of-mass energies. The efficiency ε_{had} decreases with an increase of the 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

ECAL cell length / mm	ClosestHitDistance / mm
3	5
5	10
7	10
10	10
15	20
20	20

Table 6.6: Optimised values of ClosestHitDistance parameters in PhotonFragmentRemoval algorithm as a function of the ECAL square cell sizes.

pair from π^0 decay. The reduction in the ability to separate photon pairs leads to a degradation of the classification performance.

The efficiency ε_{had} decreases with the increasing ECAL cell sizes. The change in the ECAL cell size will change the ECAL 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 a larger ECAL cell size.

Table 6.7 lists the achieved hadronic decay mode separation as measured by ε_{had} with 3 mm and 20 mm ECAL cells 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 3 mm and 20 mm ECAL cell lengths for four different centre-of-mass energies.

Figure 6.10 shows the correct classification efficiencies (ε_i) for five tau hadronic decay modes as a function of the ECAL square cell sizes for four different centre-of-mass energies. For the ECAL square cells, the cell size is the squared of the cell length. The tau decay mode correct classification efficiencies generally decrease with an increase of centre-of-mass energies and an increase of ECAL cell sizes.

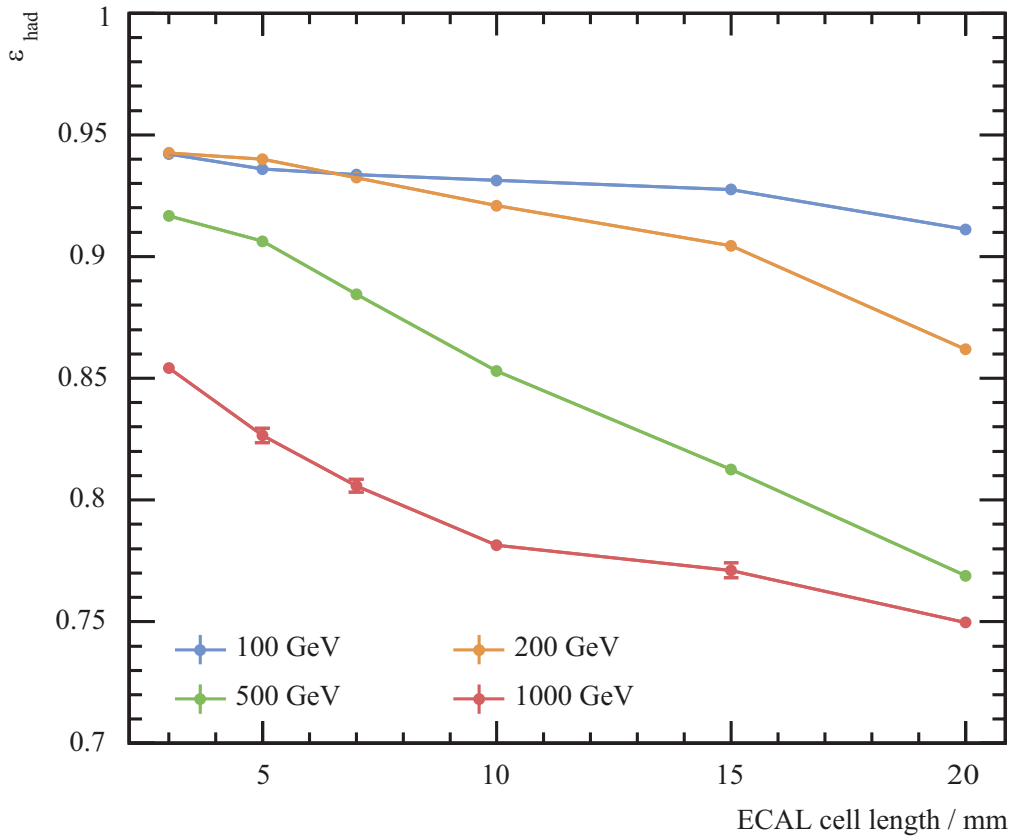


Figure 6.9: The weighted average tau hadronic decay correct classification efficiency, ε_{had}, as a function of the ECAL cell sizes for four different centre-of-mass energies. The blue, orange, green, and red points show ε_{had} at centre-of-mass energies of 100, 200, 500, and 1000 GeV, respectively.

For the $\tau^- \to \rho(\pi^-\pi^0)\nu_\tau$ decay mode, the efficiency at $\sqrt{s} = 1000\,\text{GeV}$ 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 $\tau^- \to \text{a}_1(\pi^-\pi^0\pi^0)\nu_\tau$ decay mode at the same centre-of-mass energy. For this reason, ε_{had} gives a better picture of true performance.

For the $\tau^- \to \text{a}_1(\pi^-\pi^0\pi^0)\nu_\tau$ decay mode, the loss of efficiency with an increasing ECAL cell size and an increasing centre-of-mass energy is most significant compared to other decay modes. With more photons in the final state, it is the most challenging decay mode to reconstruct and thus most sensitive to the change in cell sizes and centre-of-mass energies.

For the $\tau^- \to \text{a}_1(\pi^+\pi^-\pi^-)\nu_\tau$ 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 detector performance, which is not affected by different ECAL cell sizes.

A previous study [96] on the tau decay mode classification was performed using the ILD detector on $\pi^-\nu_\tau$, $\rho\nu_\tau$, and $a_1(\pi^-\pi^0\pi^0)\nu_\tau$ decay modes. Samples used were $e^+e^- \to Z^0 \to \tau^+\tau^-(\gamma)$ at 250 GeV. Pre-selection required that events with photon converted to electron pairs in the tracking detector were discarded. GARLIC [97, 98] photon reconstruction was used to reconstruct photons in the events. The main differences between the previous analysis and the current one are the pre-selection cuts, the photon reconstruction algorithm, and the number of classified decay modes. Table 6.8 lists the correct classification rates of $\pi^-\nu_\tau$, $\rho\nu_\tau$, and $a_1(\pi^-\pi^0\pi^0)\nu_\tau$ decay modes for current analysis with PandoraPFA photon reconstruction at a centre-of-mass energies of 200 GeV and the previous analysis using GARLIC photon reconstruction. Similar correct classification rates are achieved for $\pi^-\nu_\tau$, $\rho\nu_\tau$, and $a_1(\pi^-\pi^0\pi^0)\nu_\tau$ decay modes.

Decay mode	PandoraPFA $\sqrt{s} = 200$ GeV	GARLIC $\sqrt{s} = 250$ GeV
$\pi^-\nu_\tau$	$94.6\% \pm 0.1\%$	$96.8\% \pm 0.2\%$
$\rho\nu_\tau$	$94.2\% \pm 0.1\%$	$90.5\% \pm 0.2\%$
$a_1(\pi^-\pi^0\pi^0)\nu_\tau$	$88.6\% \pm 0.2\%$	$91.1\% \pm 0.4\%$

Table 6.8: Correct classification rates of $\pi^-\nu_\tau$, $\rho\nu_\tau$, and $a_1(\pi^-\pi^0\pi^0)\nu_\tau$ decay modes for current analysis with PandoraPFA photon reconstruction and previous analysis using GARLIC photon reconstruction. Values for the previous analysis are taken from [96].

6.8 Summary

For the ILC at $\sqrt{s} = 250$ GeV or CLIC at $\sqrt{s} = 350$ GeV, an ECAL size of 10 mm or fewer is sufficient to achieve a ε_{had} of 92%. For a linear collider operating at a centre-of-mass energy above a few hundred GeV, such as the ILC at $\sqrt{s} = 500$ GeV or CLIC at $\sqrt{s} = 1.4$ TeV or 3 TeV, it is preferable to have a small ECAL cell size, i.e. 3 mm, for the best tau decay mode separation, as ε_{had} decreases drastically with an increasing ECAL cell size.

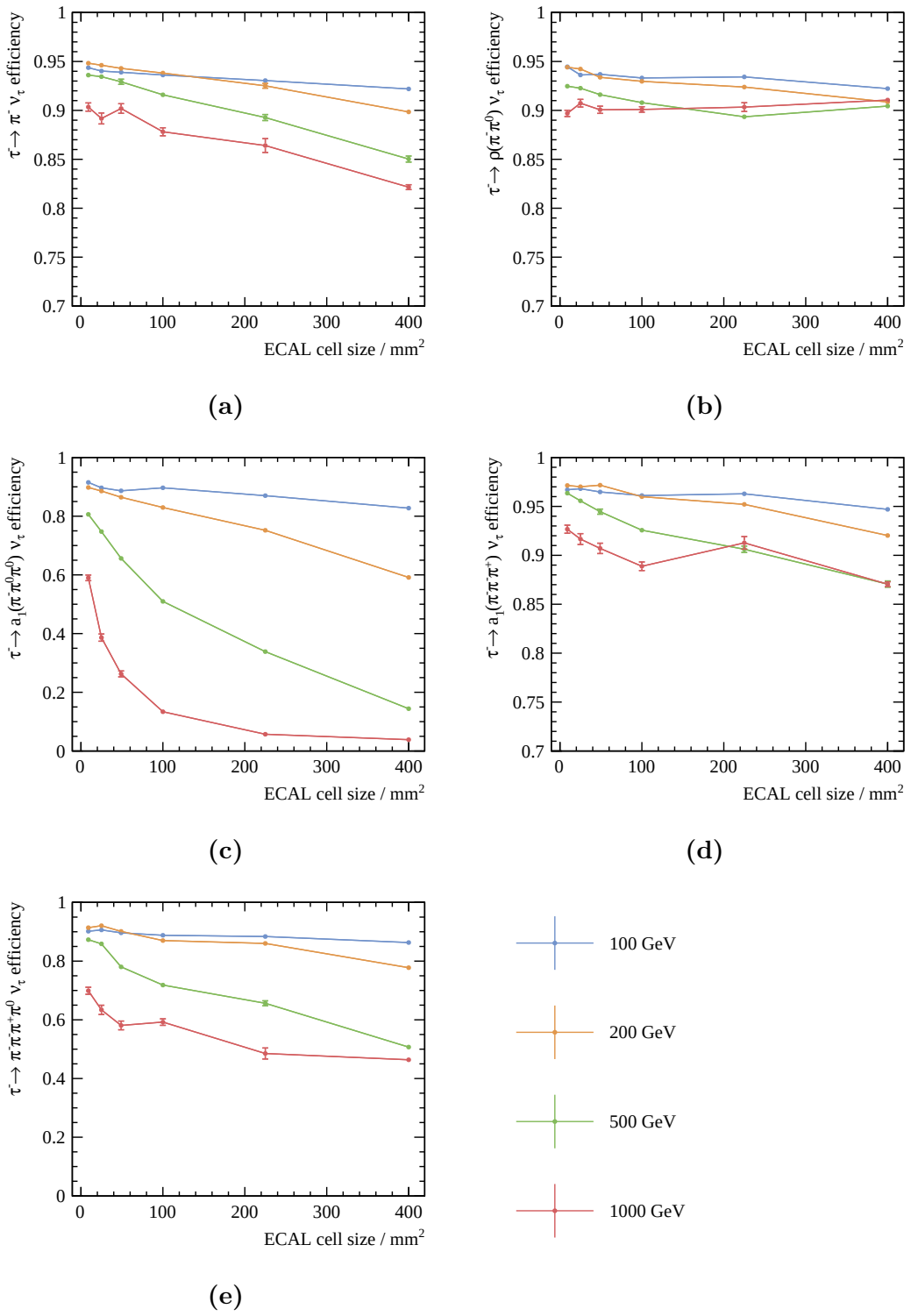


Figure 6.10: The correct classification efficiencies as a function of the ECAL square cell sizes for: a) $\tau^- \to \pi^- \nu_\tau$ decays; b) $\tau^- \to \rho(\pi^-\pi^0)\nu_\tau$ decays; c) $\tau^- \to a_1(\pi^-\pi^0\pi^0)\nu_\tau$ decays; d) $\tau^- \to a_1(\pi^+\pi^-\pi^-)\nu_\tau$ decays; and e) $\tau^- \to \pi^+\pi^-\pi^-\pi^0\nu_\tau$ decays. Results are shown for centre-of-mass energies of 100, 200, 500, and 1000 GeV.

Chapter 7

Tau Pair Polarisation Correlation

'Where can I find a man who has forgotten words so I can have a word with him?'

— Zhuang Zi, 369 BC – 286 BC

This chapter 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 spin-0 scalar Higgs boson can decay to $\tau_{\mathrm{L}}^{+}\tau_{\mathrm{L}}^{-}$ or $\tau_{\mathrm{R}}^{+}\tau_{\mathrm{R}}^{-}$, whereas the spin-1 Z boson can decay to $\tau_{\mathrm{L}}^{+}\tau_{\mathrm{R}}^{-}$ or $\tau_{\mathrm{R}}^{+}\tau_{\mathrm{L}}^{-}$, where L, R denote the tau lepton helicities. 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.

Here a proof-of-principle analysis is performed to reconstruct the polarisation correlation of the tau pairs in the $\mathrm{Z} \to \tau^{+}\tau^{-}$ process, where both $\tau^{\pm} \to \pi^{\pm}\nu_{\tau}$. $\tau^{\pm} \to \pi^{\pm}\nu_{\tau}$ decay mode is selected using the tau decay mode classifier developed in chapter 6, which utilises the photon reconstruction in a highly granular calorimeter developed in chapter 5. The charged pion decay mode is chosen because the correlation between $E_{\pi^{+}}/E_{\tau^{+}}$ and $E_{\pi^{-}}/E_{\tau^{-}}$ is very different if the parent boson is a Z or a H boson, as suggested in figure 2.5.

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 $\tau^{\pm} \to \pi^{\pm}\nu_{\tau}$ 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 7.1 shows the main steps in this proof-of-principle demonstration of the tau pair polarisation correlations using $e^+e^- \to ZZ \to \tau^+\tau^-qq$ events, where both $\tau^\pm \to \pi^\pm \nu_\tau$.

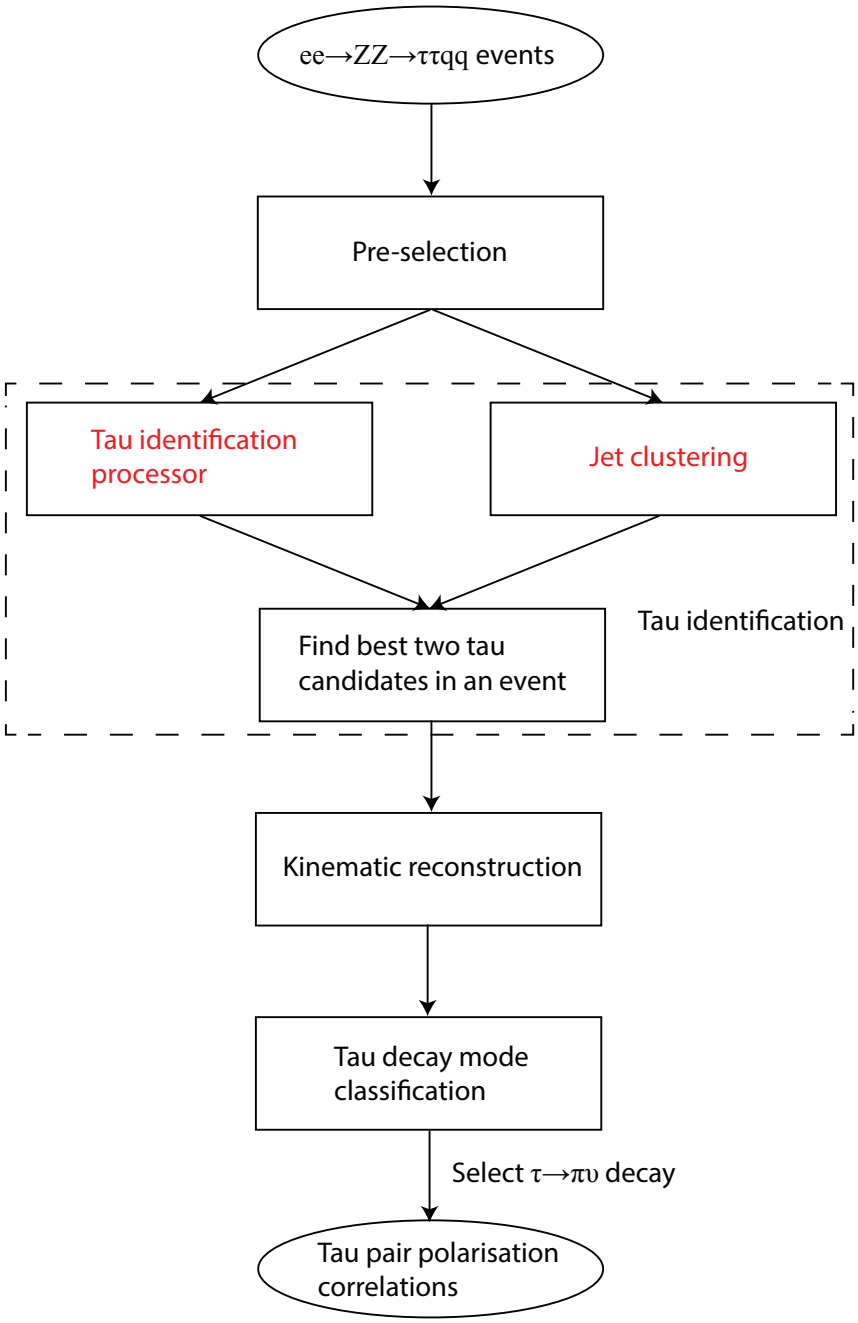


Figure 7.1: Main steps in the proof-of-principle demonstration of the tau pair polarisation correlations.

7.1 Event generation and simulation

For this proof-of-principle study, $e^+e^- \to ZZ \to \tau^+\tau^-qq$ events were generated at a centre-of-mass energy of 350 GeV using WHIZARD [50] generator without ISR. TAUOLA [54] was used to describe the tau lepton decay with correct spin correlations of the tau decay products. Beam effects, such as the initial state radiation and the beam induced background, were not included. The $e^+e^- \to ZZ \to \tau^+\tau^-qq$ events were simulated using the ILD detector model as described in chapter 4.

7.2 Event reconstruction

Events were reconstructed with iLCSoft version v01-17-07 [94] and PandoraPFA version 3 [66], using the photon reconstruction algorithms described in chapter 5. Seven tau decay modes defined in section 6.1 were considered in this analysis: $\tau^- \to e^-\overline{\nu}_e\nu_\tau$; $\tau^- \to \mu^-\overline{\nu}_\mu\nu_\tau$; $\tau^- \to \pi^-\nu_\tau$; $\tau^- \to \rho(\pi^-\pi^0)\nu_\tau$; $\tau^- \to a_1(\pi^-\pi^0\pi^0)\nu_\tau$; $\tau^- \to a_1(\pi^+\pi^-\pi^-)\nu_\tau$; and $\tau^- \to \pi^+\pi^-\pi^-\pi^0\nu_\tau$.

7.3 Pre-selection

Two pre-selection cuts defined in section 6.3 are used: the tau decay products with photon conversion to electron pairs in the tracking detector are not considered; and tau decays with the generated polar angle of the tau lepton in the region $0.6 < \left|\theta_\tau^{MC}\right| < 0.9$ rad are not considered.

7.4 Tau identification

The $e^+e^- \to ZZ \to \tau^+\tau^-qq$ final state contains two tau leptons and two quark jets. Identifying the tau decay products in $e^+e^- \to ZZ \to \tau^+\tau^-qq$ events is challenging as a low multiplicity quark-jet could be topologically similar to a tau hadronic decay. Hence the tau decay product identification processor and the jet clustering algorithm are both used to find tau decay products. Figure 7.2 shows an event display of a $e^+e^- \to ZZ \to \tau^+\tau^-qq$ event. The two brown cones indicate the tau decay products found by the tau identification processor. The four blue cones indicate the four jets found by the jet clustering algorithm. Particles associated with the tau decay products found by the tau identification processor are different to the particles associated with jets found by the jet clustering algorithm.

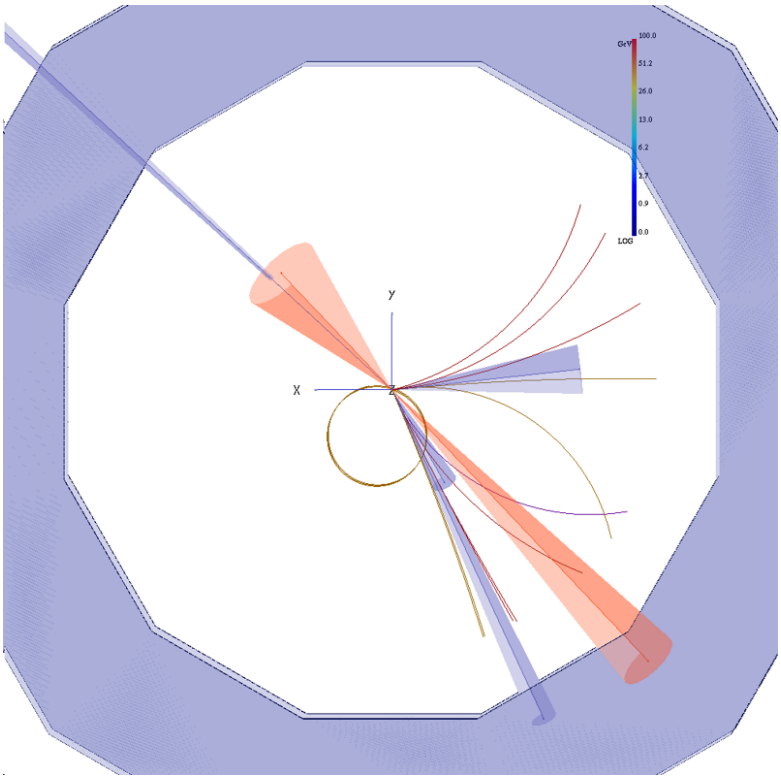


Figure 7.2: An event display of a $e^+e^- \to ZZ \to \tau^+\tau^-qq$ event. The two brown cones indicate the tau decay products found by the tau identification processor. The four blue cones indicate the four jets found by the jet clustering algorithm. The blue outer region shows the HCAL.

