

# **Detectors and Physics at a Future Linear Collider**

Boruo Xu  
of King's College

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



## Abstract

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



## Declaration

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

Boruo Xu



## Acknowledgements

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



## **Preface**

This will be my preface. Where is Wolly?



# Contents

<b>1</b>	<b>Let's make introduction great again</b>	<b>1</b>
1.1	Future Linear Colliders . . . . .	1
1.2	Motivation . . . . .	1
<b>2</b>	<b>Theoretical overview</b>	<b>3</b>
2.1	Overview of the Standard Model . . . . .	3
2.2	Notations and conventions . . . . .	5
2.3	Quantum electrodynamics . . . . .	5
2.4	Quantum chromodynamics . . . . .	6
2.5	The electroweak interaction . . . . .	6
2.6	Higgs Mechanism . . . . .	8
2.7	Yukawa couplings . . . . .	8
2.8	Standard Model Higgs boson . . . . .	9
2.9	Higgs beyond the Standard Model . . . . .	10
2.10	Tau pair polarisation correlations as signature of Higgs boson . . . . .	14
<b>3</b>	<b>Detectors</b>	<b>17</b>
3.1	The ILC . . . . .	17
3.2	The CLIC . . . . .	18
3.3	Physics at future linear colliders . . . . .	18
3.4	Impact of physics requirements on the detector design . . . . .	19
3.4.1	Jet energy resolution requirements on the detector design . . . . .	19
3.4.2	Other requirements on the detector design . . . . .	21
3.5	The International Large Detector . . . . .	22
3.6	Overview of ILD sub-detectors . . . . .	23
3.6.1	Vertex Detector . . . . .	24
3.6.2	Tracking Detectors . . . . .	24
3.6.3	Electromagnetic Calorimeter . . . . .	24
3.6.4	Hadronic Calorimeter . . . . .	26

3.6.5	Solenoid, Yoke and Muon system . . . . .	27
3.6.6	Very Forward Calorimeters . . . . .	27
3.7	The CLIC versus the ILC . . . . .	28
3.7.1	The CLIC_ILD versus the ILD . . . . .	29
<b>4</b>	<b>Simulation, reconstruction and analysis software</b>	<b>33</b>
4.1	Monte Carlo event generation . . . . .	34
4.2	Event Simulation . . . . .	34
4.3	Event Reconstruction . . . . .	34
4.4	PandoraPFA event reconstruction . . . . .	35
4.4.1	Track selection . . . . .	35
4.4.2	Calorimeter selection . . . . .	36
4.4.3	Particle Identification . . . . .	36
4.4.4	Clustering . . . . .	37
4.4.5	Topological cluster association . . . . .	38
4.4.6	Track-cluster association . . . . .	38
4.4.7	Re-clustering . . . . .	39
4.4.8	Fragment removal . . . . .	40
4.4.9	Particle Flow Object Creation . . . . .	40
4.5	The CLIC specific simulation and reconstruction . . . . .	41
4.5.1	Luminosity spectrum . . . . .	41
4.5.2	Beam induced backgrounds . . . . .	41
4.5.3	CLIC simulated particle masses . . . . .	43
4.6	Analysis software . . . . .	43
4.6.1	Monte Carlo truth linker . . . . .	44
4.6.2	Jet algorithms . . . . .	44
4.6.3	Event shape variables . . . . .	46
4.7	Multivariate Analysis . . . . .	47
4.7.1	Optimisation and overfitting . . . . .	48
4.7.2	Choice of models . . . . .	49
4.7.3	Multiple classes . . . . .	55
<b>5</b>	<b>Photon Reconstruction in PandoraPFA</b>	<b>57</b>
5.1	Overview of photon reconstruction in PandoraPFA . . . . .	58
5.2	Electromagnetic shower . . . . .	58
5.3	PHOTON RECONSTRUCTION algorithm . . . . .	60
5.3.1	Form photon clusters . . . . .	61

5.3.2	Find photon candidates . . . . .	61
5.3.3	Photon ID test . . . . .	62
5.3.4	Photon Fragment removal . . . . .	62
5.4	Two dimensional peak finding algorithm for photon candidate . . . . .	63
5.4.1	Initialise 2D histogram . . . . .	64
5.4.2	Project hit energy to histogram . . . . .	64
5.4.3	Local peak identifying . . . . .	64
5.4.4	Associate non-peak bins to peaks . . . . .	65
5.4.5	Peak filtering . . . . .	65
5.4.6	Candidate close to track projection . . . . .	66
5.4.7	Inclusive mode . . . . .	67
5.5	Likelihood classifier for photon ID . . . . .	67
5.5.1	Projective Likelihood in PandoraPFA . . . . .	67
5.6	Photon fragment removal algorithm in the ECAL . . . . .	70
5.7	Photon fragment recovery algorithm in the HCAL . . . . .	74
5.8	Photon splitting algorithm . . . . .	76
5.9	Characterise the performance . . . . .	77
5.10	Compare with no photon reconstruction . . . . .	78
5.11	Compare with photon reconstruction in PandoraPFA version 1 . . . . .	80
5.12	Understand photon reconstruction improvement . . . . .	84
5.13	Current photon reconstruction performance . . . . .	85
<b>6</b>	<b>Tau Lepton Decay Modes Classification</b> . . . . .	<b>87</b>
6.1	Overview of the analysis . . . . .	88
6.2	Decay modes . . . . .	89
6.3	Simulation and reconstruction . . . . .	89
6.4	Event pre-selection . . . . .	90
6.5	Select single tau decay . . . . .	91
6.6	Discriminative variables . . . . .	91
6.6.1	$\rho(\pi^-\pi^0)$ and $\rho(\pi^-\pi^0)$ resonances reconstruction . . . . .	94
6.6.2	Separate $e^-$ from $\pi^-$ . . . . .	95
6.7	Multivariate Analysis . . . . .	96
6.8	Classification Efficiency . . . . .	96
6.9	Electromagnetic calorimeter optimisation . . . . .	98
6.9.1	Tau hadronic decay correct classification efficiency . . . . .	100