7.4.1 Tau identification processor

The IsolatedTauIdentifer processor is a modified version of the tau decay product identification software used in the double Higgs analysis (section 8.3.2). The processor identifies high transverse momentum (p_T) 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 referred to as the search cone, which contains potential 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 number of charged particles in the temporary tau candidate, N_{X^+}, should be one or three. The invariant mass of the temporary tau candidate, m_c, should be less than 3 GeV. The temporary tau candidate also needs to pass the isolation condition to be identified as a tau candidate, which requires the opening angle between the temporary search cone and the 2^{nd} closest charged particle, $\theta^c_{X'^+}$, is larger than 0.6 rad.

The iterative particle addition procedure stops when the cone opening angle, θ_S, is larger than $\cos^{-1}(0.99)$. If multiple temporary tau candidates of the same tau seed pass

the isolation condition, the one with the smallest opening angle is chosen to form the final tau candidate. Table 7.1 lists the parameters of the IsolatedTauIdentifer processor.

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	$\theta^c_{X'^+} > 0.6\,\mathrm{rad}$

Table 7.1: Optimised parameters of the modified IsolatedTauIdentifer.

The event is discarded if the IsolatedTauIdentifer processor finds fewer than two tau candidates. If more than two tau candidates are found, the best two are selected by choosing the tau candidates that gives the smallest value of the function in equation 7.1, resulting in well reconstructed $\mathrm{Z} \to \mathrm{qq}$ decays:

$$\left(m_{\mathrm{qq}} - m_{\mathrm{Z}}\right)^2 + \left(E_{\mathrm{qq}} - \frac{\sqrt{s}}{2}\right)^2, \tag{7.1}$$

where $\sqrt{s}$ is the centre-of-mass energy; variable m_{Z} is the mass of Z boson from reference [3]; m_{qq} is the invariant mass of particles that do not belong to the two tau candidates; and E_{qq} is the total energy of particles that do not belong to the two tau candidates. This function is considered for all pairs of tau candidates. In the generated $\mathrm{e^+e^-} \to \mathrm{ZZ} \to \tau^+\tau^-\mathrm{qq}$ events, the energy of each Z boson is half of the centre-of-mass energy. The invariant mass of two quarks from the Z decay should be close to Z mass.

7.4.2 Jet clustering

The Durham algorithm (section 4.4.2) was run in the exclusive mode to force the reconstructed particles in $\mathrm{e^+e^-} \to \mathrm{ZZ} \to \tau^+\tau^-\mathrm{qq}$ events into exactly four jets. The two tau candidate jets are identified by selecting two jets that gives the smallest value of the function in equation 7.1. Here m_{qq} is the invariant mass of particles that are not in the two tau candidates jets and E_{qq} is the total energy of particles that are not in the two tau candidates jets. Other variables are defined in the same way as in section 7.4.1.

7.4.3 Selecting the best tau candidates in an event

If both IsolatedTauIdentifer processor and the jet clustering method find two tau candidates, the best pair of tau candidates should result in well reconstructed $Z \to qq$ decays, defined by:

$$\left|m_{qq} - m_Z\right| < 10\,\text{GeV}, \quad \left|E_{qq} - \frac{\sqrt{s}}{2}\right| < 10\,\text{GeV}. \tag{7.2}$$

The selection of best pair of tau candidates in an event proceeds as follows:

1. if pairs of tau candidates from both IsolatedTauIdentifer processor and the jet clustering method satisfy equation 7.2, the pair with the smallest value of equation 7.1 is selected;
2. otherwise, the pair of tau candidates that satisfies equation 7.2 is selected;
3. otherwise, if one jet from the jet clustering is close to the beam pipe and there are exactly two tau candidates obtained from IsolatedTauIdentifer, then the two tau candidates from IsolatedTauIdentifer are selected. This choice is motivated by the fact that 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 jet reconstruction;
4. otherwise, the two jets with the fewest number of particles are selected.

Table 7.2 lists the numbers of events with unmatched and matched taus identified in each of four steps in 107 manually scanned $e^+e^- \to ZZ \to \tau^+\tau^-qq$ events. An event with matched tau requires that MC particles contributing the most to identified best tau candidates are both true taus. The opposite is an event with unmatched taus. In 107 events, 93 events have matched taus and 14 events have unmatched taus.

7.5 Kinematic reconstruction of tau energy

The pion energy fractions, E_{π^+}/E_{τ^+} and E_{π^-}/E_{τ^-}, are the appropriate kinetic variables to illustrate the tau pair polarisation correlation in $e^+e^- \to ZZ \to \tau^+\tau^-qq$ events, where both taus decay $\tau^\pm \to \pi^\pm \nu_\tau$, motivated in section 2.9. To obtain E_{π^+}/E_{τ^+} and E_{π^-}/E_{τ^-}, the energies of the taus, E_{τ^+} and E_{τ^-}, are required. Energies of the taus are also required for the calculation of the energy variables used in the tau decay mode classification.

Step	Unmatched taus	Matched taus
1	0	37
2	2	26
3	0	5
4	12	25
total	14	93

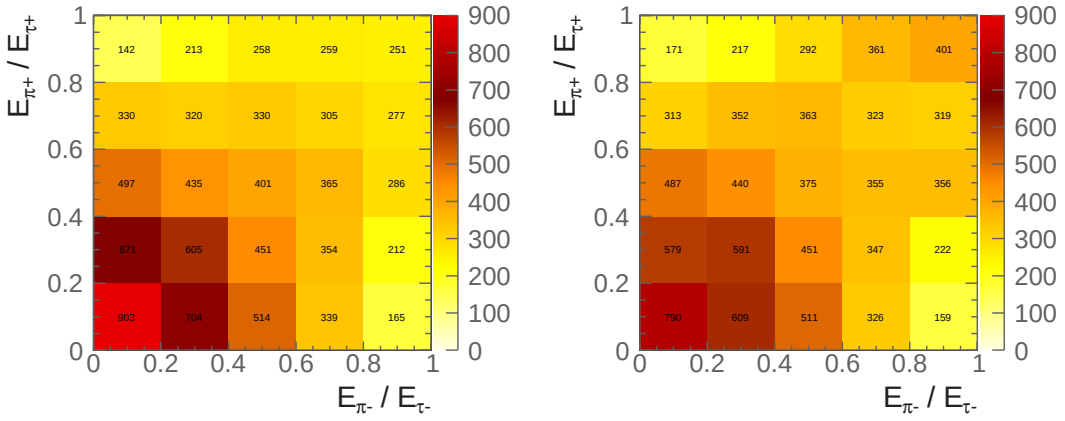
Table 7.2: Numbers of events with unmatched and matched taus identified in each of four steps to select best tau candidates in 107 $e^+e^- \to ZZ \to \tau^+\tau^-qq$ events.

In the laboratory frame, the energies of the taus from $Z \to \tau^+\tau^-$ can not be determined easily. However, in the Z rest frame, the energies of the taus are simply the half of the energy of the Z, i.e. $E'_{\tau^+} = E'_{\tau^-} = \frac{1}{2}E'_Z$.

Because the energies of the taus can be obtained in the Z rest frame, kinematic variables used in tau decay mode classification, for example $\tilde{E}_C$, are calculated in the Z rest frame as well. Therefore, tau decay products need to be boosted into the Z rest frame for the calculation of the kinematic variables.

To boost tau decay products into the Z rest frame, the four-momentum of the Z is needed. Since there are two Zs in the $e^+e^- \to ZZ \to \tau^+\tau^-qq$ event, $Z_{\tau\tau}$ refers to the Z decaying to a tau pair and Z_{qq} refers to the Z decaying to a quark pair. The four-momentum of the $Z_{\tau\tau}$ can be obtained from the recoil four-momentum of Z_{qq} against the centre-of-mass energy, where the four-momentum of Z_{qq} is measured directly from the particles that are not the tau candidates. The beam crossing angle of 14 mrad is taken in account.

The estimation of the four-momentum of $Z_{\tau\tau}$ is improved by correcting the energy of the $Z_{\tau\tau}$ to be half of the centre-of-mass energy, which improves the estimation of the tau energies. Figure 7.3 shows the two-dimensional distributions of E_{π^+}/E_{τ^+} as a function of E_{π^-}/E_{τ^-} from $Z \to \tau^+\tau^-$ decays where both $\tau^\pm \to \pi^\pm\nu_\tau$. $\tau^\pm \to \pi^\pm\nu_\tau$ decay mode is determined using the truth information. True Monte Carlo tau decay particles are used to generate distributions shown in the figures. Figure 7.3a shows the distribution without the improved four-momentum of $Z_{\tau\tau}$. Figure 7.3b shows the distribution with the improved four-momentum of $Z_{\tau\tau}$. A better match with the distributions obtained in the generator level study in figure 2.5a is achieved with the improved four-momentum of $Z_{\tau\tau}$, which motivates the correction of the $Z_{\tau\tau}$ energy to improve the measurement of the distributions of E_{π^+}/E_{τ^+} and E_{π^-}/E_{τ^-}.



(a) (b)

Figure 7.3: Two-dimensional distributions of E_{π^+}/E_{τ^+} as a function of E_{π^-}/E_{τ^-} from $Z \to \tau^+\tau^-$ decay where both $\tau^\pm \to \pi^\pm \nu_\tau$, with: a) the unimproved four-momentum of $Z_{\tau\tau}$; and b) the improved four-momentum of $Z_{\tau\tau}$. $\tau^\pm \to \pi^\pm \nu_\tau$ decay mode is determined using the truth information. True Monte Carlo tau decay particles are used to generate distributions in figures.

7.6 Tau decay mode classification

The tau decay mode classifier developed in chapter 6 can now be used to select the $\tau^\pm \to \pi^\pm \nu_\tau$ decay mode. In the classifier, 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.

7.7 Tau pair polarisation correlations

Figure 7.4 shows the one-dimensional distributions of E_{π^+}/E_{τ^+} and E_{π^-}/E_{τ^-} using generated $e^+e^- \to ZZ \to \tau^+\tau^-qq$ events. Both taus decay by $\tau^\pm \to \pi^\pm \nu_\tau$. $E_{\tau^\pm}$ is half of the energy of $Z_{\tau\tau}$ in $Z_{\tau\tau}$ rest frame. At the generator level, the shape of the distribution decreases gradually with the increasing $E_{\pi^\pm}/E_{\tau^\pm}$. The decreasing shape is largely preserved in the full detector simulation. In the full detector simulation, events with $E_{\pi^\pm}/E_{\tau^\pm}$ close to 0, which fall in the first bin in the figures, are not reconstructed correctly. When the energies of the tau are mostly carried away by the neutrinos, it is difficult to identify low-energy charged pions as tau decay products.

Figure 7.5 shows the two-dimensional distributions of E_{π^+}/E_{τ^+} versus E_{π^-}/E_{τ^-}, using generated $e^+e^- \to ZZ \to \tau^+\tau^-qq$ events. Both taus decay by $\tau^\pm \to \pi^\pm \nu_\tau$. Figure 7.5a shows the two-dimensional tau decay product energy fraction distribution obtained with

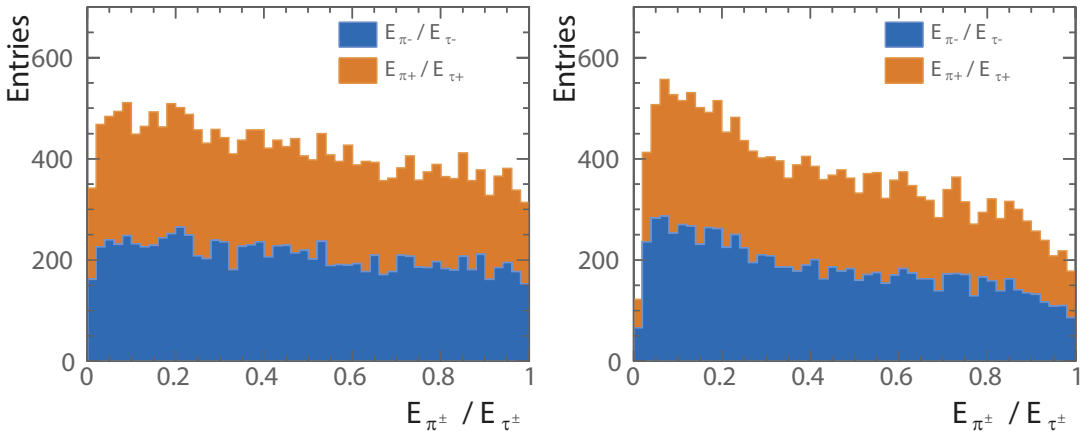


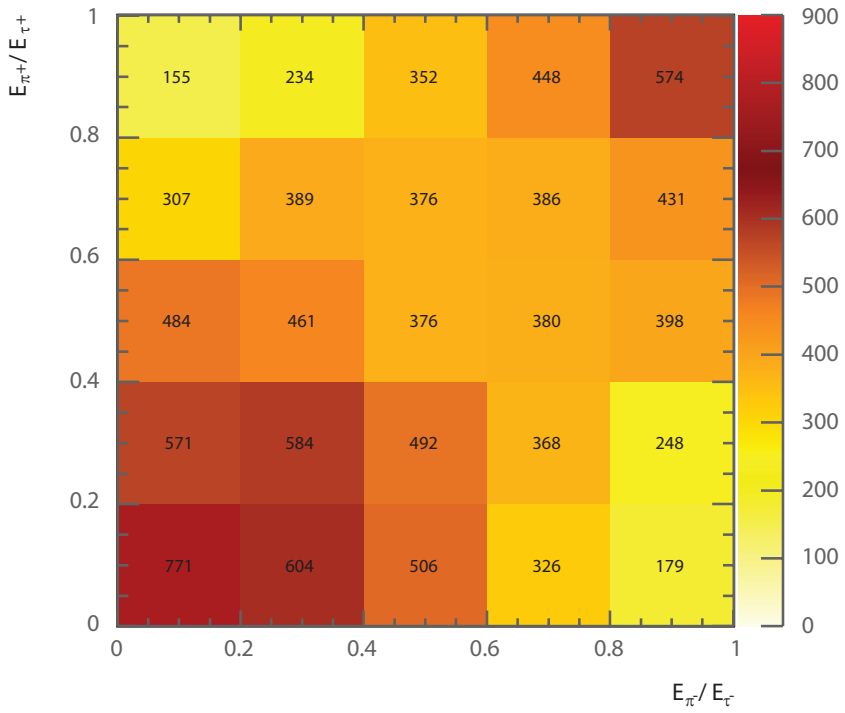
Figure 7.4: Distributions of $E_{\pi^\pm}/E_{\tau^\pm}$ from $Z \to \tau^+\tau^-$ decays where both $\tau^\pm \to \pi^\pm \nu_\tau$, in $e^+e^- \to ZZ \to \tau^+\tau^-qq$ events for: a) generator-level Monte Carlo particles, and b) the full detector simulation. The $\tau^\pm \to \pi^\pm \nu_\tau$ decay mode is selected using: a) the truth information; and b) the tau decay mode classifier.

the generator-level Monte Carlo particles. Figure 7.5b 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 $\tau^\pm \to \pi^\pm \nu_\tau$, 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 due to the failure of identifying the correct tau decay products.

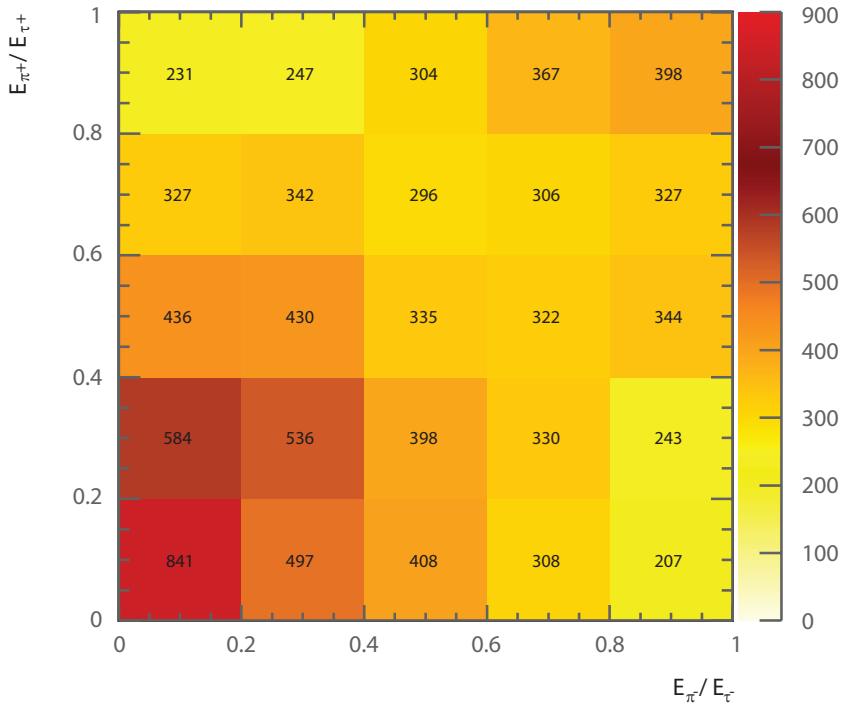
The fact that the effects of tau spin correlations are presented in figure 7.5 provides a demonstration of the method. If the analysis had been repeated with $e^+e^- \to HZ \to \tau^+\tau^-qq$, an anti-correlation would be seen the two-dimensional energy fraction plot compared to figure 7.5b.

7.8 Summary

This is a proof-of-principle demonstration that the generator-level pion energy fraction correlation can be reconstructed at the analysis level. The analysis contains several important steps: tau identification; kinematic reconstruction of the energies of the taus; the classification of the $\tau^\pm \to \pi^\pm \nu_\tau$ decay mode; and the reconstruction of the tau pair polarisation correlations discussed in section 2.9.



(a)



(b)

Figure 7.5: Two-dimensional distributions of E_{π^+}/E_{τ^+} as a function of E_{π^-}/E_{τ^-} from $Z \to \tau^+\tau^-$ decays where both $\tau^\pm \to \pi^\pm \nu_\tau$, in $e^+e^- \to ZZ \to \tau^+\tau^-qq$ events for: a) generator-level Monte Carlo particles; and b) the full detector simulation. The $\tau^\pm \to \pi^\pm \nu_\tau$ decay mode is selected using: a) the truth information; and b) the tau decay mode classifier.

Chapter 8

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 [10, 11], 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 [12, 13]. The study of double Higgs production would process 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 [20, 21, 99].

In e^+e^- collisions, there are two main challenges with the study of the double Higgs production process, $e^+e^- \to HH\nu_e\overline{\nu}_e$. 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 in the forward region, 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 process, $e^+e^- \to HH\nu_e\overline{\nu}_e$, 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 CLIC. Part of this analysis has been published in [23].

8.1 Analysis strategy overview

Leading-order Feynman diagrams for double Higgs production via W^+W^- fusion are shown in figure 8.1. The diagram shown in figure 8.1a contains the triple Higgs vertex, which is sensitive to the Higgs trilinear self coupling g_{HHH}. The diagram in the figure 8.1b is sensitive to the quartic coupling g_{WWHH}. Figures 8.1c and 8.1d show the Feynman diagrams for irreducible background processes in the study of g_{HHH} and g_{WWHH}.

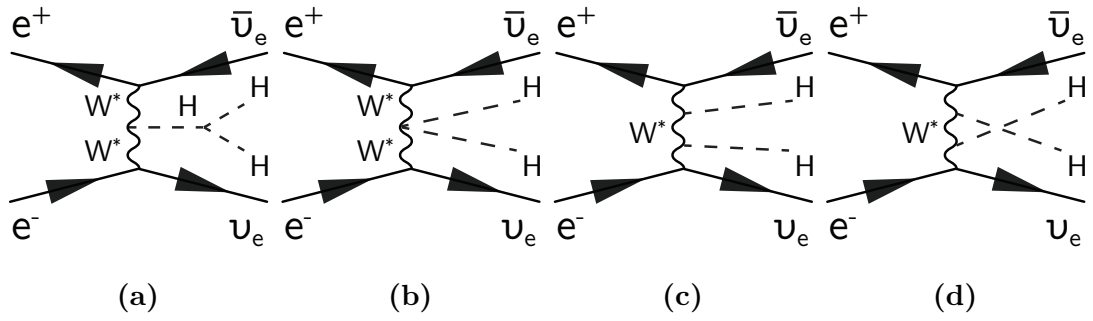


Figure 8.1: The main Feynman diagrams for the leading-order $e^+e^- \to HH\nu_e\overline{\nu}_e$ processes at CLIC.

Double Higgs production can also be produced via $e^+e^- \to ZHH$. This also contributes to the $HH\nu\overline{\nu}$ final state for Z decaying to $\nu\overline{\nu}$. The ZHH process has been studied in e^+e^- collisions at $\sqrt{s} = 500\,\text{GeV}$ [100]. However, for the CLIC energies of $\sqrt{s} = 1.4\,\text{TeV}$ and 3 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 bosons 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 CLIC 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, $HH\nu_e\overline{\nu}_e \to b\bar{b}W^+W^-\nu_e\overline{\nu}_e \to b\bar{b}qqqq\nu_e\overline{\nu}_e$, 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 processes 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 the jets are used as inputs for pre-selection and multivariate analysis.

The event analysis 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.

8.2 Monte Carlo sample generation

A full list of generated samples with their cross sections can be found in table 8.1. 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 interaction cross sections 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 [101].

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 such background processes are $e^+e^- \to qqqq\nu\overline{\nu}$ and $e^\pm\gamma \to \nu qqqq$. For the same reason, single Higgs boson production, such as $e^+e^- \to qqH\nu\overline{\nu}$, has a similar final state to the signal events and 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 [102], as the values are accounted for measure Higgs boson mass.

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 four-momenta of all quarks is above 50 GeV or 120 GeV. This restricts the event 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.

8.3 Lepton identification

For the signal process, $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$, there is no primary lepton in the final state, whilst many background processes, such as $qqqq\ell\nu$, contain primary leptons. 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.

8.3.1 Electron and muon identification

Two approaches to electron and muon identification were utilised, which are described below. The performance is summarised in table 8.6.

Process $\sqrt{s} = 1.4\,\text{TeV}$	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
$e^+e^- \to q_l q_l H\nu\overline{\nu}$	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 8.1: 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. γ (EPA) represents a "quasi-real" photon, simulated with the Equivalent Photon Approximation. For processes labelled with * and †, events are generated with the invariant mass of the total momenta of all quarks above 50 and 120 GeV, respectively.

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 CLIC collaborator, Rosa Simoniello, using the $HH \to b\bar{b}b\bar{b}$ as the signal process and the $e^+e^- \to qqqq\ell\nu$ as the background process, as the background processes are the same for this analysis with $HH \to b\bar{b}W^+W^-$ process.

Optimal values of the parameters of the IsolatedLeptonFinder are listed in table 8.2: 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 8.2: Optimised parameters of the IsolatedLeptonFinder processor.

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 identity information provided by the PandoraPFA reconstruction to identify leptons.

Table 8.3 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 $r_0 < 0.015\,\text{mm}$. Afterwards, the lepton should either have $p_\text{T} > 40\,\text{GeV}$, or $E \geqslant 23\,\text{GeV}^{\frac{1}{2}} \times \sqrt{E_{cone1}} + 5\,\text{GeV}$. Figure 8.2a and 8.2b show the distributions of the p_T of identified electrons after E and r_0 cuts, for 4000 $\text{HH} \to \text{b}\bar{\text{b}}\text{W}^+\text{W}^- \to \text{b}\bar{\text{b}}\text{qqqq}$ signal events and 4000 $\text{e}^+\text{e}^- \to \text{qqqq}\ell\nu$ background events respectively. A cut of $p_\text{T} > 40\,\text{GeV}$ preserves most signal events. Figure 8.2c and 8.2d show the distributions of $23\,\text{GeV}^{\frac{1}{2}} \times \sqrt{E_{cone1}} + 5\,\text{GeV}$ as a function of E of identified electrons after E and r_0 cuts, for $\text{HH} \to \text{b}\bar{\text{b}}\text{W}^+\text{W}^- \to \text{b}\bar{\text{b}}\text{qqqq}$ signal events and $\text{e}^+\text{e}^- \to \text{qqqq}\ell\nu$ background events respectively. A cut along the two-dimensional histogram would discard background events and leave signal events intact. The choice of $E \geqslant 23\,\text{GeV}^{\frac{1}{2}} \times \sqrt{E_{cone1}} + 5\,\text{GeV}$ allows more energy in the isolation cone for a high energy lepton.

The second set of cuts is similar to the first set of cuts. Apart from the differences in the values of the cuts, lepton ID is determined using the fraction of the energy deposited in the ECAL relative to 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.

IsolatedLeptonIdentifer	Selection
High Energy	$E > 10\,\text{GeV}$
$\text{e}^\pm$ 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	$p_\text{T} > 40\,\text{GeV}$
b) Isolation	$E \geqslant 23\,\text{GeV}^{\frac{1}{2}} \times \sqrt{E_{cone1}} + 5\,\text{GeV}$
High Energy	$E > 10\,\text{GeV}$
$\text{e}^\pm$ 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	$p_\text{T} > 40\,\text{GeV}$
b) Isolation	$E \geqslant 28\,\text{GeV}^{\frac{1}{2}} \times \sqrt{E_{cone2}} + 30\,\text{GeV}$

Table 8.3: Optimised parameters of 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).

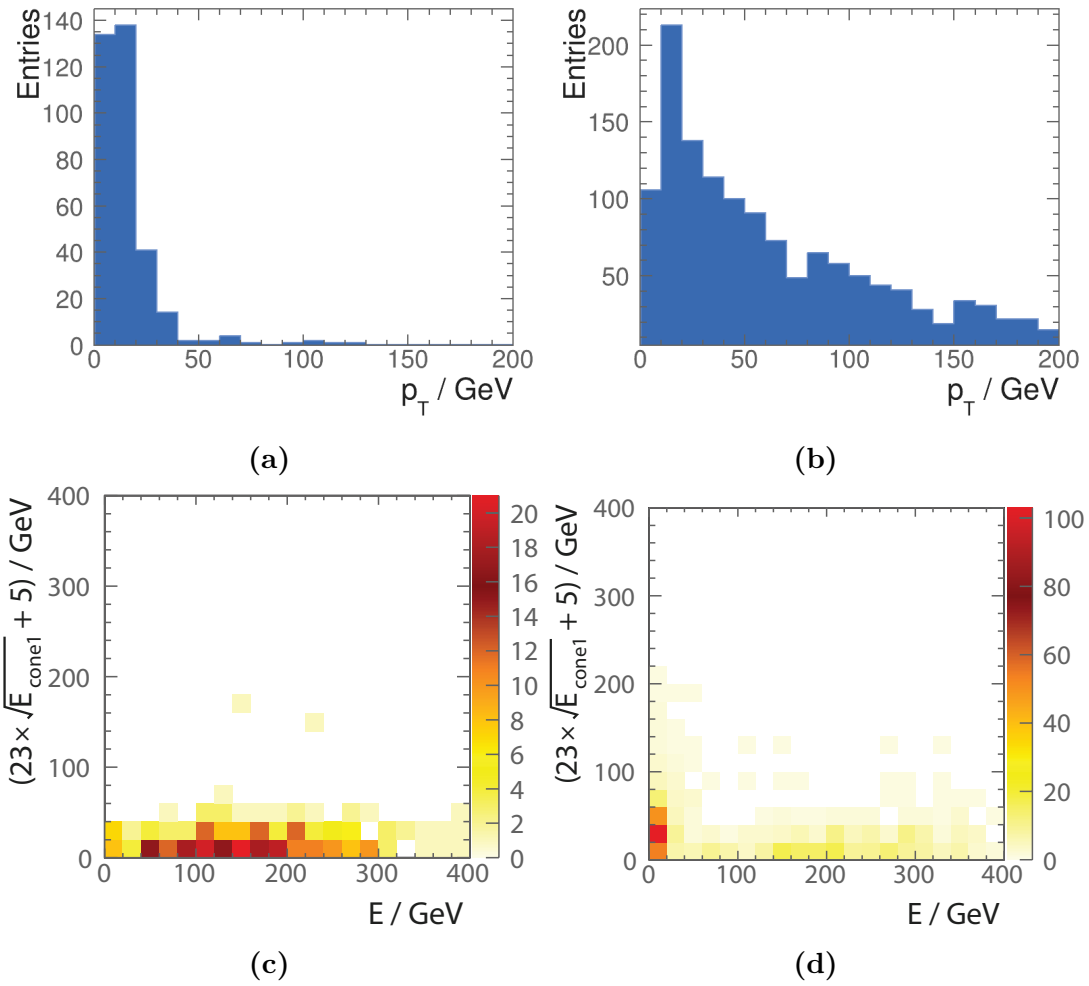


Figure 8.2: Distributions of the p_{T} of identified electrons after E and r_0 cuts, for: a) $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-} \to \mathrm{b\bar{b}qqqq}$ signal process; and b) $\mathrm{e^+e^-} \to \mathrm{qqqq}\ell\nu$ background process. Distributions of $23\,\mathrm{GeV}^{\frac{1}{2}} \times \sqrt{E_{cone1}} + 5\,\mathrm{GeV}$ 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 events; and d) $\mathrm{e^+e^-} \to \mathrm{qqqq}\ell\nu$ background events.

8.3.2 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 [103] reconstruction package has been optimised. In addition, a package, IsolatedTauIdentifer, was developed to provide additional tau lepton identification.

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 fewer 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 8.4. The optimisation is performed by the CLIC collaborator, Rosa Simoniello, using $\mathrm{HH} \to \mathrm{b\bar{b}b\bar{b}}$ signal process and the $\mathrm{e^+e^-} \to \mathrm{qqqq}\ell\nu$ background process, 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_τ are the number of charged particles and the number of PFOs in the tau cone respectively; m_τ 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 8.4: Optimised parameters of the TauFinder processor.

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 referred to as the search cone, which contains potential 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 decays and 3-prong decays, reflecting different topologies of tau decay final states. The isolation criteria 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.

Table 8.4 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.

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

Table 8.5: Optimised parameters of IsolatedTauIdentifer processor

Figure 8.3 shows the distributions of variables used in isolation criterion of tau candidate for 600 $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-} \to \mathrm{b\bar{b}qqqq}$ signal events (blue) and 5000 $\mathrm{e^+e^-} \to \mathrm{qqq}\ell\nu$ background events (orange). Figure 8.3a shows the distribution of the transverse mo-

mentum of the particles in the search cone, after selecting $N_{cone1} = 0$, used in the isolation criterion 1. The cut at $p_{\mathrm{T}\,cone} \geqslant 10\,\mathrm{GeV}$ selects more tau candidates in background events than in the signal events, where there should be no true high-energy isolated tau leptons in signal events. Figure 8.3b shows the distributions of r_0 after selecting $N_{X^+} = 1$ and $N_{cone1} = 1$. The cut at $r_0 > 0.01\,\mathrm{mm}$ used in the isolation criterion 2 selects more true tau candidates in background events. Figure 8.3a shows the distribution of the transverse momentum of the particles in the search cone, after requiring $N_{X^+} = 3$, $N_{cone1} = 1$ and $\theta_S < \cos^{-1}(0.9995)$ for isolation criteria 3. The cut at $p_{\mathrm{T}\,cone} \geqslant 10\,\mathrm{GeV}$ selects tau candidates in background events.

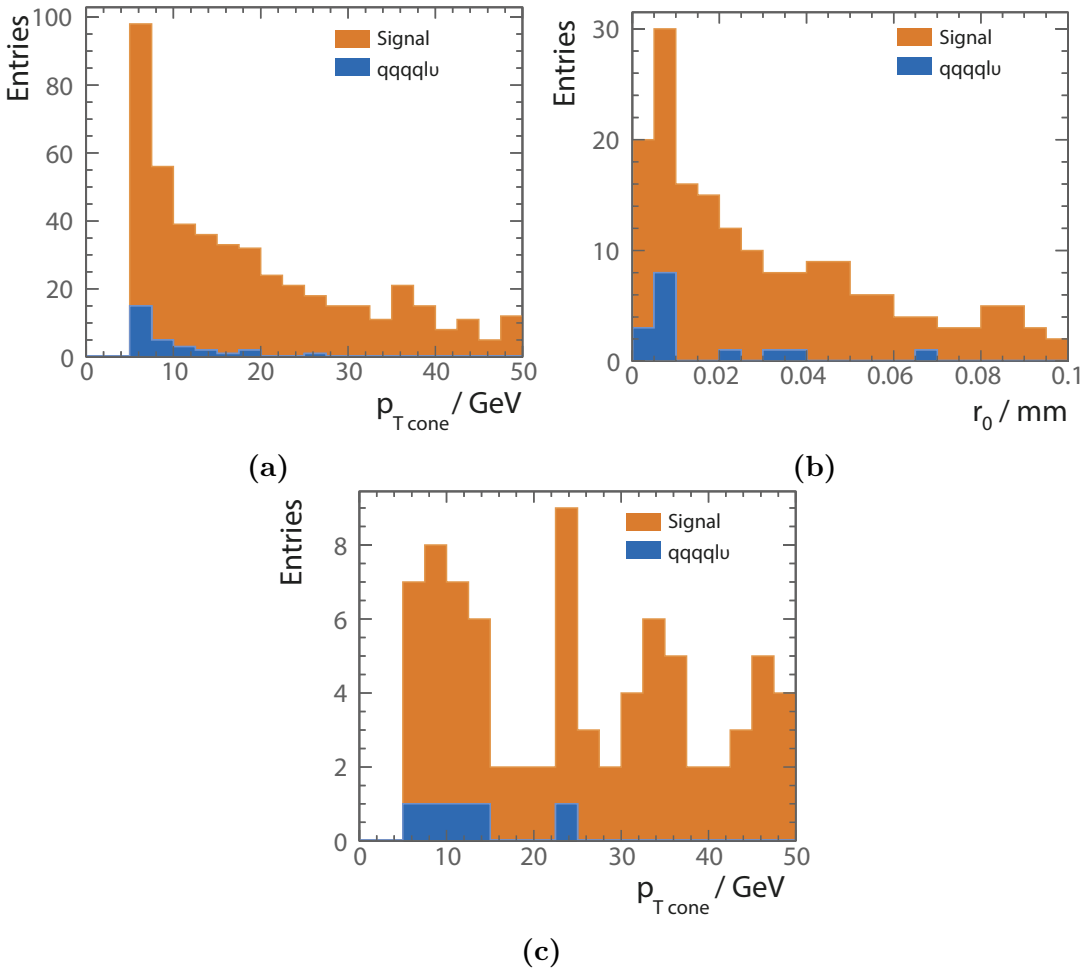


Figure 8.3: Distributions to show isolation criteria: a) $p_{\mathrm{T}\,cone}$ for isolation criterion 1 after selecting $N_{cone1} = 0$; b) r_0 for isolation criterion 2 after selecting $N_{X^+} = 1$, $N_{cone1} = 1$; and c) $p_{\mathrm{T}\,cone}$ for isolation criteria 3 after selecting $N_{X^+} = 3$, $N_{cone1} = 1$, $\theta_S < \cos^{-1}(0.9995)$. Distributions are shown for tau candidates in $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-} \to \mathrm{b\bar{b}qqqq}$ signal events (blue) and $\mathrm{e^+e^-} \to \mathrm{qqqq}\ell\nu$ background events (orange).

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 tau search cone size.

8.3.3 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, beam induced background in the forward calorimeters are not simulated, due to the high demand on the computational resources. Particles in the forward calorimeters are not reconstructed for the same reason. Instead, studies have been performed with particles simulated in the forward calorimeters to understand the primary lepton identification efficiencies [49, 104, 105]. The studied primary lepton identification efficiencies are then parameterised as a function of lepton energies. The parametrisation approach is adopted in this analysis.

Figure 8.4a shows the primary electron identification efficiencies in the BeamCAL as a function of polar angle for a 500 GeV electron. An external processor [104] has been 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 energies to use the external processor.

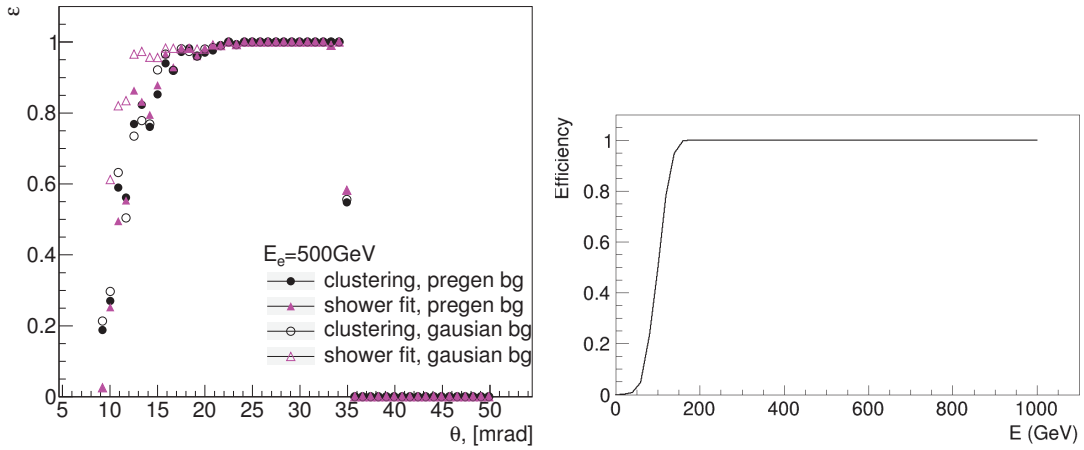
Figure 8.4b shows the primary electron identification efficiencies in the LumiCAL as a function of electron energy for a polar angle $\theta = 50\,\text{mrad}$. The efficiency is obtained from a full simulation study [105], assuming a background integrated over 100 bunch crossings. In this analysis, the primary electron identification efficiency as a function of electron energy is assumed to be parameterised by the curve in figure 8.4b. The polar angle dependency of the efficiency is not considered, due to the lack of study. The primary

electron identification efficiency curve in figure 8.4b takes the functional form of:

$$\varepsilon = \begin{cases} 0, & \text{if } E < 50\,\text{GeV}, \\ 0.99 \times \frac{(\text{erf}(E/\text{GeV}-100)+1)}{2}, & \text{otherwise}, \end{cases} \tag{8.1}$$

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

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



(a) **(b)**

Figure 8.4: a) 500 GeV electron identification efficiency in the BeamCAL as a function of polar angles, with different methods to model backgrounds: pre-generated and Gaussian, and two methods to identify electrons: clustering algorithm and shower fitting algorithm, obtained from a full simulation study in [104]. b) electron tagging efficiency in the LumiCAL as a function of the electron energy, for a polar angle $\theta = 50\,\text{mrad}$, obtained from a full simulation study in [105].

8.3.4 Summary of lepton identification performance

The performances of the different lepton finding processors for signal events and the selected background processes are shown in table 8.6 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. Only 1% of signal events are rejected, but 47.4% of the $e^-\gamma(\text{BS}) \to e^-qqqq$ background events are rejected.

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

Table 8.6: The performances of the lepton finding algorithms for the signal events and selected background events at $\sqrt{s} = 1.4\,\text{TeV}$. γ (BS) represents a real photon from beamstrahlung. 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 $\sqrt{s} = 1.4\,\text{TeV}$. The same set of parameters is also effective for $\sqrt{s} = 3\,\text{TeV}$. The performances of the lepton finders at $\sqrt{s} = 3\,\text{TeV}$ are summarised in table 8.7.

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 tend to be boosted more and the spatial separation between particles is smaller due to the higher multiplicities. Consequently 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.

8.4 Jet reconstruction

The signal process, $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$, 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.

Efficiency (3 TeV)	Signal	$e^+e^- \to qqqq\ell\nu$	$e^-\gamma(BS) \to e^-qqqq$
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 8.7: The performances of the lepton finding algorithms for the signal events and selected background events at $\sqrt{s} = 3\,\text{TeV}$. γ (BS) represents a real photon from beamstrahlung. Numbers represent the fractions of events where no leptons are identified by the individual lepton finder.

8.4.1 Jet reconstruction optimisation

Jet reconstruction algorithms cluster particles into jets. Jet reconstruction is important at 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. For this analysis, the 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 size of the jet. 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, $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$ events are processed through k_t jet algorithm in the six-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 ($m_{H_{bb}}$ and $m_{H_{WW^*}}$) and two W masses (m_W and m_{W^*}). Here W^* indicates the off-mass-shell W boson. The MC paring is used to optimise the choice of parameters. It is not used in the subsequent analysis.

Three invariant mass distributions are considered: $m_{H_{bb}}$, $m_{H_{WW^*}}$, and m_W. The optimal jet reconstruction should produce sharp mass peaks around the simulated particle masses. For example, figure 8.5 shows the $m_{H_{bb}}$ invariant mass distribution for $R = 1.3$ using the loose PFO collection for samples at $\sqrt{s} = 3\,\text{TeV}$. An analytical functional form is fitted to 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

$$f(m) = A \exp\left\{-\frac{(m-\mu)^2}{g}\right\}, \tag{8.2}$$

$$g = \begin{cases} 2\sigma_L + \alpha_L(m-\mu), & \text{if } m < \mu, \\ 2\sigma_R + \alpha_R(m-\mu), & \text{if } m \geqslant \mu, \end{cases} \tag{8.3}$$

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.