---

6.10 Separate H from Z with tau pair decay . . . . .	103
6.10.1 Event pre-selection . . . . .	103
6.10.2 Identify tau pairs . . . . .	103
6.10.3 Variables . . . . .	106
6.10.4 Multivariate analysis . . . . .	106
6.10.5 Result . . . . .	107
<b>7 Double Higgs Boson Production Analysis</b>	<b>109</b>
7.1 Analysis Straggly Overview . . . . .	110
7.2 Monte Carlo sample generation . . . . .	111
7.3 Lepton identification . . . . .	112
7.3.1 Electron and muon identification . . . . .	114
7.3.2 Tau lepton identification . . . . .	115
7.3.3 Very forward electron identification . . . . .	118
7.3.4 Lepton identification performance . . . . .	120
7.4 Jet reconstruction . . . . .	122
7.4.1 Jet reconstruction optimisation . . . . .	122
7.5 Jet flavour tagging . . . . .	127
7.5.1 Mutually exclusive cuts for $\text{HH} \rightarrow b\bar{b}W^+W^-$ and $\text{HH} \rightarrow b\bar{b}b\bar{b}$ . . . . .	128
7.6 Jet pairing . . . . .	130
7.7 Pre-selection . . . . .	131
7.8 MVA variables . . . . .	135
7.8.1 Invariant mass variables . . . . .	137
7.8.2 Energy and momentum variables . . . . .	137
7.8.3 Lab-frame angle variables . . . . .	138
7.8.4 Boosted-frame angle variables . . . . .	138
7.8.5 Event shape variables . . . . .	139
7.8.6 b and c tag variables . . . . .	140
7.8.7 PFOs number variables . . . . .	140
7.8.8 Cuts to aid the MVA . . . . .	142
7.9 Multivariate analysis . . . . .	142
7.10 Signal selection results . . . . .	142
7.11 $e^+e^- \rightarrow \text{HH}\nu_e\bar{\nu}_e \rightarrow b\bar{b}W^+W^-\nu\bar{\nu}$ hadronic decay at $\sqrt{s} = 3 \text{ TeV}$ analysis .	143
7.12 $e^+e^- \rightarrow \text{HH}\nu_e\bar{\nu}_e \rightarrow b\bar{b}W^+W^-\nu\bar{\nu}$ semi-leptonic decay at $\sqrt{s} = 3 \text{ TeV}$ analysis	151
7.13 Result interpretation . . . . .	152
7.14 Combined results . . . . .	155

7.15 Simultaneous couplings extraction . . . . .	155
<b>Bibliography</b>	<b>163</b>
<b>List of figures</b>	<b>169</b>
<b>List of tables</b>	<b>175</b>



*“Two bags of pork scratchings are worth  
a bag of gold.”*

— Joris the Dutch



# Chapter 1

## Let's make introduction great again

*“The journey of a thousand miles begins with a single step.”*

— Lao Zi, 604 BC - 531 BC

→ Introduction

### 1.1 Future Linear Colliders

Basic intro. LHC.

Next challenge

Future Options. FCC vs LC

LC options

### 1.2 Motivation

Photon - passage through matter. Photon electromagnetic shower

Since Higgs discovery in the LHC in 2012, Higgs

Ha there is a higgs.

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

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

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

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

# Chapter 2

## Theoretical overview

*“I believe it is impossible to be sure of anything.”*

— Han Fei Zi, 280 BC-233 BC

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

### 2.1 Overview of the Standard Model

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

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

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

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

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

The SM has enjoyed great success with theoretical predictions being experimentally verified. Some highlights included the discovery of the top quark in 1995 [?], the tau neutrino in 2000 [?] and the Higgs bosons in 2012 [1]. However, there are observations which are not explained by the SM. One issue is that the SM does not incorporate the gravitational force. There have been attempts to modify the SM but no conclusive theory exists yet. Another issue is that the SM does not allow neutrino masses and mixings. There have been many theories beyond the Standard Model (BSM). One such example is the generalisation of the Higgs theory to allow non-SM coupling strengths. This will be discussed in section 2.9.