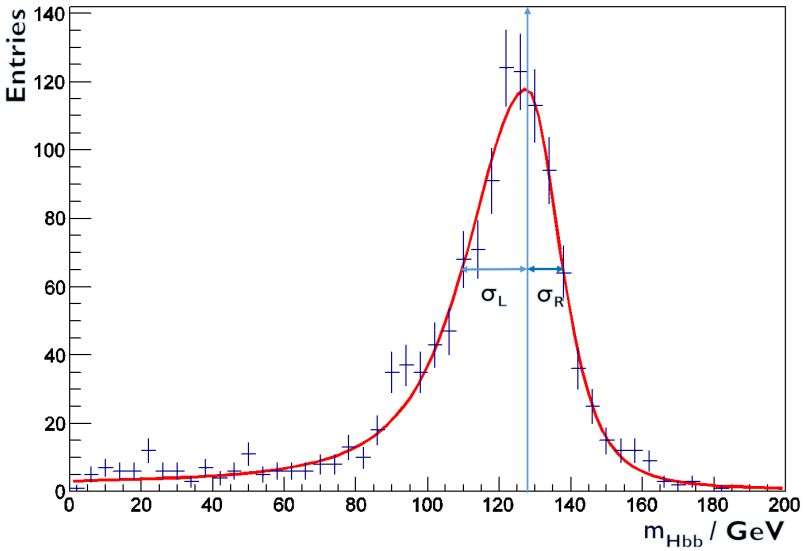


Figure 8.5: A typical example of the reconstructed $m_{\mathrm{H_{bb}}}$ mass distribution for $R = 1.3$ using loose PFO collection for $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-} \to \mathrm{b\bar{b}qqqq}$ samples at $\sqrt{s} = 3\,\mathrm{TeV}$. The fitting function is superimposed in red. The arrow shows the fitted peak position.

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 8.6 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 radii. For the relative width, the values for $\mathrm{H_{bb}}$ increase with increasing

jet radii, but the values for H_{WW^*} decrease with increasing jet radii. This is due to a compensating effect; the invariant mass for H_{WW^*} is formed from four jets, which prefers a large jet radius, whereas the invariant mass for 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 8.6, it was decided to use $R = 0.7$ with the selected PFO collection. This choice gives good fitted mass peak positions for H_{bb}, H_{WW^*} and W. The extracted fitted parameters of optimal jet reconstructions are summarised in table 8.8.

Fitted jet parameter	$\sqrt{s} = 1.4\,\text{TeV}$	$\sqrt{s} = 3\,\text{TeV}$
$\mu_{H_{bb}}$	$122.3_{\pm 0.2}$	$119.1_{\pm 0.3}$
$\sigma_{L,H_{bb}}$	$15.2_{\pm 0.2}$	$15.0_{\pm 0.3}$
$\sigma_{R,H_{bb}}$	$7.6_{\pm 0.2}$	$8.4_{\pm 0.2}$
$\mu_{H_{WW^*}}$	$125.7_{\pm 0.2}$	$123.0_{\pm 0.3}$
$\sigma_{L,H_{WW^*}}$	$29.4_{\pm 0.3}$	$36.6_{\pm 0.6}$
$\sigma_{R,H_{WW^*}}$	$7.2_{\pm 0.2}$	$7.4_{\pm 0.2}$
μ_{W}	$80.5_{\pm 0.2}$	$78.1_{\pm 0.3}$
$\sigma_{L,W}$	$16.2_{\pm 0.3}$	$13.1_{\pm 0.4}$
$\sigma_{R,W}$	$9.0_{\pm 0.2}$	$9.5_{\pm 0.2}$

Table 8.8: 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.

A separate jet reconstruction optimisation is performed for $\sqrt{s} = 3\,\text{TeV}$ analysis. Figure 8.7 shows the variation of fitted mass peak positions and the relative mass resolutions for H_{bb}, H_{WW^*}, and W as function of R and PFO collections. The relative mass resolution of W boson quickly degrades with an increasing R. The fitted mass peak positions also increases more rapidly with the increase of R, compared with the fitted mass peak 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

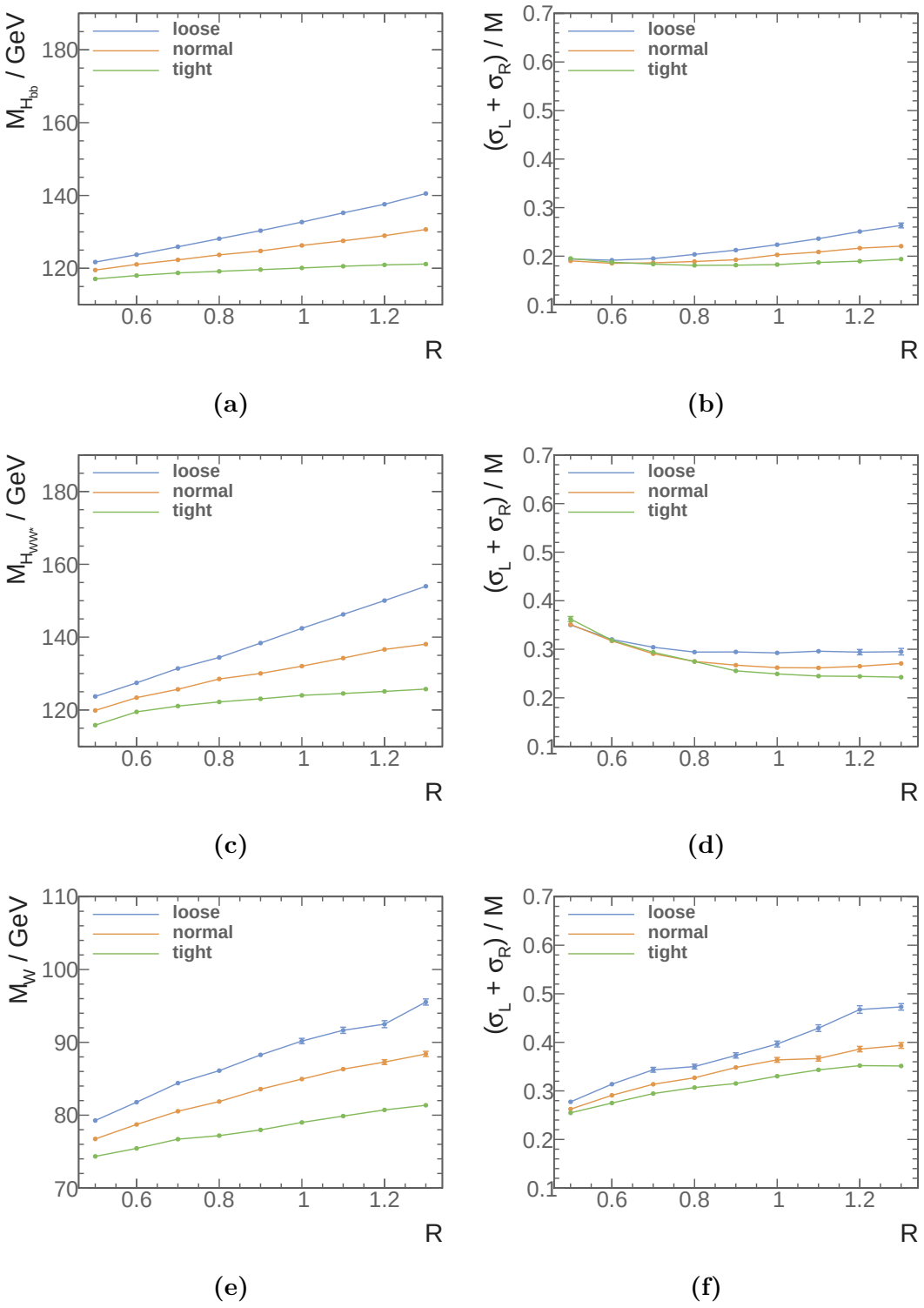


Figure 8.6: Distributions of: a) fitted mass peak positions for H_{bb}; b) relative mass peak widths for H_{bb}; c) fitted mass peak positions for H_{WW^*}; d) relative mass peak widths for H_{WW^*}; e) fitted mass peak positions for W; and f) relative mass peak widths for W. All plots show the variation 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 $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$ samples.

$\sqrt{s} = 3\,\text{TeV}$ analysis. With the 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 8.8.

8.5 Jet flavour tagging

As the signal process, $\text{HH} \to \text{b}\bar{\text{b}}\text{W}^+\text{W}^- \to \text{b}\bar{\text{b}}\text{qqqq}$, 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 [106] is used. The processor is based on the LcfiVertex package [107], which was used in the simulation studies for the ILC Letter of Intent [29, 108] and the CLIC Concept Design Report [2].

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

The vertex finding algorithms perform vertex fitting and identify primary and secondary vertices. There are two vertex refining algorithm. The 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 Durham algorithm. An additional constraint requires the decay products from the semi-leptonic decay of the quarks, secondary vertices and the muons, fall into the same jet as the parent quarks. This ensures the topology of the jet remains consistent with the hadronic decays of heavy quarks.

Having obtained re-clustered jets, the next step is 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 [79]. 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: b jet, c jet, or a light flavour quark (u, d or s) jet.

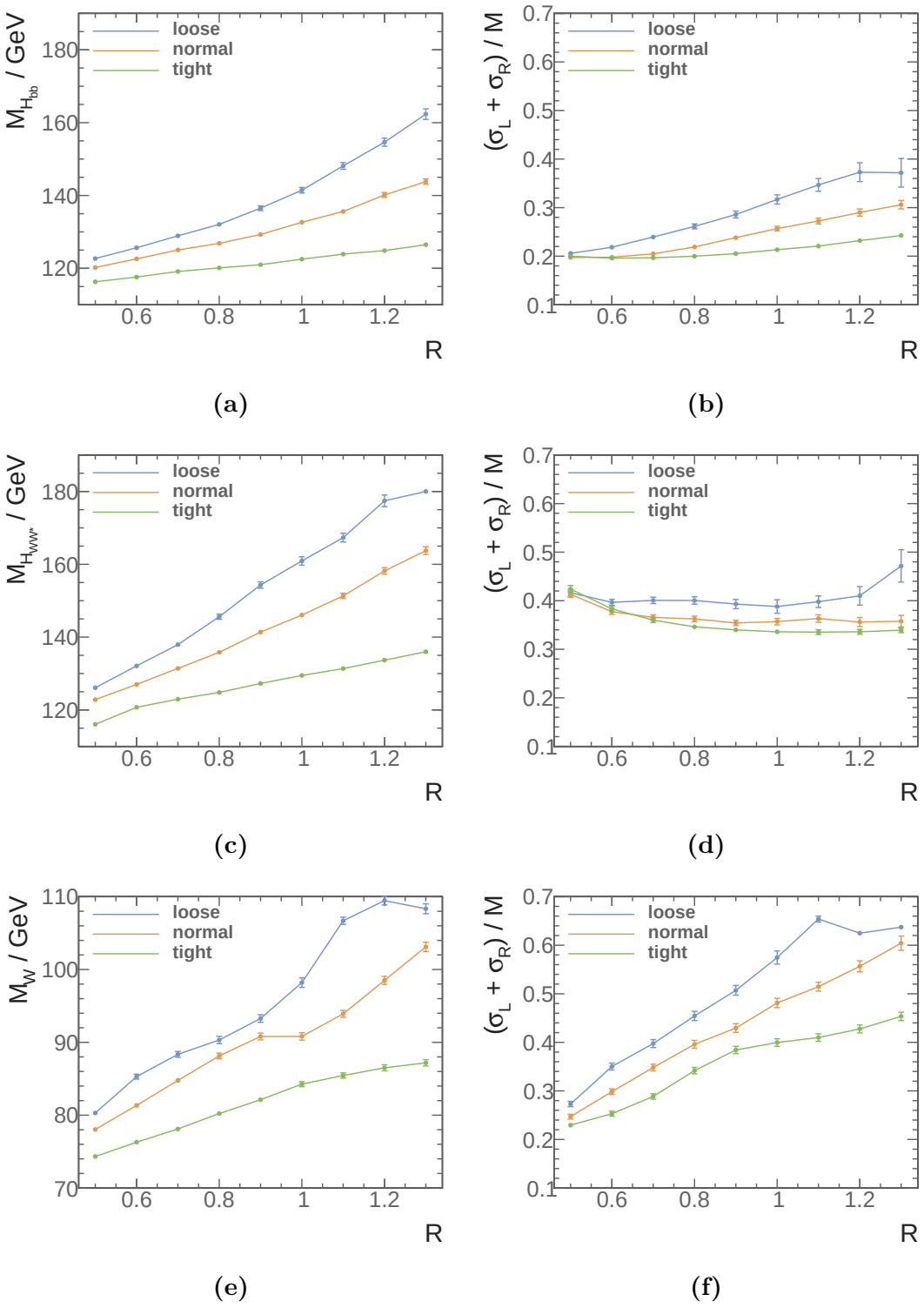
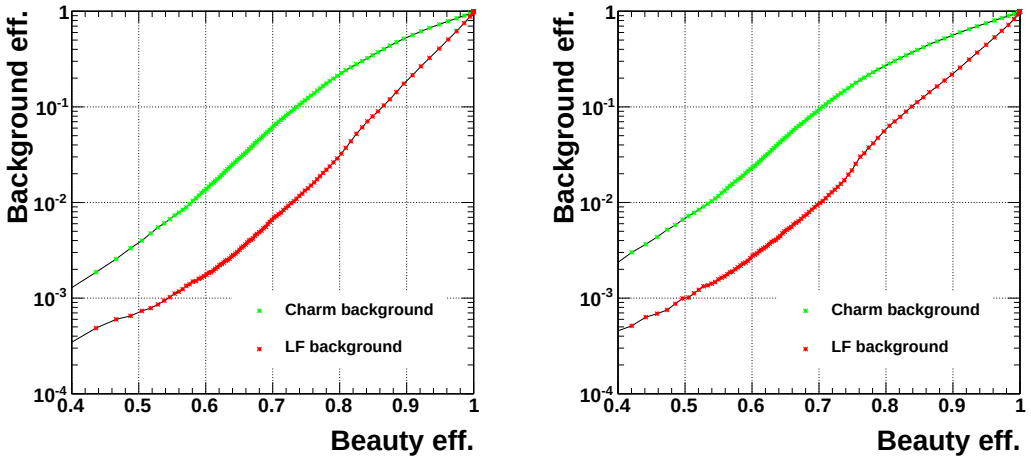


Figure 8.7: Distributions of: a) fitted mass peak positions of H_{bb}; b) relative mass peak widths of H_{bb}; c) fitted mass peak positions of H_{WW^*}; d) relative mass peak widths of H_{WW^*}; e) fitted mass peak positions of W; and f) relative mass peak widths of W. All plots show the variation of 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 $\text{HH} \to \text{b}\bar{\text{b}}\text{W}^+\text{W}^- \to \text{b}\bar{\text{b}}\text{qqqq}$ samples.

The MVA MultiClass classifier was trained with $e^+e^- \to Z\overline{\nu}\nu$ events 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 classification efficiency with the training samples is shown in figure 8.8a.

Having trained the MVA classifier, the MVA classifier is applied to the samples. Under the signal hypothesis, the re-clustering 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 8.9.

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 8.8b. 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.



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

Figure 8.8: 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) $\sqrt{s} = 3\,\text{TeV}$.

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

The two $e^+e^- \to HH\nu_e\overline{\nu}_e$ final states with the largest branching fractions are $HH \to b\overline{b}b\overline{b}$ (31.5%) and $HH \to b\overline{b}W^+W^-$ (25.9%). These two final states have different topologies

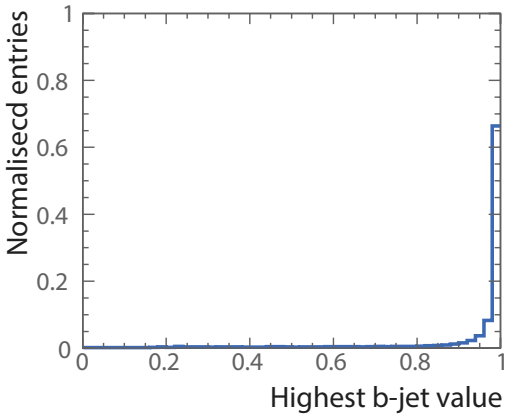


Figure 8.9: The distribution of the highest b-jet value for the $HH \to b\bar{b}W^+W^- \to b\bar{b}qqqq$ events at $\sqrt{s} = 1.4\,\text{TeV}$. The area under the curve is normalised to unity.

and are subjects 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 [23]. Because the results of the two studies are subsequently combined, a set of cuts is 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 differences between the $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$ sub-channels are 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.

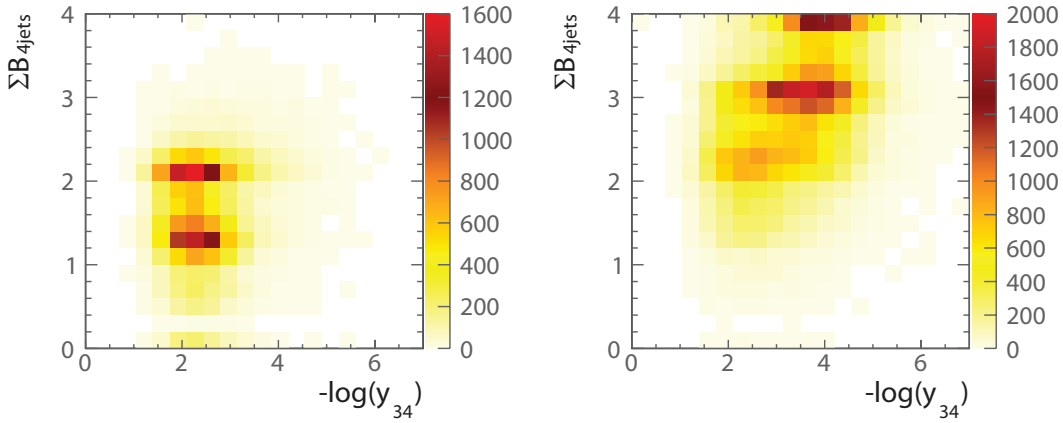
Figure 8.10 shows the sum of b-jet tag values, when the event is clustered into four jets, as a function of $-\log(y_{34})$ for the hadronic W^+W^- decay in $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$. As expected, the two sub-channels can be clearly separated in this two dimensional phase 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 phase space S, and the $HH \to b\bar{b}b\bar{b}$ events should be contained in phase space $\neg S$.

The optimal cuts are chosen such that they maximise:

$$\varepsilon = \frac{N_{HH \to b\bar{b}W^+W^-,\,\text{hadronic}} \in S}{N_{HH \to b\bar{b}W^+W^-,\,\text{hadronic}}} \times \frac{N_{HH \to b\bar{b}b\bar{b}} \in \neg S}{N_{HH \to b\bar{b}b\bar{b}}}, \tag{8.4}$$

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 8.9.



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

Figure 8.10: The two-dimensional distribution of sum of b-jet tag values against $-\log(y_{34})$ for: 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 $\sqrt{s} = 1.4\,\text{TeV}$. The sum of b-jet tag values is calculated for the cases where events are clustered into four jets.

8.6 Jet pairing

All events are reconstructed assuming the $HH \to b\bar{b}W^+W^-$ 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 $H \to b\bar{b}$, 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.

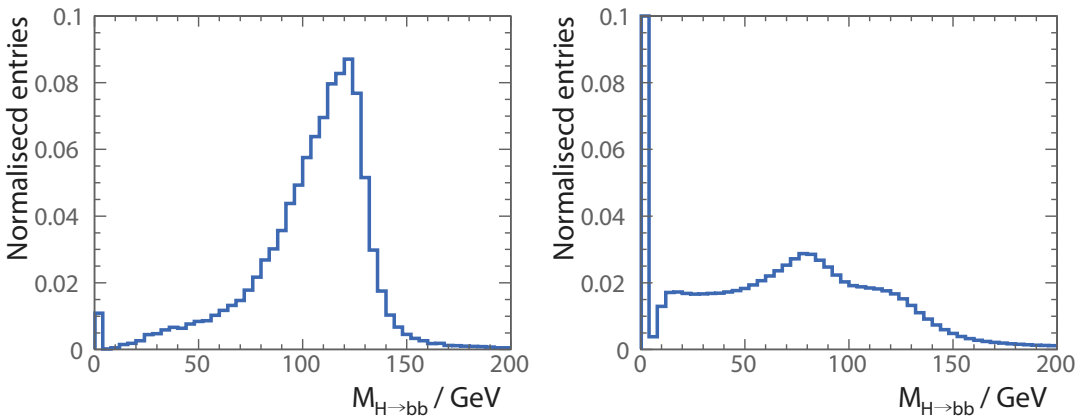
Six jets are associated to H_{bb}, W and W*. There are 90 possible permutations for associating six jets to H_{bb}, W and W*. The best permutation is obtained by choosing the permutation that gives the smallest χ^2 quantity, representing the consistency of the hypothesis with the signal topology:

$$\chi^2 = \left(\frac{m_{ij} - \mu_{H_{bb}}}{\sigma'_{H_{bb}}}\right)^2 + \left(\frac{m_{klmn} - \mu_{H_{WW^*}}}{\sigma'_{H_{WW^*}}}\right)^2 + \left(\frac{m_{kl} - \mu_W}{\sigma'_W}\right)^2, \tag{8.5}$$

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 8.8, defined as:

$$\sigma'_{\mathrm{H_{bb}}} = \begin{cases} \sigma_{L,\mathrm{H_{bb}}}, & \text{if } m_{ij} < \mu_{\mathrm{H_{bb}}}, \text{ etc.} \\ \sigma_{R,\mathrm{H_{bb}}}, & \text{otherwise,} \end{cases} \tag{8.6}$$

A jet pairing is only considered when at least one of the jets associated to the $\mathrm{H_{bb}}$ decay has a b-jet tag > 0.2. Of these combinations of jets, the jet pairing giving smallest χ^2 is selected. Figure 8.11 shows the normalised distributions of $m_{\mathrm{H_{bb}}}$ after jet pairing, for a) the signal process, $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$ and b) the sum of all background processes. For the signal process, the distribution peaks around the expected mass of $m_{\mathrm{H_{bb}}}$. Around 1% of signal events have no solutions for the jet pairing, i.e. 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 8.9 for signal $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$ process and all background processes.