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

## 2.2 Notations and conventions

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

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

## 2.3 Quantum electrodynamics

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

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

the Lagrangian is fixed to be

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

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

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

## 2.4 Quantum chromodynamics

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

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

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

$$G_{\mu\nu}^a = \partial_\mu \gamma_\nu^a - \partial_\nu \gamma_\mu^a - g_s f_{abc} G_\mu^b G_\nu^c. \quad (2.4)$$

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

## 2.5 The electroweak interaction

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

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

$$\mathcal{L}_{Electroweak} = \mathcal{L}_{Boson} + \mathcal{L}_{Fermion} \quad (2.5)$$

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

$$\mathcal{L}_{Boson} = -\frac{1}{4} W_{\mu\nu}^i W^{i\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu}, \quad (2.6)$$

where

$$W_{\mu\nu}^i = \partial_\nu W_\mu^i - \partial_\mu W_\nu^i - g\varepsilon^{ijk}W_\mu^j W_\nu^k, \quad (2.7)$$

$$B_{\mu\nu} = \partial_\nu B_\mu - \partial_\mu B_\nu. \quad (2.8)$$

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

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

$$\mathcal{L}_{Fermion} = \sum_{\psi \in fermions} \bar{\psi}_L \gamma^\mu D_\mu^L \psi_L + \bar{\psi}_R \gamma^\mu D_\mu^R \psi_R. \quad (2.9)$$

$D_\mu^L$  and  $D_\mu^R$  are defined as

$$D_\mu^L = \partial_\mu + ig\frac{\tau_i}{2}W_\mu^i + ig'Y_\psi B_\mu. \quad (2.10)$$

$$D_\mu^R = \partial_\mu + ig'Y_\psi B_\mu. \quad (2.11)$$

This Lagrangian allows  $W$  and  $B$  fields to couple with left-handed fermions. But only the  $B$  field couples to right-handed fermions. The  $\tau_i$  matrices are the generators of SU(2). (Pauli spin matrices are one of the representations.)  $Y_\psi$  is the hypercharge associated with the fermion field  $\psi$ .  $g'$  is the  $B$  field strength.

Two physical bosons,  $W^+$  and  $W^-$ , only couples to left-handed fermions. However  $Z$  and  $\gamma$  couples to both left-handed and right-handed fermions. Hence  $W^1$  and  $W^2$  are associated with  $W^+$  and  $W^-$ . Mass eigenstates for  $Z$  and  $\gamma$ ,  $Z_\mu$  and  $A_\mu$ , are mixtures of  $W_\mu^3$  and  $B_\mu$ .  $Z_\mu$  and  $A_\mu$  are defined as

$$Z_\mu = \cos(\theta_W) W_\mu^3 - \sin(\theta_W) B_\mu, \quad (2.12)$$

$$A_\mu = \sin(\theta_W) W_\mu^3 + \cos(\theta_W) B_\mu. \quad (2.13)$$

where  $\theta_W$  is the Weinberg mixing angle [7], which is determined experimentally. So far the  $SU(2) \otimes U(1)$  gauge theory explains the parity violating nature of the weak interaction. Explicit fermion mass is not allowed in the gauge symmetry. However, the higgs mechanism via spontaneous symmetry breaking would introduce mass terms for fermions.

## 2.6 Higgs Mechanism

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

$$\mathcal{L}_{Higgs} = (D_\mu \Phi_H)^\dagger (D^\mu \Phi_H) - \mu^2 \Phi_H^\dagger \Phi_H - \lambda (\Phi_H^\dagger \Phi_H)^2, \quad (2.14)$$

with

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

For negative  $\mu^2$ , the Higgs field potential

$$\mu^2 \Phi_H^\dagger \Phi_H + \lambda (\Phi_H^\dagger \Phi_H)^2 \quad (2.16)$$

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

$$\frac{(g\nu)^2}{4} W_\mu^+ W^{-\mu} + \frac{(g^2 + g'^2) \mu^2}{8} Z_\mu Z^\mu. \quad (2.17)$$

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

## 2.7 Yukawa couplings

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

$$\mathcal{L}_{Yukawa} = -\lambda^u \bar{q}_L \Phi_H^c u_R - -\lambda^d \bar{q}_L \Phi_H d_R - \lambda^e \bar{l}_L \Phi_H e_R + h.c. \quad (2.18)$$

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

$$m_u = \frac{\lambda^u \nu}{\sqrt{2}}, \quad m_d = \frac{\lambda^d \nu}{\sqrt{2}}, \quad m_e = \frac{\lambda^e \nu}{\sqrt{2}}. \quad (2.19)$$