(a) **(b)**

Figure 8.11: The distribution of $m_{\mathrm{H_{bb}}}$ for: a) the signal process, hadronic $\mathrm{W^+W^-}$ decay of $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$; and b) the sum of all background processes normalised to the respective cross sections. The area under the curve is normalised to unity. All plots are shown for $\sqrt{s} = 1.4\,\mathrm{TeV}$.

8.7 Pre-selection

After the association of jets to candidate bosons, kinematic and topological variables can be calculated. A set of pre-selection cuts are placed to discard the phase space dominated

$\sqrt{s} = 1.4\,\text{TeV}$	N	Lepton veto	$b\bar{b}W^+W^-$/ $b\bar{b}W^+W^-$ 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 8.9: 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. The 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.

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 $m_{\mathrm{HH}} > 150\,\mathrm{GeV}$, as shown in Figure 8.12, removes a small faction of signal events but discards many background events, especially $\gamma\gamma \to \mathrm{qqqq}$ events.

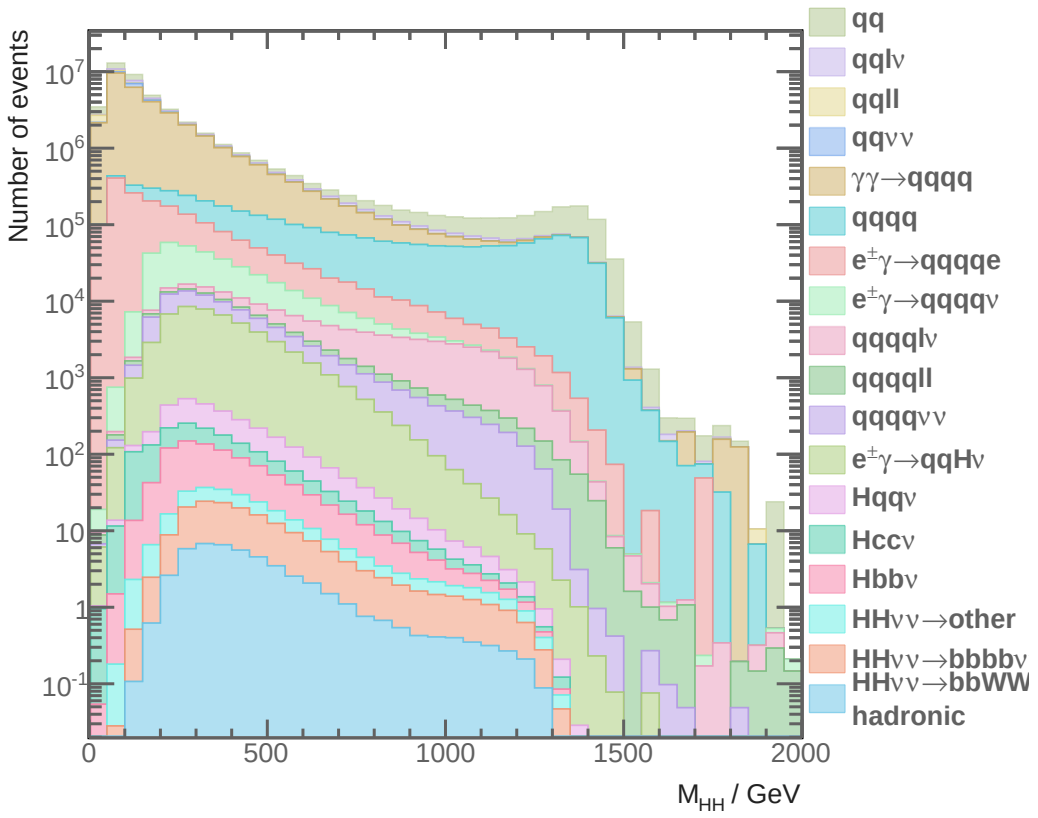


Figure 8.12: Distributions of the invariant mass of the two Higgs system for $\sqrt{s} = 1.4\,\mathrm{TeV}$, assuming an integrated luminosity of $1500\,\mathrm{fb}^{-1}$.

Many background events do not have b quarks in the final state. Therefore, by requiring the second highest b-jet tag value greater than 0.2, as shown in Figure 8.13, background events with no b quarks 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

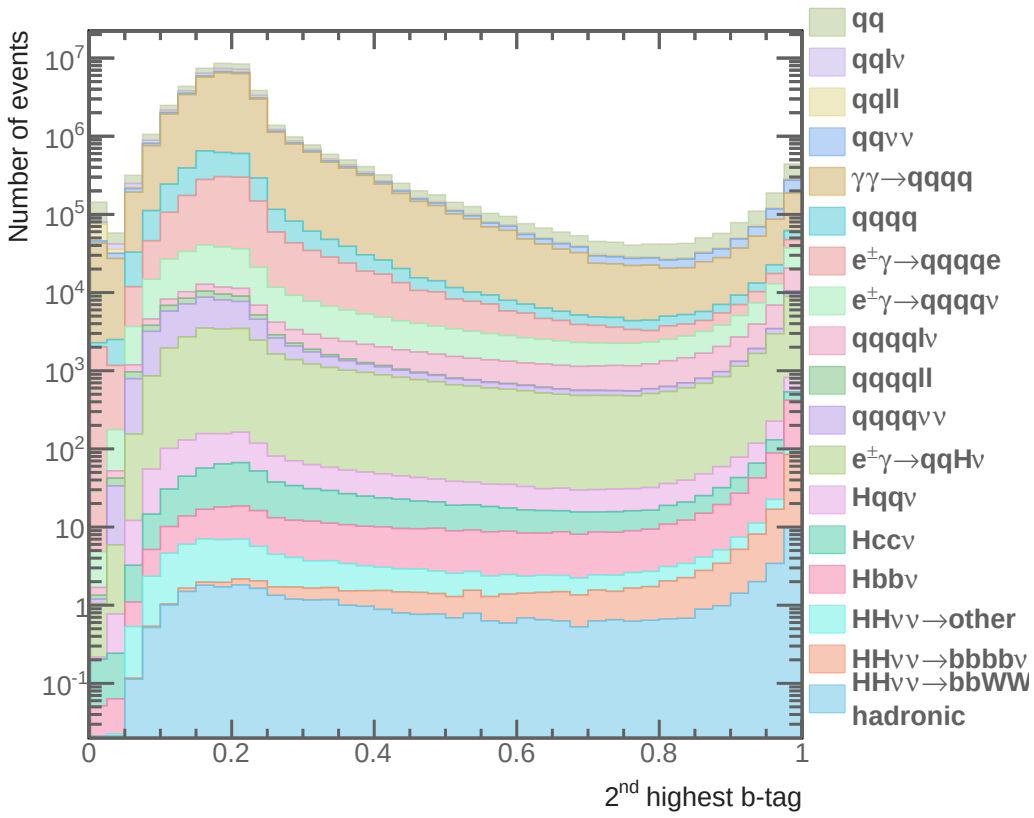


Figure 8.13: Distributions of the second highest b-jet tag value for $\sqrt{s} = 1.4\,\text{TeV}$, assuming an integrated luminosity of $1500\,\text{fb}^{-1}$.

$p_T > 30\,\text{GeV}$, as shown in figure 8.14, is extremely effective against background processes with no neutrinos in the final state.

The full list of fraction of events surviving after each pre-selection cut can be found in table 8.10.

8.7.1 Cuts to aid the MVA

Occasionally, invariant masses variables have extreme values. By limiting the range of the variables, the MVA classifier can focus on the phase space with high signal event density. 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 $< 1400\,\text{GeV}$. Events that failed cuts are discarded.

Process	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 8.10: 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. The table shows the signal and background events at $\sqrt{s} = 1.4$ TeV, assuming an integrated luminosity of 1500 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.

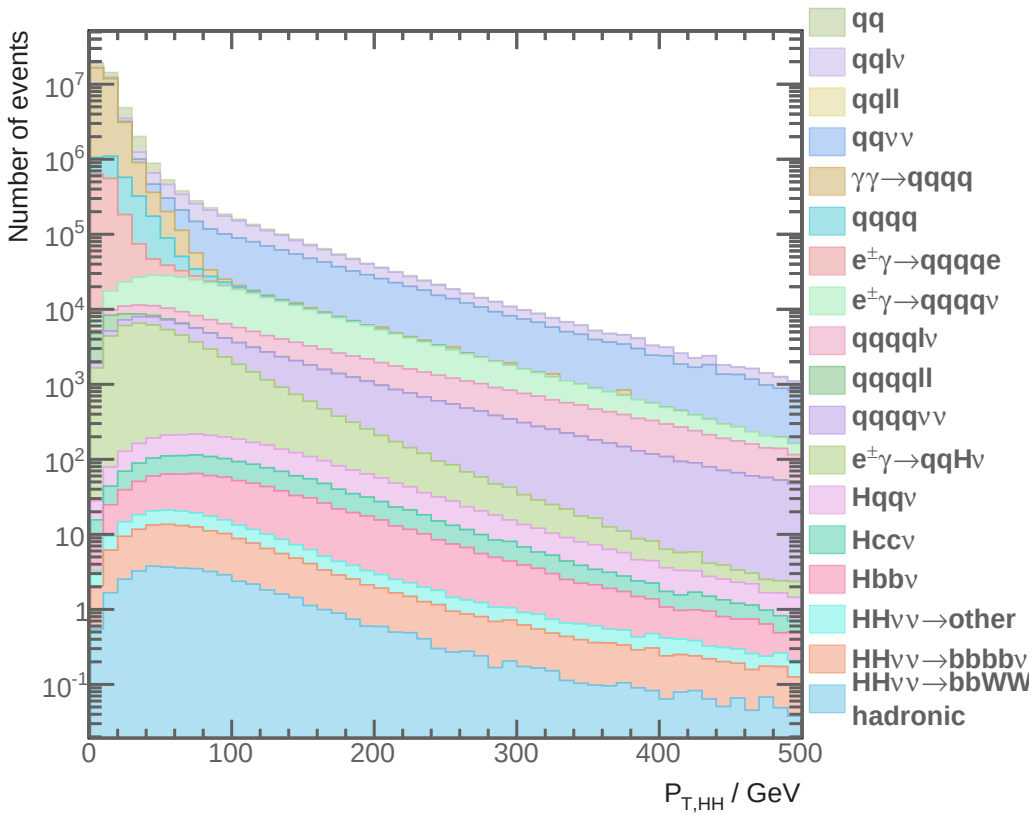


Figure 8.14: Distributions of the transverse momentum of the two Higgs system for $\sqrt{s} = 1.4\,\text{TeV}$, assuming an integrated luminosity of $1500\,\text{fb}^{-1}$.

8.8 MVA variables

Having extracted information about leptons, b-jets, and jet pairing, a number of variables are used to differentiate the signal to the background events. These variables are the basis of the subsequent MVA event selection listed in table 8.11. The distributions of the four most powerful discriminators are shown in figure 8.15.

8.8.1 Invariant mass variables

Four invariant masses are used in the MVA event selection: the invariant mass of H_{bb} ($m_{\text{H}_{\text{bb}}}$), the invariant mass of H_{WW^*} ($m_{\text{H}_{\text{WW}^*}}$), 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

expected masses, whereas the distributions of the background events do not have such peak structure. Shown in figure 8.15a, 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 8.15b, 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.

8.8.2 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} ($p_{T H_{bb}}$), the transverse momentum of H_{WW^*} ($p_{T H_{WW^*}}$), 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 a powerful discriminant variable against background events with no neutrinos in the final states. The missing momentum 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.

8.8.3 Laboratory-frame angular variables

Four laboratory-frame angular variables are used 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

$$\eta_{mis} \equiv -\ln\left[\tan\left(\frac{\theta_{mis}}{2}\right)\right], \tag{8.7}$$

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

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

$$A_{ij} = \pi - \cos^{-1}\left(\hat{\mathbf{p}}_i \cdot \hat{\mathbf{p}}_j\right), \tag{8.8}$$

where $\hat{\mathbf{p}}_i$ is the unit momentum three-vector of momenta i. The distribution of the $A_{\mathrm{H_{bb}}}$, shown in figure 8.15c, 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 $A_{\mathrm{H_{WW^*}}}$ and A_{HH} are different for the signal and the background events.

8.8.4 Rest-frame angular variables

The MVA event selection also uses five rest-frame angular variables: the angle between two jets associated with $\mathrm{H_{bb}}$ in the $\mathrm{H_{bb}}$ decay rest frame ($\cos\theta^*_{\mathrm{H_{bb}}}$), the angle between two Ws associated with $\mathrm{H_{WW^*}}$ in the $\mathrm{H_{WW^*}}$ decay rest frame ($\cos\theta^*_{\mathrm{H_{WW^*}}}$), the angle between two jets associated with W in the W decay rest frame ($\cos\theta^*_{\mathrm{W}}$), the angle between two jets associated with $\mathrm{W^*}$ in the $\mathrm{W^*}$ decay rest frame ($\cos\theta^*_{\mathrm{W^*}}$), and the angle between two Higgs bosons in two Higgs bosons decay rest frame ($\cos\theta^*_{\mathrm{HH}}$).

These variables are some of the most powerful variables used in the MVA. For example, $\cos\theta^*_{\mathrm{H_{bb}}}$ for the signal events has a unform distribution, shown in figure 8.15d, as it is equally likely for two quarks to decay in any opening angle in the $\mathrm{H_{bb}}$ decay rest frame. For the background events, the $\cos\theta^*_{\mathrm{H_{bb}}}$ distribution for the background events peaks at 1.

8.8.5 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, $\mathbf{S}$, is a measure of the spherical symmetry of the event, which will be different for the signal and background events. The sphericity is derived from the

sphericity tensor [109], which is defined as

$$\boldsymbol{S}^{\alpha\beta} = \frac{\sum_i p_i^\alpha p_i^\beta}{\sum_i |\vec{p_i}|^2}, \tag{8.9}$$

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 axes. Eigenvalues of $\boldsymbol{S}$, 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, $\mathbf{S}$, is defined in terms of λ,

$$\mathbf{S} = \frac{3}{2}(\lambda_1+\lambda_2). \tag{8.10}$$

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

8.8.6 b-jet and c-jet 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_{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_{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 quarks in the final states.

8.8.7 Particle number variables

The last four variables used in the MVA event selection are particle numbers: the number of particles associated with $\mathrm{H_{bb}}$ ($N_{\mathrm{H_{bb}}}$), the number of particles associated with $\mathrm{H_{WW^*}}$ ($N_{\mathrm{H_{WW^*}}}$), the number of particles associated with W (N_{W}), the number of particles associated with W* ($N_{\mathrm{W^*}}$). These variables are effective to differentiate the signal events from the background events with fewer than six quarks in final states.

An 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}, $p_{\mathrm{T\,H_{bb}}}$, $p_{\mathrm{T\,H_{WW^*}}}$, $p_{\mathrm{T\,W}}$, $p_{\mathrm{T\,HH}}$
laboratory-frame angles	η_{mis}, $A_{\mathrm{H_{bb}}}$, A_{W}, A_{HH}
Rest-frame angles	$\cos\theta^*_{\mathrm{H_{bb}}}$, $\cos\theta^*_{\mathrm{H_{WW^*}}}$, $\cos\theta^*_{\mathrm{W}}$, $\cos\theta^*_{\mathrm{W^*}}$, $\cos\theta^*_{\mathrm{HH}}$
Event shape	$\lvert\mathbf{S}\rvert$, $-\ln(y_{23})$, $-\ln(y_{34})$, $-\ln(y_{45})$, $-\ln(y_{56})$
b-jet and c-jet tag	$B^1_{\mathrm{H_{bb}}}$, $B^2_{\mathrm{H_{bb}}}$, B^1_{W}, $B^1_{\mathrm{W^*}}$, $C^1_{\mathrm{H_{bb}}}$, C^1_{W}
Particle number	$N_{\mathrm{H_{bb}}}$, $N_{\mathrm{H_{WW^*}}}$, N_{W}, $N_{\mathrm{W^*}}$

Table 8.11: Variables used in the MVA event selection for $\sqrt{s} = 1.4\,\mathrm{TeV}$

8.9 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 [79]. The parameters of the boosted decision tree classifier 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 8.12.

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

8.10 Signal selection results

The numbers of events passed the MVA event selection at $\sqrt{s} = 1.4\,\mathrm{TeV}$, assuming an integrated luminosity of $1500\,\mathrm{fb}^{-1}$, are listed in table 8.13 for individual processes. A few background processes have non-zero events after the event selection; $\mathrm{e^+e^- \to q\bar{q}H\nu\overline{\nu}}$ events are difficult to discard because its topology, one Higgs and neutrinos, is very similar to the signal event topology. Similarly, $\mathrm{e^+e^- \to qqqq\ell\nu}$ events can be confused with the signal events when the lepton is undetected in the forward region, or the energy

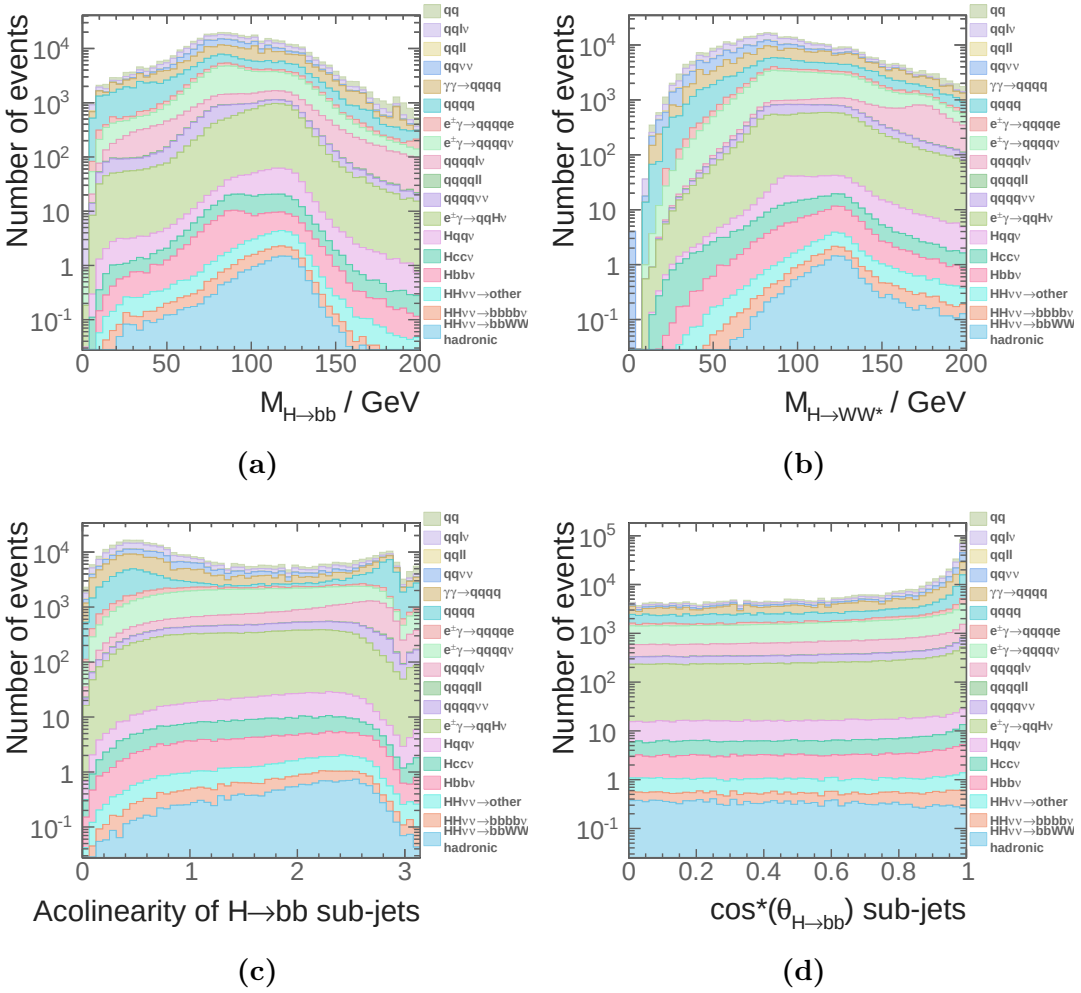


Figure 8.15: Distributions of the four variables with highest discriminating power: a) the invariant mass of H_{bb}; b) the invariant mass of H_{WW^*}; c) the acollinearity of the two jets associated with H_{bb}; and d) the opening angle of the two jets associated with H_{bb} in the decay rest frame of the H_{bb}. All plots assume an integrated luminosity of 1500 fb^{-1} at $\sqrt{s} = 1.4$ TeV after all pre-selection cuts applied before the MVA event selection.

of the lepton is too low for the lepton to be tagged. In addition, $e^+e^- \to qqqq\nu\overline{\nu}$ events can also have a similar topology to the signal events. Other background processes that are not discarded after the MVA are the electron−photon and photon−photon interactions with the same final states as the processes above.