## 2.8 Standard Model Higgs boson

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

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

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

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

$$\mathcal{L} \supset -\frac{m_i}{\nu} h \bar{\psi}_i \psi_i. \quad (2.21)$$

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

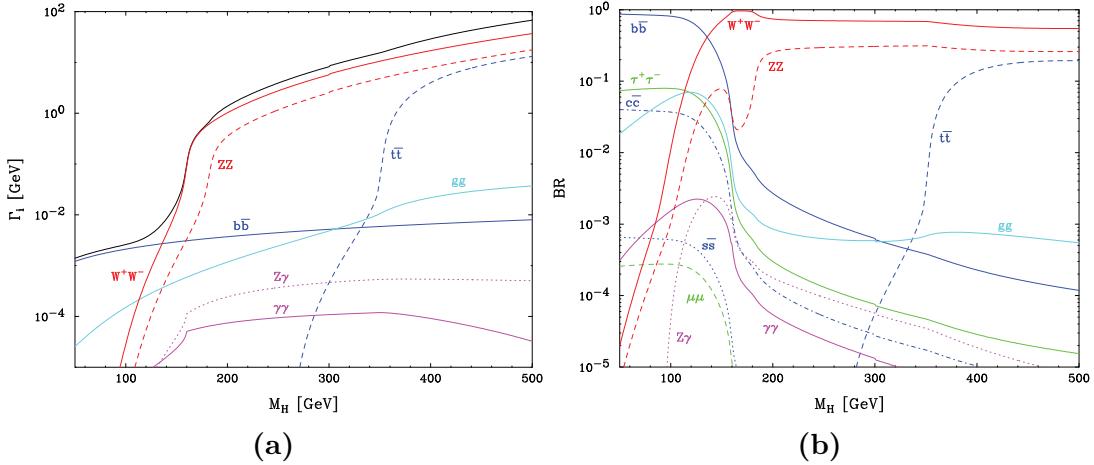
$$\mathcal{L} \supset m_W^2 \left( \frac{2h}{\nu} + \frac{h^2}{\nu^2} \right) W_\mu^+ W^{-\mu} + \frac{m_Z^2}{2} \left( \frac{2h}{\nu} + \frac{h^2}{\nu^2} \right) Z_\mu Z^\mu. \quad (2.22)$$

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

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

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

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



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

## 2.9 Higgs beyond the Standard Model

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

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

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

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

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

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

$$V(h) = \frac{1}{2}m_h^2 h^2 + d_3 \left(\frac{m_h^2}{2\nu}\right) h^3 + d_4 \left(\frac{m_h^2}{8\nu^2}\right) h^4 + \dots \quad (2.25)$$

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

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

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

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

where  $A_{SM}$  is the SM amplitude and

$$\delta_b \equiv 1 - \frac{b}{a^2}, \quad (2.27)$$

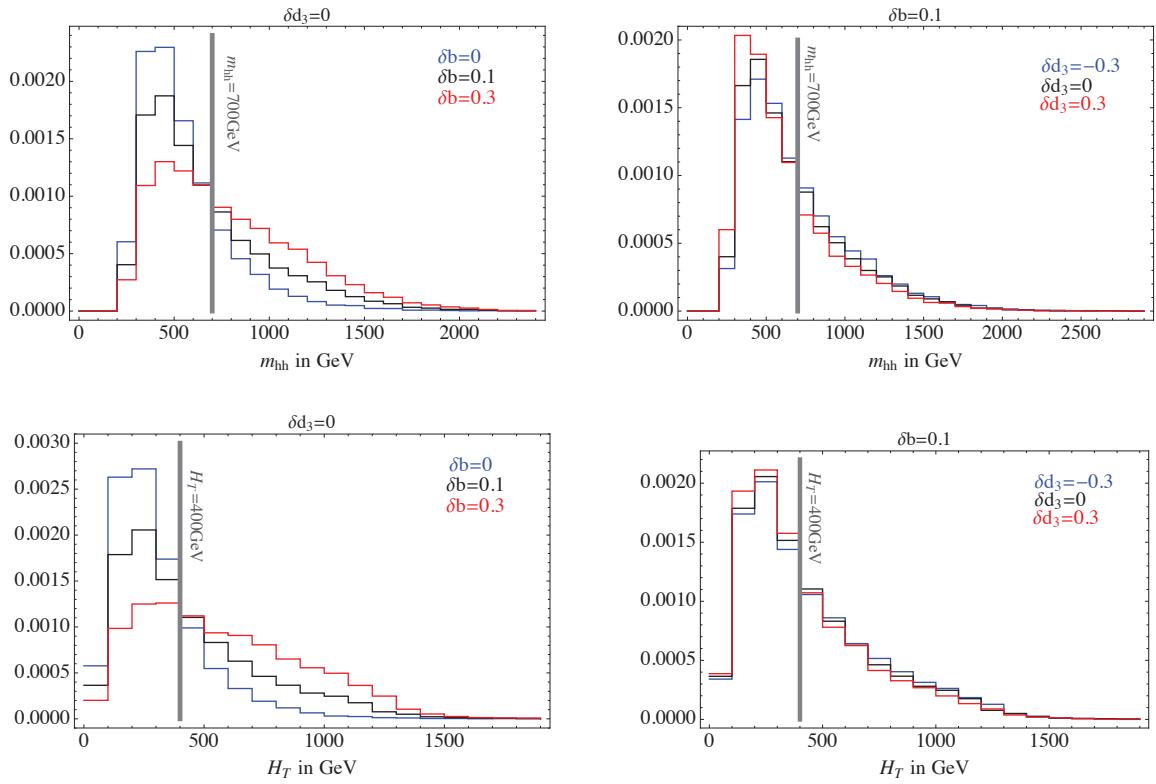
$$\delta_{d_3} \equiv 1 - \frac{d_3}{a}. \quad (2.28)$$