Before interpreting the result for analysis at $\sqrt{s} = 1.4$ TeV, the analyses at $\sqrt{s} = 3$ 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.

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	$x \in [0, 1]$
DoPreSelection	yes

Table 8.12: Optimised parameters of the boosted decision tree classifier used in the MVA event selection. See section 4.5.6 for detailed explanations of variables.

8.11 $\sqrt{s} = 3\,\text{TeV}$ analysis

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 in the analyses, which have not been mentioned, will be highlighted in this section.

Cross sections of the samples considered in this study are listed in table 8.14. 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 8.16 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 $\Sigma B_{4jets} < 2.3$, $-\log(y_{34}) < 3.6$. The selection efficiencies after lepton veto, the mutually exclusive cuts, and the valid jet pairing for individual processes are shown in table B.1.

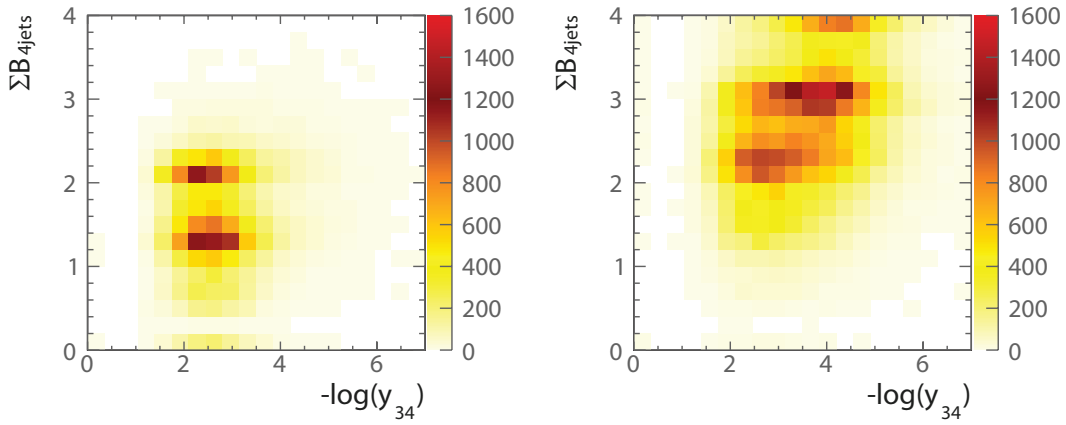
The pre-selection cuts at $\sqrt{s} = 3\,\text{TeV}$ use the same cut on m_{HH}. The cut on b-jet tag is different because the performance of flavour tagging is worse at $\sqrt{s} = 3\,\text{TeV}$ in comparison to the performance at $\sqrt{s} = 1.4\,\text{TeV}$. Figure 8.18 shows the distribution of the highest b-jet tag value, where the cut above 0.7 helps to reduce background events with no b quarks in final states. Figure 8.17 shows the distribution of the invariant mass of the two Higgs system, where the cut above 150 GeV is effective against samples with

$\sqrt{s} = 1.4\,\text{TeV}$	N	ε_{presel}	ε_{MVA}	N_{MVA}
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$				
$b\overline{b}W^+W^-\nu_e\overline{\nu}_e$, hadronic	27.9	59.8%	8.2%	1.29
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$				
$b\overline{b}b\overline{b}\nu_e\overline{\nu}_e$	67.6	15.4%	0.5%	0.05
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$ other	128.0	20.4%	1.7%	0.45
$e^+e^- \to q_lq_lH\nu\overline{\nu}$	1290	39.5%	0.05%	0.29
$e^+e^- \to c\overline{c}H\nu\overline{\nu}$	540	31.6%	0.1%	0.16
$e^+e^- \to b\overline{b}H\nu\overline{\nu}$	465	24.7%	0.3%	0.37
$e^+e^- \to qqqq$	1867650	3.3%	-	-
$e^+e^- \to qqqq\ell\ell$	93150	0.3%	-	-
$e^+e^- \to qqqq\ell\nu$	165600	9.8%	0.01%	2.06
$e^+e^- \to qqqq\nu\overline{\nu}$	34800	16.5%	0.002%	0.10
$e^+e^- \to qq$	6014250	0.8%	-	-
$e^+e^- \to qq\ell\nu$	6464550	0.9%	-	-
$e^+e^- \to qq\ell\ell$	4088700	0.08%	-	-
$e^+e^- \to qq\nu\nu$	1181550	4.0%	-	-
$e^\pm\gamma(BS) \to e^\pm qqqq$	2606625	0.3%	-	-
$e^\pm\gamma(EPA) \to e^\pm qqqq$	861000	0.3%	-	-
$e^\pm\gamma(BS) \to \nu qqqq$	178987.5	25.7%	0.005%	2.05
$e^\pm\gamma(EPA) \to \nu qqqq$	52050	12.5%	0.004%	0.27
$e^\pm\gamma(BS) \to qqH\nu$	35437.5	30.7%	0.02%	2.16
$e^\pm\gamma(EPA) \to qqH\nu$	10170.0	16.1%	0.06%	0.95
$\gamma(BS)\gamma(BS) \to qqqq$	2054951.5	0.2%	-	-
$\gamma(BS)\gamma(EPA) \to qqqq$	4521037.5	0.4%	-	-
$\gamma(EPA)\gamma(BS) \to qqqq$	4539150.0	0.3%	-	-
$\gamma(EPA)\gamma(EPA) \to qqqq$	1129500.0	0.3%	-	-

Table 8.13: List of signal and background events with selection efficiency and number of events at $\sqrt{s} = 1.4\,\text{TeV}$, assuming an integrated luminosity 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. The entries marked with "-" represent numbers less than 0.01. q can be u, d, s, b or t.

Process	$\sigma(\sqrt{s} = 3\,TeV)$ / fb
$e^+e^- \to HH\nu_e\overline{\nu}_e$	0.588
$e^+e^- \to HH\nu_e\overline{\nu}_e \to b\overline{b}W^+W^-$,hadronic	0.07
$e^+e^- \to HH\nu_e\overline{\nu}_e \to b\overline{b}b\overline{b}\nu_e\overline{\nu}_e$	0.19
$e^+e^- \to HH \to$ others	0.34
$e^+e^- \to q_l\overline{q}_l H\nu\overline{\nu}$	3.06
$e^+e^- \to c\overline{c}H\nu\overline{\nu}$	1.15
$e^+e^- \to b\overline{b}H\nu\overline{\nu}$	1.78
$e^+e^- \to qqqq$	546.5*
$e^+e^- \to qqqq\ell\ell$	169.3*
$e^+e^- \to qqqq\ell\nu$	106.6*
$e^+e^- \to qqqq\nu\overline{\nu}$	71.5*
$e^+e^- \to qq$	2948.9
$e^+e^- \to qq\ell\nu$	5561.1
$e^+e^- \to qq\ell\ell$	3319.6
$e^+e^- \to qq\nu\nu$	1317.5
$e^\pm\gamma(\text{BS}) \to e^\pm qqqq$	2536.3*
$e^\pm\gamma(\text{EPA}) \to e^\pm qqqq$	575.7*
$e^\pm\gamma(\text{BS}) \to \nu qqqq$	524.8*
$e^\pm\gamma(\text{EPA}) \to \nu qqqq$	108.4*
$e^\pm\gamma(\text{BS}) \to qqH\nu$	117.1*
$e^\pm\gamma(\text{EPA}) \to qqH\nu$	22.4*
$\gamma(\text{BS})\gamma(\text{BS}) \to qqqq$	13050.3*
$\gamma(\text{BS})\gamma(\text{EPA}) \to qqqq$	2420.6*
$\gamma(\text{EPA})\gamma(\text{BS}) \to qqqq$	2423.1*
$\gamma(\text{EPA})\gamma(\text{EPA}) \to qqqq$	402.7*

Table 8.14: 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. 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 8.16: The two-dimensional distributions of sum of b-jet tag values against $-\log(y_{34})$ for: 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 cases where events are clustered into four jets.

two-quark final states. The fractions of events passing each pre-section cut for individual processes are listed in table B.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 double Higgs system due to a higher centre-of-mass energy. 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 of the Boosted Decision Tree classifier are the same. The efficiencies of the MVA event selection and the numbers of events after the MVA event selection are listed in table 8.15. Background processes that are dominant after the MVA event selection are almost identical to those at $\sqrt{s} = 1.4\,\text{TeV}$. Hence see section 8.10 for discussion.

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

The final analysis is the semi-leptonic W^+W^- decay of $e^+e^- \to HH\nu_e\overline{\nu}_e \to b\bar{b}W^+W^-\nu\overline{\nu}$ at $\sqrt{s} = 3\,\text{TeV}$. The semi-leptonic decay analysis at $\sqrt{s} = 1.4\,\text{TeV}$ was also performed, but only the semi-leptonic decay analysis at $\sqrt{s} = 3\,\text{TeV}$ is presented.

$\sqrt{s} = 3\,\text{TeV}$	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	146.0	61.7%	11.6%	9.89
$e^+e^- \to HH\nu_e\overline{\nu}_e \to$				
$b\bar{b}b\bar{b}\nu_e\overline{\nu}_e$	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
$e^+e^- \to q_l q_l H\nu\overline{\nu}$	6120	36.0%	0.4%	9.42
$e^+e^- \to c\bar{c}H\nu\overline{\nu}$	2300	26.3%	0.5%	3.13
$e^+e^- \to b\bar{b}H\nu\overline{\nu}$	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%	-	-
$e^\pm\gamma(\text{BS}) \to e^\pm qqqq$	4007354	0.7%	-	-
$e^\pm\gamma(\text{EPA}) \to e^\pm qqqq$	1151200	0.4%	-	-
$e^\pm\gamma(\text{BS}) \to \nu qqqq$	829184	16.4%	0.04%	61.0
$e^\pm\gamma(\text{EPA}) \to \nu qqqq$	216800	7.6%	0.04%	6.0
$e^\pm\gamma(\text{BS}) \to qqH\nu$	185018	30.2%	0.2%	121.7
$e^\pm\gamma(\text{EPA}) \to qqH\nu$	46800.0	15.3%	0.2%	18.1
$\gamma(\text{BS})\gamma(\text{BS}) \to qqqq$	18009414	1.6%	-	-
$\gamma(\text{BS})\gamma(\text{EPA}) \to qqqq$	3824548	1.0%	-	-
$\gamma(\text{EPA})\gamma(\text{BS}) \to qqqq$	3828498	1.0%	-	-
$\gamma(\text{EPA})\gamma(\text{EPA}) \to qqqq$	805400	0.6%	-	-

Table 8.15: 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. The entries marked with "-" represent numbers 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.

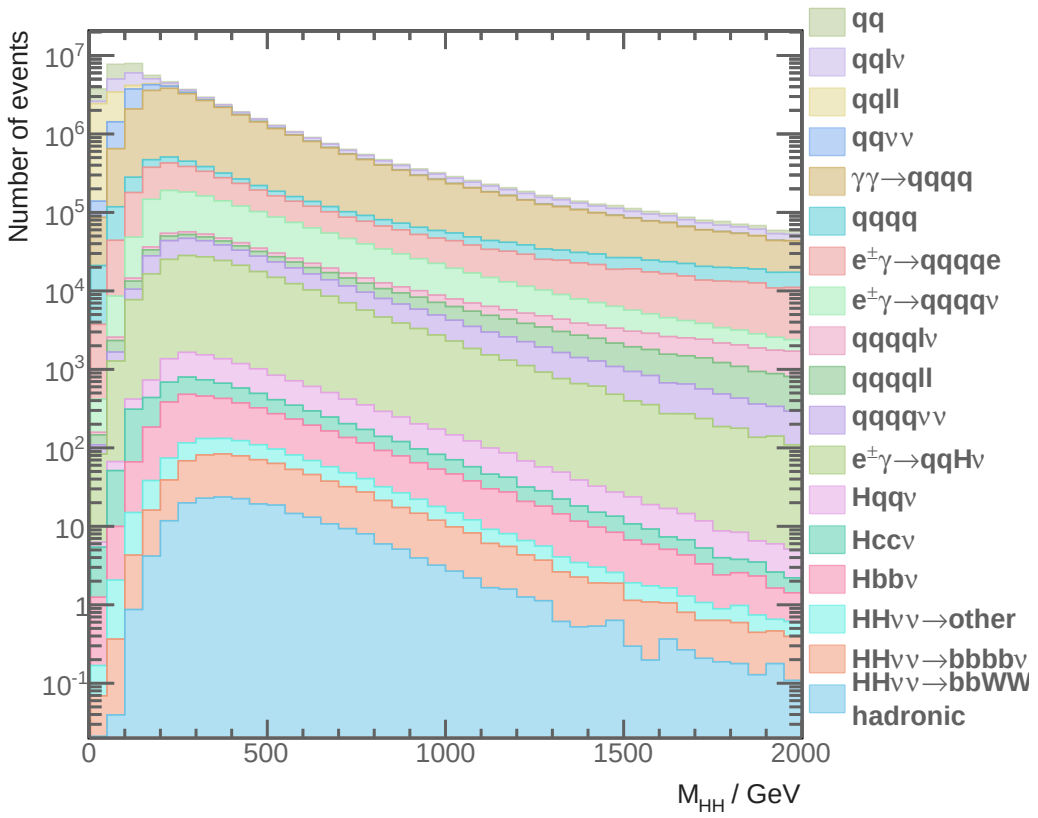


Figure 8.17: Distributions of the invariant mass of the two Higgs system for $\sqrt{s} = 3\,\text{TeV}$, assuming an integrated luminosity of $2000\,\text{fb}^{-1}$.

The strategy of the semi-leptonic decay analysis is very similar to the hadronic decay analysis. The main differences are that there is one lepton in the final state and the final state has four quarks instead of six. The H_{WW^*} and W bosons can not be reconstructed due to the leptonic decay of one of the Ws. 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 $\sqrt{s} = 3\,\text{TeV}$. There are no mutually exclusive cuts since there is no semi-leptonic analysis for the $\text{HH} \to \text{b}\bar{\text{b}}\text{b}\bar{\text{b}}$ sub-channel.

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.

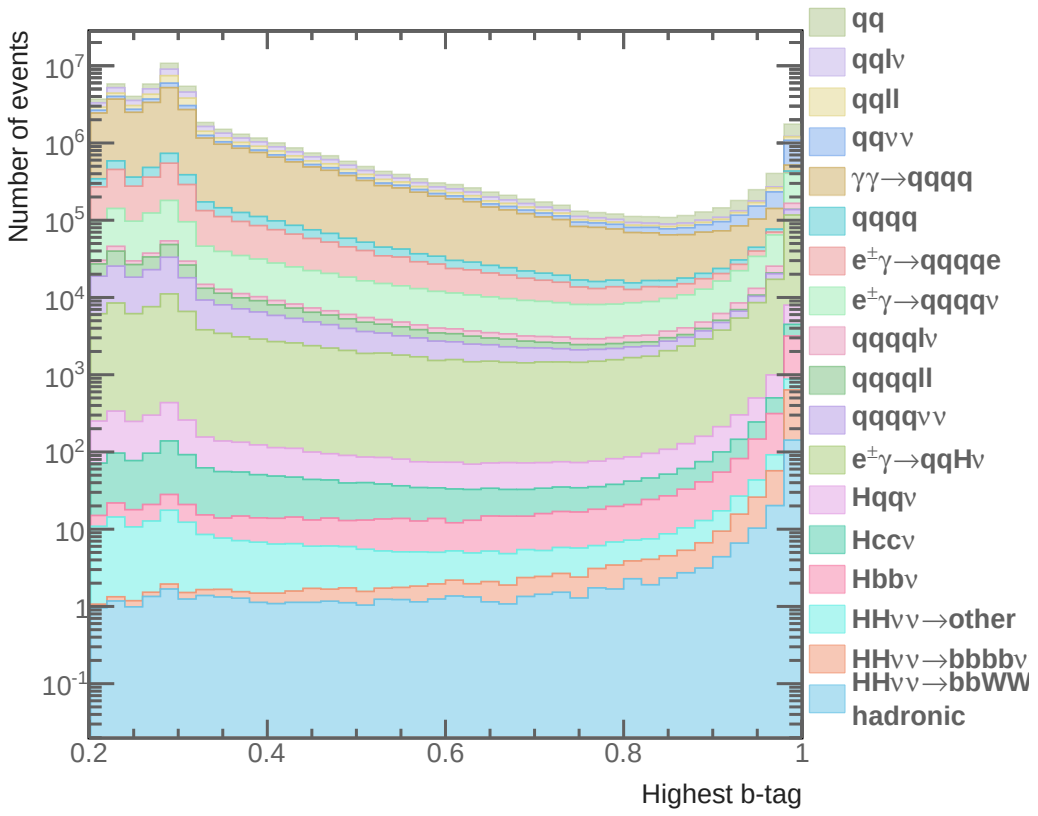


Figure 8.18: Distributions of the highest b-jet tag value for $\sqrt{s} = 3\,\text{TeV}$, assuming an integrated luminosity of $2000\,\text{fb}^{-1}$.

Variables used in the MVA classifier, listed in table 8.16, are a subset of a reduced set of the variables used in the hadronic decay analysis, as H_{WW^*} and one W can not be reconstructed in the semi-hadronic decay analysis. For the same reason, the cuts to aid the MVA are only require that the invariant mass of $H_{bb} < 500\,\text{GeV}$ and the invariant mass of the double Higgs system $< 3000\,\text{GeV}$.

Table 8.17 lists the selection efficiency and number of events after the MVA event selection for individual processes at $\sqrt{s} = 3\,\text{TeV}$, assuming an integrated luminosity of $2000\,\text{fb}^{-1}$. The dominant background processes are almost identical to those for the hadronic decay analysis at $\sqrt{s} = 3\,\text{TeV}$.

Category	Variable
Invariant mass	$m_{\mathrm{H_{bb}}}$, m_{W}, m_{HH}
Energy and momentum	E_{mis}, $p_{\mathrm{T}\mathrm{H_{bb}}}$, $p_{\mathrm{T}\mathrm{W}}$, $p_{\mathrm{T}\mathrm{HH}}$
laboratory-frame angles	θ_{mis}, $A_{\mathrm{H_{bb}}}$, A_{W}, A_{HH}
Rest-frame frames	$\cos\theta^*_{\mathrm{H_{bb}}}$, $\cos\theta^*_{\mathrm{HH}}$
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_{W}, $C^1_{\mathrm{H_{bb}}}$, C^1_{W}
Particle number	$N_{\mathrm{H_{bb}}}$, N_{W}

Table 8.16: 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 $\sqrt{s}=3\,\mathrm{TeV}$.

8.13 Results and interpretation

The numbers of signal events and background events and signal 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 8.18. The significance is defined as the number of signal events divided by the square root of the sum of the signal and background events. The analysis at $\sqrt{s}=1.4\,\mathrm{TeV}$ assumes an integrated luminosity of $1500\,\mathrm{fb}^{-1}$, while 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 numbers of signal and background events after the MVA selection are 1.79 and 8.41 respectively. For the hadronic analysis at $\sqrt{s}=3\,\mathrm{TeV}$, the respective numbers of the signal and background events after the MVA selection are 15.45 and 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 numbers of the signal and background events after the MVA selection are 31.24 and 3612.39 respectively.

The expected uncertainties on the measurement of the double Higgs production cross sections are estimated to be the inverse of the significance [3]:

$$\frac{\Delta\left[\sigma\left(\mathrm{HH}\nu_{\mathrm{e}}\overline{\nu}_{\mathrm{e}}\right)\right]}{\sigma\left(\mathrm{HH}\nu_{\mathrm{e}}\overline{\nu}_{\mathrm{e}}\right)}=\begin{cases}179\%, & \text{at } \sqrt{s}=1.4\,\mathrm{TeV},\\ 92\%, & \text{at } \sqrt{s}=3\,\mathrm{TeV}.\end{cases} \tag{8.11}$$

The expected uncertainty 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 8.17: 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. The entries marked with "-" represent numbers 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.