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

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

where  $\sigma_{SM}$  is the cross section predicted by the SM. With suitable kinematic cuts, high-energy behaviour can be disentailed from the physics at the threshold, allowing the extraction of  $\delta_b$ ,  $\delta_{d_3}$  and hence the coupling strengths,  $g_{VVH}$  and  $g_{HHH}$ . Suitable observables are variables that increase with increasing centre-of-mass energy. Two examples of such variables are the invariant mass of the two Higgs system,  $m_{hh}$ , and the sum of their transverse momenta,  $H_T$ . Figure 2.2 shows that the  $m_{hh}$  and  $H_T$  distributions are sensitive to the values of  $\delta_b$  and  $\delta_{d_3}$ . The figure shows the result of a general level study performed in [13].

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



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

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



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

## 2.10 Tau pair polarisation correlations as signature of Higgs boson

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

Tau pair polarisation correlations can be studied using various decay modes. Here reference [16] is followed and  $\tau^- \rightarrow \pi^- \nu_\tau$  decay mode is used as the example. The boson

decay can be represented as:

$$X \rightarrow \tau_\alpha^+ \tau_\beta^- \rightarrow \pi^+ \pi^- + \nu' s, \quad (2.30)$$

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

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

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

$$\frac{1}{\Gamma_{P\tau\mu}} \frac{d\Gamma}{dz} = Br_{\pi^-} f(\tau_\alpha^- \rightarrow \pi^-; z), \quad (2.32)$$

where  $Br_{\pi^-}$  is the branching fraction of  $\tau^- \rightarrow \pi^- \nu_\tau$ . The form  $f$  can be obtained from literature [17]:

$$f(\tau_\alpha^- \rightarrow \pi^-; z) = 1 + P_\alpha(2z - 1), \quad (2.33)$$

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

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

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

$$C_{LL}^H = C_{RR}^H = \frac{1}{2}, \quad (2.35)$$

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

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

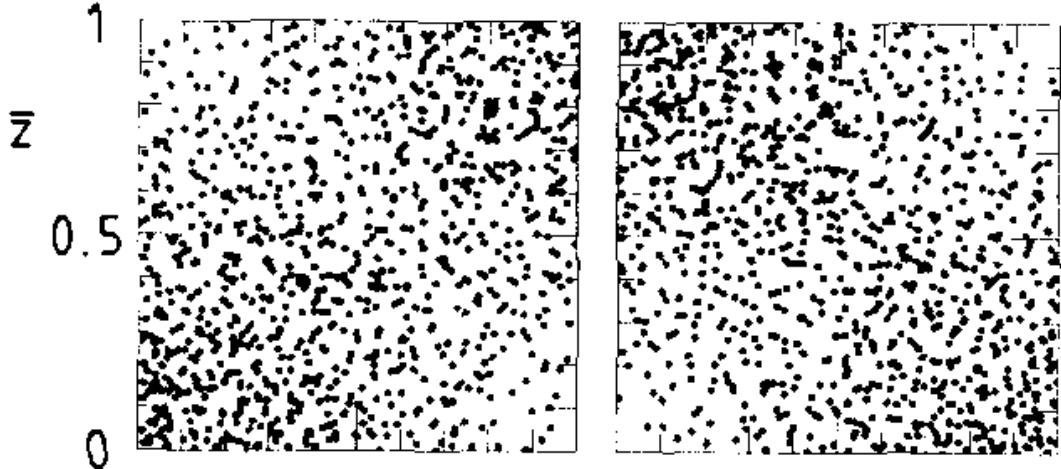
where tau polarisation,  $P_\tau$ , is due to the non-parity-conserving tau decay in the SM. For Z:

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

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

$$a = -\frac{1}{2}, \quad (2.39)$$

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



**Figure 2.4:** Two-dimensional distribution of  $Z \rightarrow \tau^+\tau^-$  on the left, and  $H \rightarrow \tau^+\tau^-$  on the right as a function of the energy fractions  $\bar{z} = E_{\pi^+}/E_{\tau^+}$ ,  $z = E_{\pi^-}/E_{\tau^-}$ . The figure is adapted from reference [17].