Process	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 8.18: 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 analysis at $\sqrt{s} = 1.4\,\text{TeV}$ assumes an integrated luminosity of $1500\,\text{fb}^{-1}$, whilst that of $\sqrt{s} = 3\,\text{TeV}$ 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 [23]:

$$\frac{\Delta g_{HHH}}{g_{HHH}} \approx \kappa \cdot \frac{\Delta\left[\sigma\left(HH\nu_e\overline{\nu}_e\right)\right]}{\sigma\left(HH\nu_e\overline{\nu}_e\right)}, \tag{8.12}$$

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 8.19 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 $\sqrt{s} = 3\,\text{TeV}$. At the SM g_{HHH} value, the coefficient κ is 1.22 at $\sqrt{s} = 1.4\,\text{TeV}$, and 1.47 at $\sqrt{s} = 3\,\text{TeV}$.

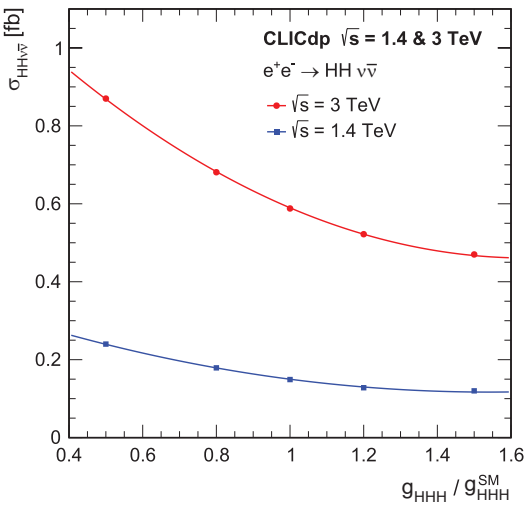


Figure 8.19: Cross sections of $e^+e^- \to HH\nu_e\overline{\nu}_e$ process as a function of the ratio g_{HHH}/g^{SM}_{HHH} at $\sqrt{s} = 1.4\,\text{TeV}$ and 3 TeV, adapted from [23].

The uncertainties on measurement of the Higgs trilinear self coupling, g_{HHH}, from $e^+e^- \to HH\nu_e\overline{\nu}_e \to b\overline{b}W^+W^-\nu_e\overline{\nu}_e$ analyses, are hence obtained via equation 8.12:

$$\frac{\Delta g_{HHH}}{g_{HHH}} \approx \begin{cases} 218\%, & \text{at } \sqrt{s} = 1.4\,\text{TeV}, \\ 135\%, & \text{at } \sqrt{s} = 3\,\text{TeV}. \end{cases} \tag{8.13}$$

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

$$\frac{\Delta g_{HHH}}{g_{HHH}} \approx \begin{cases} 163\%, & \text{at } \sqrt{s} = 1.4\,\text{TeV}, \\ 97\%, & \text{at } \sqrt{s} = 3\,\text{TeV}. \end{cases} \tag{8.14}$$

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 $HH \to b\overline{b}b\overline{b}$ [23] sub-channels are combined, the expected uncertainties on the double Higgs production cross section measurements are:

$$\frac{\Delta\left[\sigma\left(HH\nu_e\overline{\nu}_e\right)\right]}{\sigma\left(HH\nu_e\overline{\nu}_e\right)} = \begin{cases} 44\%, & \text{at } \sqrt{s} = 1.4\,\text{TeV}, \\ 20\%, & \text{at } \sqrt{s} = 3\,\text{TeV}, \end{cases} \tag{8.15}$$

This translates to uncertainties on the measurement of the Higgs trilinear self coupling g_{HHH}, via equation 8.12, with unpolarised beams:

$$\frac{\Delta g_{HHH}}{g_{HHH}} \approx \begin{cases} 54\%, & \text{at } \sqrt{s} = 1.4\,\text{TeV}, \\ 29\%, & \text{at } \sqrt{s} = 3\,\text{TeV}. \end{cases} \tag{8.16}$$

8.14 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 attempted using the results of the hadronic analysis at $\sqrt{s} =$

3 TeV. The integrated luminosity in this section is assumed to be 3000 fb^{-1} at $\sqrt{s}$ = 3 TeV to reflect the updated CLIC running scenario [110]. A simple scaling is applied to the number of events after the MVA event selection for the analysis at $\sqrt{s} = 3$ TeV to adapt to the change in the luminosity.

The $e^+e^- \to HH\nu_e\overline{\nu}_e$ events with non-SM g_{HHH} and g_{WWHH} couplings were generated and reconstructed in the same way as previously described. The few number of samples for varying g_{HHH} and g_{WWHH} are due to the limited computational resource. The cross section of the $e^+e^- \to HH\nu_e\overline{\nu}_e$ for certain values of g_{HHH} and g_{WWHH} is shown in figure 8.20. Cross sections are normalised to the value at the SM coupling. Around the SM coupling values, the cross section increases with the decrease of g_{HHH} and with the increase of g_{WWHH}.

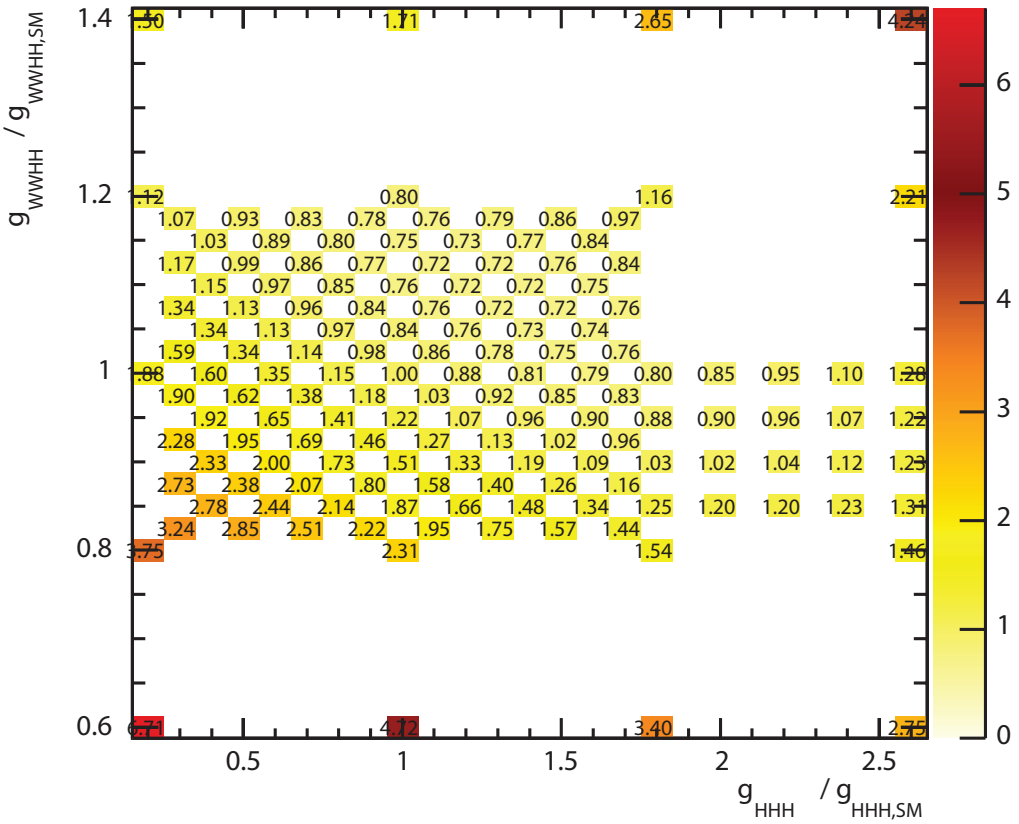


Figure 8.20: Normalised cross section for the $e^+e^- \to HH\nu_e\overline{\nu}_e$ process as a function of the g_{HHH}/g_{HHH}^{SM} and g_{WWHH}/g_{WWHH}^{SM} at $\sqrt{s} = 3$ TeV. All cross sections are normalised to the cross section at the SM couplings value.

These generated $e^+e^- \to HH\nu_e\overline{\nu}_e$ events with non-SM coupling went through the analysis chain described in this chapter with the same pre-selection cuts and the same MVA classifier applied. Variables that increase with the increasing of the centre-of-mass energies are suitable for studying the cross section dependence on g_{HHH} and g_{WWHH}. Two

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

Events are subsequently binned using the kinematic variables. 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 eight kinematic bins. The distributions of the variables H_T and m_{HH} for selected samples with different $\mathrm{g_{HHH}}$ and $\mathrm{g_{WWHH}}$ couplings are shown in figure 8.21.

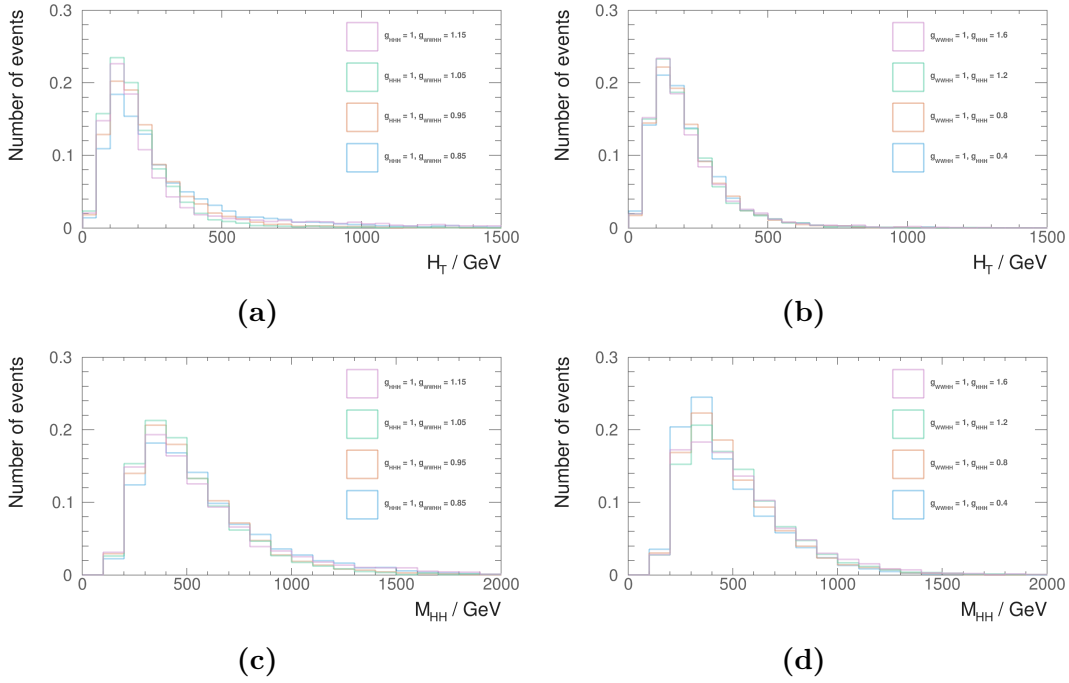


Figure 8.21: The distributions of the variables H_T and m_{HH} for selected samples with different $\mathrm{g_{HHH}}$ and $\mathrm{g_{WWHH}}$ couplings at $\sqrt{s} = 3\,\mathrm{TeV}$, using hadronic $\mathrm{W^+W^-}$ decay of $\mathrm{HH} \to \mathrm{b\bar{b}W^+W^-}$, assuming an integrated luminosity of $3000\,\mathrm{fb^{-1}}$. Area under curve is normalised to unity. The $\mathrm{g_{HHH}}$ and $\mathrm{g_{WWHH}}$ in legends refer to the ratio to the $\mathrm{g^{SM}_{HHH}}$ and $\mathrm{g^{SM}_{WWHH}}$ respectively.

The change in the m_{HH} and H_T distributions for non-SM coupling samples is quantised using a χ^2 function:

$$\chi^2 = \sum_i^8 \frac{(N_i^e - N_i^o)^2}{N_i^e}, \tag{8.17}$$

where N_i^e is the number of expected events in the kinematic bin i in a non-SM coupling sample; and N_i^o 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 function is summed over all kinematic bins. By construction, the SM coupling sample has a χ^2 of 0.

Two sub-channels, hadronic W^+W^- decay of $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$ [23], 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 sample with the SM coupling is treated as a data template set, and 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 values are calculated using these generated data sets as the observed data (N_i^e in equation 8.17). The χ^2 values are then averaged over the number of data sets (100000) and normalised such that the χ^2 at the SM couplings is 0. Figure 8.22 shows the normalised χ^2 after averaging over 100000 toy MC experiments for certain values of g_{HHH}/g_{HHH}^{SM} and g_{WWHH}/g_{WWHH}^{SM}.

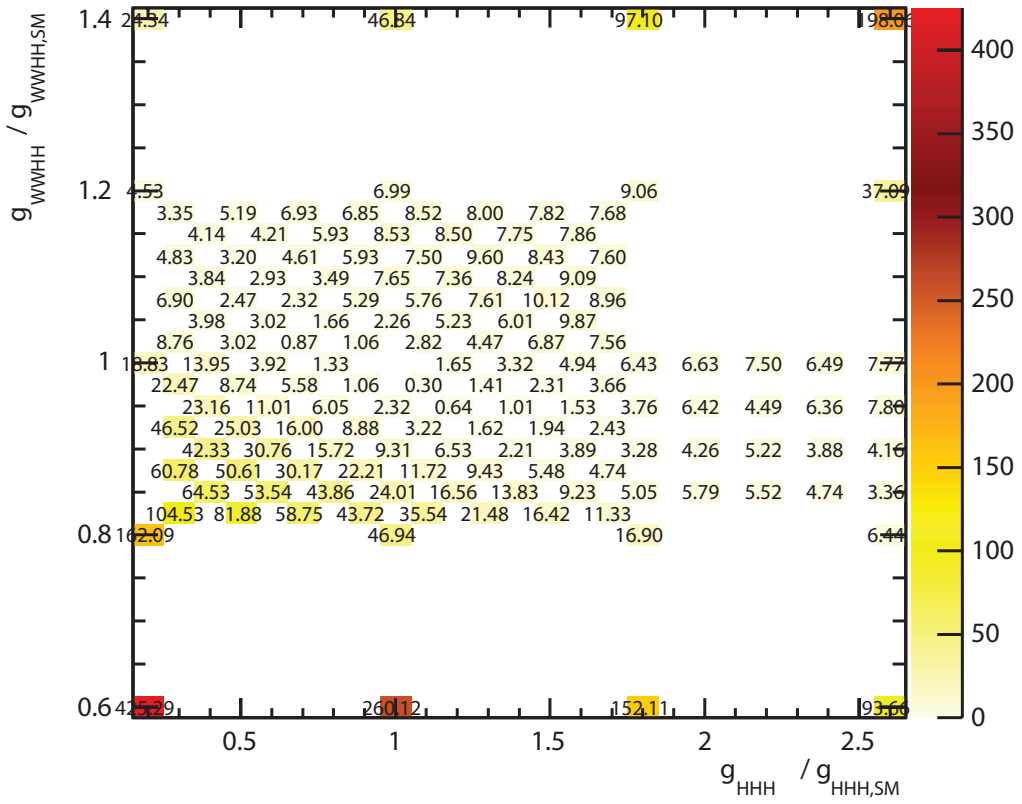


Figure 8.22: Normalised χ^2 after averaging over 100000 toy MC experiments for certain values of g_{HHH}/g_{HHH}^{SM} and g_{WWHH}/g_{WWHH}^{SM}, by combining hadronic decay $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$ sub-channels, assuming an integrated luminosity of 3000 fb^{-1}. Normalisation is set such that the χ^2 at the SM coupling point is 0.

Since there are two couplings in this χ^2 surface, the number of degrees of freedom 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 8.22. The interpolation is performed by fitting cubic functions along the points on the antidiagonal. Figure 8.23 shows the contour of 68% confidence of the measurements of the g_{HHH} and g_{WWHH}. The shape of the contour largely corresponds to the plot for the cross section. The roughness of the plot is due to the limited samples of non-SM couplings. The counter can be sliced in one dimension to extract the uncertainty of the measurements of one coupling for a given value of the other coupling. For example:

$$\frac{\Delta g_{WWHH}}{g_{WWHH}} \simeq 4.9\%, \text{ for } g_{HHH} = g^{SM}_{HHH}, \tag{8.18}$$

$$\frac{\Delta g_{HHH}}{g_{HHH}} \simeq 29\%, \text{ for } g_{WWHH} = g^{SM}_{WWHH}. \tag{8.19}$$

The statistical precisions on the measurements of g_{WWHH} and g_{HHH} are much better at CLIC than at the LHC or high-luminosity LHC [20, 21, 99]. The LHC can probe g_{HHH} to 20−60% and the high-luminosity LHC can determine the g_{HHH} within 40% uncertainty.

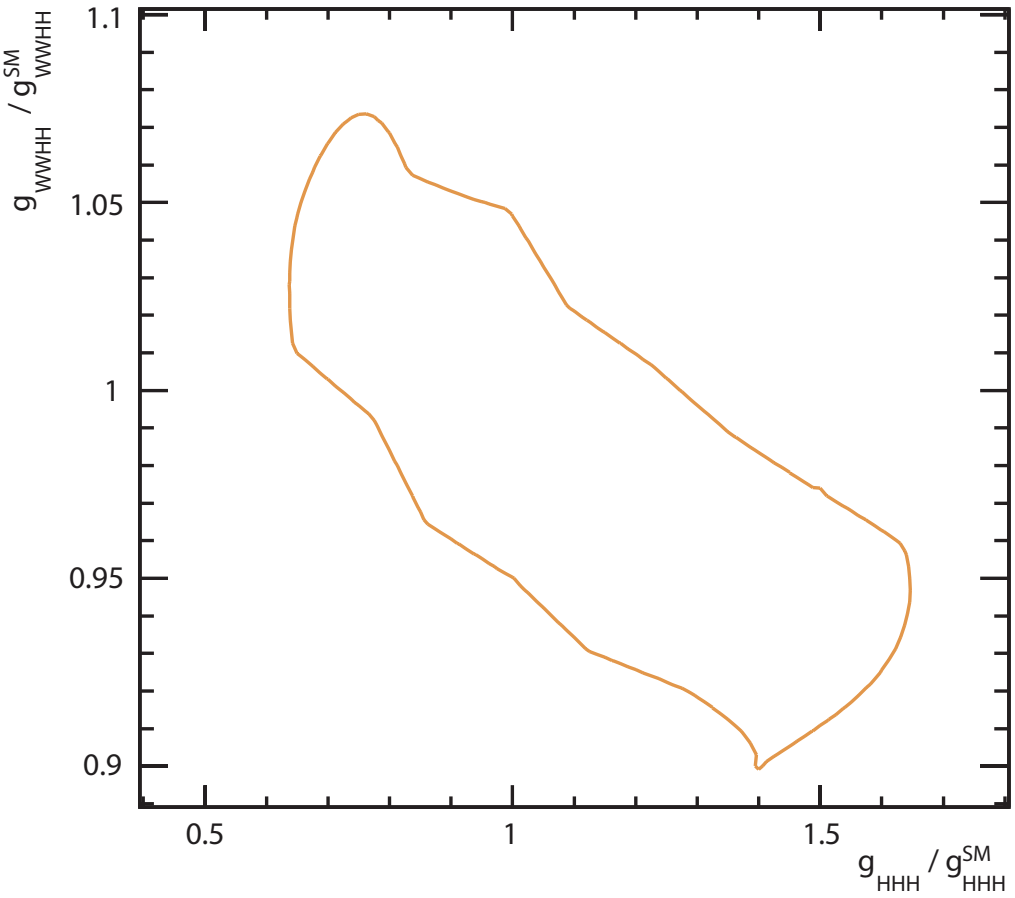


Figure 8.23: Contour plot of 68% confidence ($\chi^2 = 2.3$), after averaging toy MC experiments, as a function of g_{HHH}/g^{SM}_{HHH} and g_{WWHH}/g^{SM}_{WWHH}, after combining hadronic W^+W^- decay of $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$ sub-channels, assuming an integrated luminosity of $3000\,fb^{-1}$.

Chapter 9

Summary

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

— Confucius, 551 BC – 479 BC

In chapter 5, a set of new photon reconstruction algorithms developed in PandoraPFA are discussed. Using the ILD detector model, the single photon reconstruction efficiency is above 98% for photons with energies above 2GeVand above 99.5% for photons with energies above 100 GeV. 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 a two-photon sample with photon energies of 500 GeV and 50 GeV, the average numbers of photons and particles beyond a distance separation of 20 mm are both less than 2.05, where the true value is 2. The minimal distance separation of resolved photon pairs is reduced to 6 mm for two photons with the same energy, and 10 mm for two photons with different energies. The jet energy resolution has been improved for $\sqrt{s} = 360$ GeV and 500 GeV. The photon confusion terms, except at $\sqrt{s} = 91$ GeV, have been reduced to 0.9%.

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. The efficiency of the tau decay mode classification decreases with an increase of the centre-of-mass energy and with the increasing ECAL cell sizes. The sensitivity of ε_{had} to different cell sizes is stronger at high centre-of-mass

energies. For the ILC at $\sqrt{s} = 250\,\text{GeV}$ or CLIC at $\sqrt{s} = 350\,\text{GeV}$, an ECAL size of 10 mm or fewer is sufficient to achieve a ε_{had} of 92%. For a linear collider operating at a centre-of-mass energy above a few hundred GeV, such as the ILC at $\sqrt{s} = 500\,\text{GeV}$ or CLIC at $\sqrt{s} = 1.4\,\text{TeV}$ or 3 TeV, it is preferable to have a small ECAL cell size, i.e. 3 mm, for the best tau decay mode separation, as ε_{had} decreases drastically with an increasing ECAL cell size.