# Chapter 3

## Detectors

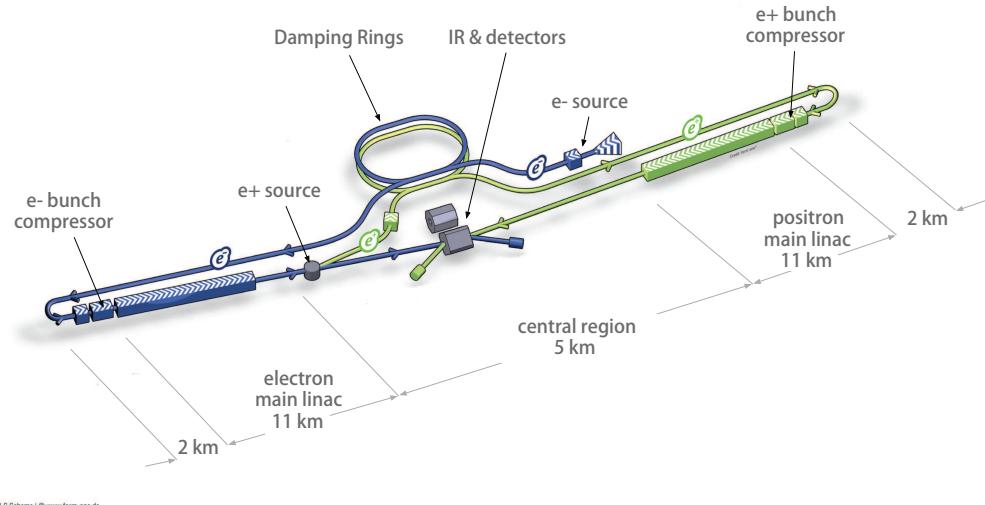
*“The great man is the one who does not lose his childqŕs heart.”*

— Mencius, 372 BCÍC 289 BC

Since the discovery of a particle consistent with the SM Higgs boson at the LHC in 2012 [1, 18], our understanding of the Standard Model has improved greatly. Yet limited by the underlying QCD interaction from proton-anti-proton collision, one has great difficulties in measuring the properties of the Higgs precisely. Next generation electron-positron linear colliders could hopefully make precise measurements of the Higgs sector and the Top quark sector [19, 20].

### 3.1 The ILC

Two leading candidates for next generation electron-positron linear colliders are the International Linear Collider (ILC) [19], and the Compact Linear Collider (CLIC) [20]. The ILC is a high-luminosity electron-positron linear collider with centre-of-mass energies from 200 GeV up to 1 TeV. The machine would be built at different stages. The first stage would have a centre-of-mass energy of 250/350 GeV. The second stage would have a centre-of-mass energy of 500 GeV with a possible upgrade to 1 TeV. Thirty years of development leads to the technical design report in 2013 [21]. A layout of the collider complex is shown in figure 3.1. The proposal contains two detector concepts, the International Large Detector (ILD) and the Silicon Detector (SiD).



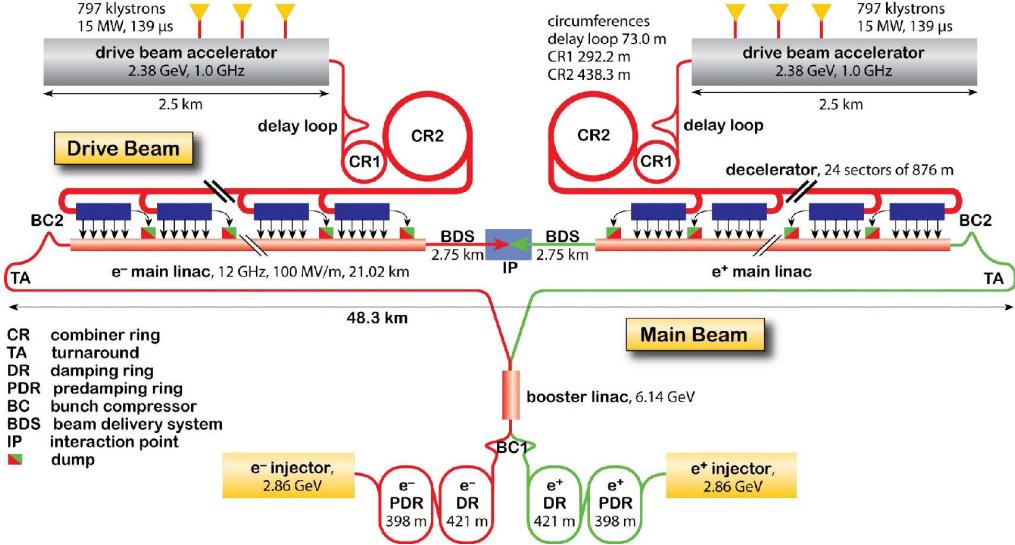
**Figure 3.1:** A layout of the International Linear Collider complex, taken from [21].

## 3.2 The CLIC

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

## 3.3 Physics at future linear colliders

The physics program for the CLIC and the ILC, which is a driving force for the detector design, have some common goals. ILC has a reach of centre-of-mass from 200 GeV to 1 TeV, whilst CLIC can reach from 350 GeV to 3 TeV. Both machines are capable of precision higgs coupling measurements, top mass and coupling measurements, and search for new physics such as supersymmetry particles. The ILC can also operate at low energy to be a Z and a H factory for ultra precision Z mass and H mass measurements.



**Figure 3.2:** A layout of the Compact Linear Collider at final stage of a centre-of-mass of 3 TeV, taken from [22].

CLIC, however, has the advantage of higher energy reach, which allows measurements of rare events, such as higgs triple self-couplings and quartic couplings. Also the discovery potential for the CLIC is greater.

## 3.4 Impact of physics requirements on the detector design

### 3.4.1 Jet energy resolution requirements on the detector design

The physics goal of jet energy resolution at the ILC and the CLIC is to separate W and Z hadronic decays by reconstructing their quark-jet masses [20, 21]. This translates to a requirement of 3.5-5% of the energy resolution. This level of precision is unlikely to be achieved with a traditional calorimetry design. A traditional energy flow approach to calorimetry measures jet energies as a sum of the energy in the calorimeters. The jet energy resolution is parameterised by:

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

The stochastic term  $a$  is typically greater than 60% and  $b$  is of the order of a few percents. For the jet energy resolution of 3.5%,  $a$  should be less than 30% which is unlikely to be achieved by a traditional calorimeter. On the contrary, the particle flow approach to calorimetry has demonstrated its ability to reach the goal [23, 24].

In a typical jet, using measurements on the particle composition from the LEP [25, 26], about 62% of the jet energy is from charged particles, 27% from photons, 10% from long-lived neutral hadrons, and 1.5% from neutrinos. In a traditional approach to calorimetry, about 72% jet energy is measured in the electromagnetic (ECAL) and the hadronic (HCAL) calorimeters combined. The jet energy resolution is thus limited by the energy resolution of the hadronic calorimeters, which typically is  $\gtrsim 55\%/\sqrt{E(\text{GeV})}$ .

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

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

Figure 3.3 shows a typical topology of a 250 GeV jet, simulated with the CLIC\\_ILD detector concept. Particles consisting of the calorimeter hits and tracks are labelled with different colours. Clusters of calorimeter hits in the highly granular ECAL and



**Figure 3.3:** A typical topology of a 250 GeV jet, simulated with the CLIC\_ILD detector concept, taken from [24].

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

### 3.4.2 Other requirements on the detector design

Other physics requirements for the detectors for the ILC and the CLIC are summarised in [20, 21]. Here important requirements are discussed as motivations for detector designs.

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

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

impact parameter resolution can be written in the form of:

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

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

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

### 3.5 The International Large Detector

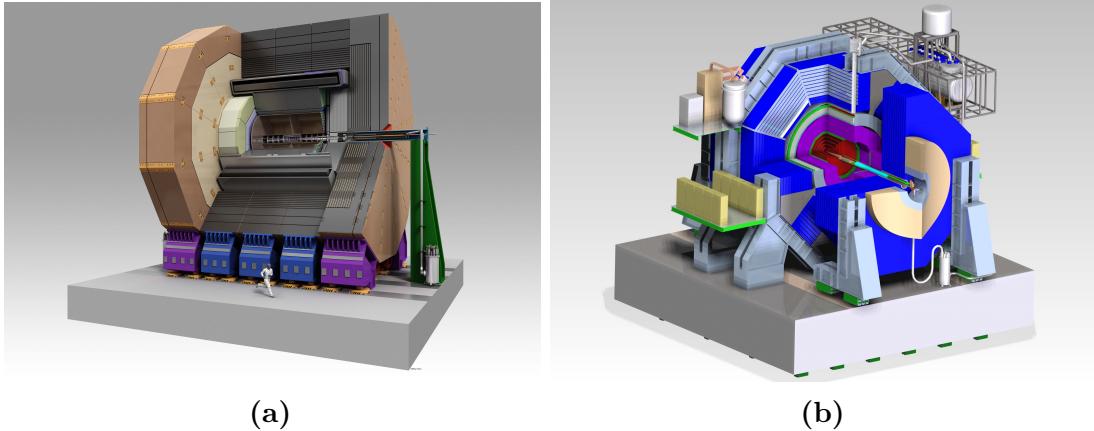
Two detector concepts have been designed for the ILC to deliver its physics program. The motivation for two detectors is to have multiple independent measurements within one collider for cross-checking, complementary measurements and competition between collaborations. The two detectors are both general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

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

The particle flow paradigm requires topological information of individual particle reconstructions. The sub-detector systems need to have the spatial resolution to separate charged particles from neutral particles. The result is a highly granular calorimeter and a central tracking system with excellent momentum resolution. Longitudinal cross section of top quadrant of the ILD detector concept, taken from [19], is shown in figure

**3.10a.** From the interaction point (IP) outwards, there is a tracking system comprising a large time projection chamber (TPC) augmented with silicon tungsten layers, highly granular electromagnetic calorimeters (ECAL) and hadronic calorimeters (HCAL), muon chambers, forward calorimeters (FCAL), magnetic coils and iron yokes.

The section below will describe the sub-systems of the ILD detector concept in the ILD technical design report [21], the ILD\_o1\_v05 option in MOKKA simulation. This detector concept has been used in studies described in subsequent chapters. The ILD detector concept has been optimised and documented in previous documents, such as the letter of intent [57]. The CLIC\_ILD detector concept for the CLIC in the conceptual design report [20] is a modified version of the ILD, adapted to the CLIC colliding environment. As they share similarities, only the differences are addressed. A comparison between ILD and CLIC\_ILD can be seen in figure 3.10 and table 3.1.