Chapter 7 presents a proof-of-principle demonstration that the generator-level pion energy fraction correlation can be reconstructed at the analysis level. The analysis contains several important steps: tau identification; kinematic reconstruction of the energies of the taus; the classification of the $\tau^{\pm} \to \pi^{\pm}\nu_{\tau}$ decay mode; and the reconstruction of the tau pair polarisation correlations.

In chapter 8, 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 $\sqrt{s} = 1.4\,\text{TeV}$ and $\sqrt{s} = 3\,\text{TeV}$ are performed. The significance of the signal events are 0.56 and 1.09, assuming an integrated luminosity of $1500\,\text{fb}^{-1}$ and $2000\,\text{fb}^{-1}$, for $\sqrt{s}$ = 1.4 TeV and $\sqrt{s} = 3\,\text{TeV}$ 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:

$$\frac{\Delta g_{HHH}}{g_{HHH}} = \begin{cases} 218\%, & \text{at } \sqrt{s} = 1.4\,\text{TeV}, \\ 135\%, & \text{at } \sqrt{s} = 3\,\text{TeV}. \end{cases} \tag{9.1}$$

When the analysis of both $HH \to b\bar{b}W^+W^-$ and $HH \to b\bar{b}b\bar{b}$ sub-channels are combined at $\sqrt{s} = 3\,\text{TeV}$, assuming an integrated luminosity of $3000\,\text{fb}^{-1}$, the simultaneous extraction of the uncertainty on the measurement of g_{HHH} and g_{WWHH} yields:

$$\frac{\Delta g_{WWHH}}{g_{WWHH}} \simeq 4.9\%, \text{ for } g_{HHH} = g^{SM}_{HHH}, \tag{9.2}$$

$$\frac{\Delta g_{HHH}}{g_{HHH}} \simeq 29\%, \text{ for } g_{WWHH} = g^{SM}_{WWHH}. \tag{9.3}$$

The statistical precisions on the measurements of g_{WWHH} and g_{HHH} are much better at CLIC than at the LHC or high-luminosity upgraded LHC [20, 21, 99].

Appendix A

Generation Parameters

'Thank you Mario! But our Princess is in another castle!'
— Toad, Super Mario Bros, 1985

Particle masses and widths used to generate SM samples for studies with CLIC detectors, used in chapter 8, are listed in table A.1. The Higgs boson mass is specified for individual samples.

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 A.1: Particle masses and widths used for the generation of SM samples for CLIC detectors, taken from [2].

Appendix B

Double Higgs Boson Production Analysis

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

Here are extra tables for hadronic decay analysis at $\sqrt{s} = 3\,\text{TeV}$ in chapter 8. Table B.1 shows the expected number of events, before cuts and after successive cuts: the lepton veto, $\text{HH} \to \text{b}\bar{\text{b}}\text{W}^+\text{W}^-/\text{HH} \to \text{b}\bar{\text{b}}\text{b}\bar{\text{b}}$ separation, and valid jet pairing, for the signal and background events at $\sqrt{s} = 3\,\text{TeV}$, assuming an integrated luminosity of $2000\,\text{fb}^{-1}$. Table B.2 shows the expected number of events after successive cuts: invariant mass of the two Higgs system $> 150\,\text{GeV}$, and the highest b-jet tag value > 0.7. All cuts include the lepton veto, $\text{HH} \to \text{b}\bar{\text{b}}\text{W}^+\text{W}^-/\text{HH} \to \text{b}\bar{\text{b}}\text{b}\bar{\text{b}}$ separation, and valid jet pairing.

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

Process	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 B.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, $\mathrm{HH} \to b\overline{b}W^+W^-/\mathrm{HH} \to b\overline{b}b\overline{b}$ separation, and valid jet pairing. The 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.

Bibliography

[1] J. Brau *et al.*, *International Linear Collider reference design report. 1: Executive summary. 2: Physics at the ILC. 3: Accelerator. 4: Detectors*, (2007).

[2] L. Linssen, A. Miyamoto, M. Stanitzki, and H. Weerts, *Physics and Detectors at CLIC: CLIC Conceptual Design Report*, (2012).

[3] C. Patrignani and P. D. Group, *Review of Particle Physics*, Chinese Physics C **40**, 100001 (2016).

[4] M. Thomson, *Modern Particle Physics*, (Cambridge University Press, 2013).

[5] D. Tong, *Lectures on Quantum Field Theory*, 2006.

[6] B. Gripaios, *Lectures on Gauge Field Theory*, 2017.

[7] SLD Electroweak Group, DELPHI, ALEPH, SLD, SLD Heavy Flavour Group, OPAL, LEP Electroweak Working Group, L3 Collaboration, S. Schael *et al.*, *Precision electroweak measurements on the Z resonance*, Phys. Rept. **427**, 257 (2006).

[8] D0 Collaboration, S. Abachi *et al.*, *Observation of the top quark*, Phys. Rev. Lett. **74**, 2632 (1995).

[9] DONUT Collaboration, K. Kodama *et al.*, *Observation of tau neutrino interactions*, Phys. Lett. **B504**, 218 (2001).

[10] ATLAS Collaboration, G. Aad *et al.*, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. **B716**, 1 (2012).

[11] CMS Collaboration, S. Chatrchyan *et al.*, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. **B716**, 30 (2012).

[12] D. B. Kaplan and H. Georgi, *SU(2) $\times$ U(1) Breaking by Vacuum Misalignment*,

Phys. Lett. **B136**, 183 (1984).

[13] W. D. Goldberger, B. Grinstein, and W. Skiba, *Distinguishing the Higgs boson from the dilaton at the Large Hadron Collider*, Phys. Rev. Lett. **100**, 111802 (2008).

[14] M. Peskin and D. Schroeder, *An Introduction to Quantum Field Theory*, (Avalon Publishing, 1995).

[15] Y. Nambu, *Quasiparticles and Gauge Invariance in the Theory of Superconductivity*, Phys. Rev. **117**, 648 (1960).

[16] J. Goldstone, *Field Theories with Superconductor Solutions*, Nuovo Cim. **19**, 154 (1961).

[17] S. Weinberg, *A Model of Leptons*, Phys. Rev. Lett. **19**, 1264 (1967).

[18] 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.

[19] G. F. Giudice, C. Grojean, A. Pomarol, and R. Rattazzi, *The Strongly-Interacting Light Higgs*, JHEP **06**, 045 (2007).

[20] R. Contino, C. Grojean, M. Moretti, F. Piccinini, and R. Rattazzi, *Strong Double Higgs Production at the LHC*, JHEP **05**, 089 (2010).

[21] R. Contino, C. Grojean, D. Pappadopulo, R. Rattazzi, and A. Thamm, *Strong Higgs Interactions at a Linear Collider*, JHEP **02**, 006 (2014).

[22] V. Barger, T. Han, P. Langacker, B. McElrath, and P. Zerwas, *Effects of genuine dimension-six Higgs operators*, Phys. Rev. **D67**, 115001 (2003).

[23] H. Abramowicz *et al.*, *Higgs Physics at the CLIC Electron−Positron Linear Collider*, (2016).

[24] ALEPH Collaboration, S. Schael *et al.*, *Branching ratios and spectral functions of tau decays: Final ALEPH measurements and physics implications*, Phys. Rept. **421**, 191 (2005).

[25] H1 Collaboration, A. Aktas *et al.*, *Tau lepton production in ep collisions at HERA*, Eur. Phys. J. **C48**, 699 (2006).

[26] DELPHI Collaboration, P. Abreu *et al.*, *A Measurement of the lifetime of the tau lepton*, Phys. Lett. **B267**, 422 (1991).

[27] B. K. Bullock, K. Hagiwara, and A. D. Martin, *Tau pair polarization correlations as a signal for Higgs bosons*, Phys. Lett. **B273**, 501 (1991).

[28] Y.-S. Tsai, *Decay Correlations of Heavy Leptons in* $e^+e^- \to \ell^+\ell^-$, Phys. Rev. **D4**, 2821 (1971), [Erratum: Phys. Rev.D13,771(1976)].

[29] Linear Collider ILD Concept Group Collaboration, T. Abe *et al.*, *The International Large Detector: Letter of Intent*, (2010).

[30] SiD Collaboration, H. Aihara *et al.*, *SiD Letter of Intent*, (2010).

[31] H. Abramowicz *et al.*, *The International Linear Collider Technical Design Report - Volume 4: Detectors*, (2013).

[32] M. Aicheler *et al.*, *A Multi-TeV Linear Collider Based on CLIC Technology*, (2012).

[33] O. S. Bruning *et al.*, *LHC Design Report Vol.1: The LHC Main Ring*, (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] CLIC Detector and Physics Study Collaboration, 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.

[36] M. Thomson, *Measurement of the mass of the W boson at LEP*, Eur. Phys. J. **C33**, S689 (2004).

[37] M. Thomson, *Particle Flow Calorimetry and the PandoraPFA Algorithm*, Nucl. Instrum. Meth. **A611**, 25 (2009).

[38] I. G. Knowles and G. D. Lafferty, *Hadronization in* Z^0 *decay*, J. Phys. **G23**, 731 (1997).

[39] M. Green, *Electron−Positron Physics at the Z*, (Taylor & Francis, 1998).

[40] J. S. Marshall, A. Münnich, and M. A. Thomson, *Performance of Particle Flow*

Calorimetry at CLIC, Nucl. Instrum. Meth. **A700**, 153 (2013).

[41] P. Mora de Freitas and H. Videau, *Detector simulation with MOKKA / GEANT4: Present and future*, p. 623 (2002).

[42] CALICE Collaboration, M. Ramilli, *Hadronic models validation in GEANT4 with CALICE highly granular calorimeters*, J. Phys. Conf. Ser. **404**, 012050 (2012).

[43] C. Adloff *et al.*, *Response of the CALICE Si-W electromagnetic calorimeter physics prototype to electrons*, Nucl. Instrum. Meth. **A608**, 372 (2009).

[44] CALICE Collaboration, C. Adloff *et al.*, *Shower development of particles with momenta from 1 to 10 GeV in the CALICE Scintillator-Tungsten HCAL*, JINST **9**, P01004 (2014).

[45] CALICE Collaboration, C. Adloff, *CALICE Report to the DESY Physics Research Committee, April 2011*, (2011).

[46] B. Parker *et al.*, *Functional Requirements on the Design of the Detectors and the Interaction Region of an e^+e^- Linear Collider with a Push-Pull Arrangement of Detectors*, (2009).

[47] CALICE Collaboration, *Construction and performance of a silicon photomultiplier/extruded scintillator tail-catcher and muon-tracker*, JINST **7**, P04015 (2012).

[48] C. Grah and A. Sapronov, *Beam parameter determination using beamstrahlung photons and incoherent pairs*, Journal of Instrumentation **3**, P10004 (2008).

[49] 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).

[50] W. Kilian, T. Ohl, and J. Reuter, *WHIZARD – Simulating Multi-Particle Processes at LHC and ILC*, Eur. Phys. J. **C71** (2011).

[51] M. Moretti, T. Ohl, and J. Reuter, *O'Mega: An Optimizing matrix element generator*, p. 1981 (2001).

[52] T. Sjostrand, *PYTHIA 5.7 and JETSET 7.4: Physics and manual*, (1995).

[53] OPAL Collaboration, G. Alexander *et al.*, *A Comparison of b and uds quark jets to gluon jets*, Z. Phys. **C69**, 543 (1996).

[54] S. Jadach, Z. Was, R. Decker, and J. H. Kuhn, *The tau decay library TAUOLA: Version 2.4*, Comput. Phys. Commun. **76**, 361 (1993).

[55] D. Schulte, *Beam-beam simulations with GUINEA-PIG*, Report No. CERN-PS-99-014-LP, CERN-PS-99-14-LP, CLIC-NOTE-387,CERN-CLIC-NOTE-387, 1999 (unpublished).

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

[57] GEANT4 Collaboration, S. Agostinelli *et al.*, *GEANT4: A Simulation toolkit*, Nucl. Instrum. Meth. **A506**, 250 (2003).

[58] M. Drees and R. M. Godbole, *Mini-jets and large hadronic backgrounds at e^+e^- supercolliders*, Phys. Rev. Lett. **67**, 1189 (1991).

[59] P. Chen, T. L. Barklow, and M. E. Peskin, *Hadron production in gamma gamma collisions as a background for e^+e^- linear colliders*, Phys. Rev. **D49**, 3209 (1994).

[60] 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.

[61] G. A. Schuler and T. Sjostrand, *A Scenario for high-energy gamma gamma interactions*, Z. Phys. **C73**, 677 (1997).

[62] T. Barklow, D. Dannheim, M. O. Sahin, and D. Schulte, *Simulation of $\gamma\gamma \to$ hadrons background at CLIC*, (2012).

[63] F. Gaede, *Marlin and LCCD: Software tools for the ILC*, Nucl. Instrum. Meth. **A559**, 177 (2006).

[64] F. Gaede, S. Aplin, R. Glattauer, C. Rosemann, and G. Voutsinas, *Track reconstruction at the ILC: the ILD tracking software*, J. Phys. Conf. Ser. **513**, 022011 (2014).

[65] D. Contardo, M. Klute, J. Mans, L. Silvestris, and J. Butler, *Technical Proposal for the Phase-II Upgrade of the CMS Detector*, Report No. CERN-LHCC-2015-010. LHCC-P-008. CMS-TDR-15-02, 2015 (unpublished).

[66] J. S. Marshall and M. A. Thomson, *The Pandora Software Development Kit for*

Pattern Recognition, Eur. Phys. J. **C75**, 439 (2015).

[67] J. S. Marshall, *Presentation on PandoraPFA with LC reconstruction*, `https://github.com/PandoraPFA/Documentation/blob/master/Pandora_LC_Reconstruction.pdf`, 2017.

[68] G. F. Sterman and S. Weinberg, *Jets from Quantum Chromodynamics*, Phys. Rev. Lett. **39**, 1436 (1977).

[69] S. Moretti, L. Lonnblad, and T. Sjostrand, *New and old jet clustering algorithms for electron−positron events*, JHEP **08**, 001 (1998).

[70] G. P. Salam, *Towards Jetography*, Eur. Phys. J. **C67**, 637 (2010).

[71] A. Ali and G. Kramer, *Jets and QCD: A Historical Review of the Discovery of the Quark and Gluon Jets and its Impact on QCD*, Eur. Phys. J. **H36**, 245 (2011).

[72] M. Cacciari, G. P. Salam, and G. Soyez, *FastJet User Manual*, Eur. Phys. J. **C72**, 1896 (2012).

[73] M. Cacciari and G. P. Salam, *Dispelling the N^3 myth for the k_t jet-finder*, Phys. Lett. **B641**, 57 (2006).

[74] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber, *Longitudinally invariant K_t clustering algorithms for hadron hadron collisions*, Nucl. Phys. **B406**, 187 (1993).

[75] S. D. Ellis and D. E. Soper, *Successive combination jet algorithm for hadron collisions*, Phys. Rev. **D48**, 3160 (1993).

[76] S. Catani, Y. L. Dokshitzer, M. Olsson, G. Turnock, and B. R. Webber, *New clustering algorithm for multi-jet cross-sections in e^+e^- annihilation*, Phys. Lett. **B269**, 432 (1991).

[77] M. Boronat, J. Fuster, I. Garcia, E. Ros, and M. Vos, *A robust jet reconstruction algorithm for high-energy lepton colliders*, Phys. Lett. **B750**, 95 (2015).

[78] M. Battaglia and F. P., *A Study of $e^+e^- \to H^0A^0 \to b\bar{b}b\bar{b}$ at 3 TeV at CLIC*, CERN Report No. LCD-Note-2010-006, 2010 (unpublished).

[79] A. Hocker *et al.*, *TMVA - Toolkit for Multivariate Data Analysis*, PoS **ACAT**, 040 (2007).

[80] TMVA Core Developer Team Collaboration, J. Therhaag, *TMVA: Toolkit for multivariate data analysis*, AIP Conf.Proc. **1504**, 1013 (2009).

[81] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, (Springer New York, 2009).

[82] Y. Freund and R. E. Schapire, *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*, Journal of Computer and System Sciences **55**, 119 (1997).

[83] 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.

[84] G. Kačarević, *Prelection for $H \to \gamma\gamma$ at 3 TeV*, `https://indico.cern.ch/event/577810/contributions/2485070/attachments/1424897/2185427/GoranKacarevic.pdf`, 2017.

[85] W. R. Nelson, T. M. Jenkins, R. C. McCall, and J. K. Cobb, *Electron-Induced Cascade Showers in Copper and Lead at 1 GeV*, Phys. Rev. **149**, 201 (1966).

[86] G. Bathow, E. Freytag, M. Koebberling, K. Tesch, and R. Kajikawa, *Measurements of the longitudinal and lateral development of electromagnetic cascades in lead, copper and aluminum at 6 gev*, Nucl. Phys. **B20**, 592 (1970).

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

[88] E. Longo and I. Sestili, *Monte Carlo Calculation of Photon Initiated Electromagnetic Showers in Lead Glass*, Nucl. Instrum. Meth. **128**, 283 (1975), [Erratum: Nucl. Instrum. Meth.135,587(1976)].

[89] M. J. Berger and S. M. Seltzer, *Tables of Energy Losses and Ranges of Electrons and Positrons. NASA SP-3012*, NASA Special Publication **3012** (1964).

[90] B. Rossi, *High-energy Particles*, (Prentice-Hall, New York, 1952).

[91] R. L. Ford and W. R. Nelson, *The Egs Code System: Computer Programs for the Monte Carlo Simulation of Electromagnetic Cascade Showers (Version 3)*, (1978).

[92] H. Hirayama, *Revision of the Sternheimer density effect coefficients in PEGS4*,

KEK Report No. KEK-INTERNAL-95-17, 1995 (unpublished).

[93] S. Berge, W. Bernreuther, and S. Kirchner, *Prospects of constraining the Higgs boson's CP nature in the tau decay channel at the LHC*, Phys. Rev. **D92**, 096012 (2015).

[94] F. Gaede and J. Engels, *Marlin et al - A Software Framework for ILC detector R&D*, EUDET Report (2007).

[95] E. Farhi, *Quantum Chromodynamics Test for Jets*, Phys. Rev. Lett. **39**, 1587 (1977).

[96] T. H. Tran, V. Balagura, V. Boudry, J.-C. Brient, and H. Videau, *Reconstruction and classification of tau lepton decays with ILD*, Eur. Phys. J. **C76**, 468 (2016).

[97] M. Reinhard and J. C. Brient, *GARLIC - GAmma Reconstruciton for the LInear Collider*, in *Linear colliders. Proceedings, International Linear Collider Workshop, LCWS08, and International Linear Collider Meeting, ILC08, Chicago, USA, Novermber 16-20, 2008*, 2009, 0902.3042.

[98] D. Jeans, J. C. Brient, and M. Reinhard, *GARLIC: GAmma Reconstruction at a LInear Collider experiment*, JINST **7**, P06003 (2012).

[99] V. Barger, L. L. Everett, C. B. Jackson, and G. Shaughnessy, *Higgs-Pair Production and Measurement of the Triscalar Coupling at LHC(8,14)*, Phys. Lett. **B728**, 433 (2014).

[100] H. Baer *et al.*, *The International Linear Collider Technical Design Report - Volume 2: Physics*, (2013).

[101] D. Lyth, *The Equivalent Photon Approximation*, Journal de Physique Colloques **35**, C2 (1974).

[102] S. Dittmaier *et al.*, *Handbook of LHC Higgs Cross Sections: 2. Differential Distributions*, (2012).

[103] A. Münnich, *TauFinder: A Reconstruction Algorithm for tau Leptons at Linear Colliders*, CERN Report No. LCD-Note-2010-009, 2010 (unpublished).

[104] CLICdp Collaboration, A. Sailer and A. Sapronov, *High Energy Electron Reconstruction in the BeamCal*, (2017).

[105] S. Lukić, *Forward electron tagging in the $H \to \mu\mu$ analysis at 1.4*

TeV, http://indico.cern.ch/event/262809/contributions/1595499/attachments/464689/643931/electronTagging.pdf, 2013.

[106] T. Suehara and T. Tanabe, *LCFIPlus: A Framework for Jet Analysis in Linear Collider Studies*, Nucl. Instrum. Meth. **A808**, 109 (2016).

[107] LCFI Collaboration, D. Bailey *et al.*, *The LCFIVertex package: vertexing, flavour tagging and vertex charge reconstruction with an ILC vertex detector*, Nucl. Instrum. Meth. **A610**, 573 (2009).

[108] H. Aihara *et al.*, *SiD Letter of Intent*, (2009).

[109] G. Hanson *et al.*, *Evidence for Jet Structure in Hadron Production by e^+e^- Annihilation*, Phys. Rev. Lett. **35**, 1609 (1975).

[110] CLICdp, CLIC Collaboration, M. J. Boland *et al.*, *Updated baseline for a staged Compact Linear Collider*, (2016).