**Figure 3.4:** Figure 3.4a and figure 3.4b show the International Large Detector and the Silicon Detector for the International Linear Collider, respectively, taken from [21].

## 3.6 Overview of ILD sub-detectors

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

### 3.6.1 Vertex Detector

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

### 3.6.2 Tracking Detectors

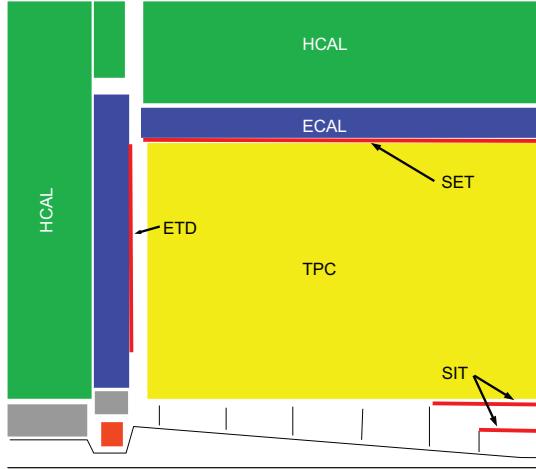
The hybrid tracking system consists of a large time projection chamber (TPC), a Silicon Inner Tracker (SIT), a Silicon External Tracker (SET) in the barrel region, a end cap tracking component (ETD) behind the endplate of the TPC, and a silicon forward tracker (ETD) in the forward region. The SIT, SET,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

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

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

### 3.6.3 Electromagnetic Calorimeter

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



**Figure 3.5:** A top quadrant view of the ILD silicon envelope system, SIT, SET, ETD, and ETD, with TPC, ECAL, and HCAL, adapted from the figure in [21].

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

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

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



**Figure 3.6:** A cross section through the electromagnetic calorimeter layers, taken from [21].

### 3.6.4 Hadronic Calorimeter

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

The principal role of the HCAL is to separate neutral hadron showers from other particles, and to measure neutral hadron energies. The neutral hadron contribution of the jet energy is around 10% on average. A moderate fine granular HCAL is a good balance between cost and performance. The chosen layout is 48 longitudinal layers with 3 by 3 cm scintillator tiles, using an analogue read out system. The layout of a technological prototype, the "EUDET prototype" [27] is shown in figure 3.7a.

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

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



**Figure 3.7:** Figure 3.7a shows the schematic view of a CALICE AHCAL technological prototype module. Figure 3.7b shows the jet energy resolution as a function of the hadronic calorimeter scintillator cell sizes, with different energies. Both figures are taken from [21].

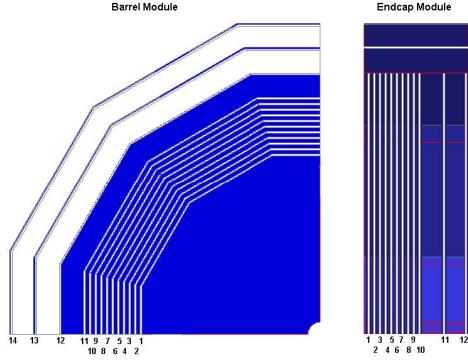
### 3.6.5 Solenoid, Yoke and Muon system

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

An iron yoke instrumented with scintillator strips as active layers returns the magnetic flux. The yoke also acts as a muon detector and tail catcher calorimeter at the same time. The layout is shown in figure 3.8. The agreed maximum magnetic field at 15 m radial distance from the detector is 50 Gauss to ensure safety [28]. A highly efficient muon detector is provided by the 3 by 3 cm scintillator strips. As a tail catcher calorimetry, the first layer of the muon detector, catches the energy leakage from the HCAL and the ECAL. It has been shown that a 10% improvement of single particle energy resolution is possible with the tail catcher [29].

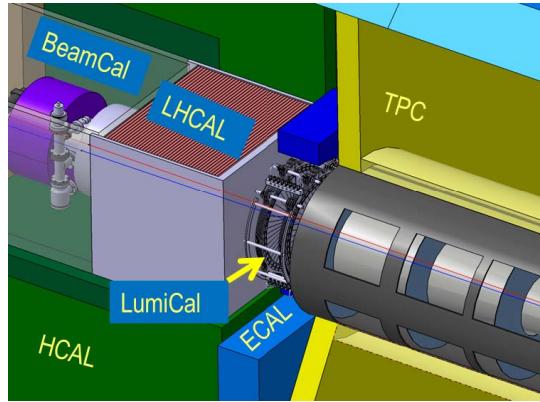
### 3.6.6 Very Forward Calorimeters

The forward region detectors provide luminosity measurements and forward coverage of calorimeters. A system of precision and radiation resistant calorimeters are required. The luminosity calorimeter counts Bhabha scattering to measure the luminosity to precision



**Figure 3.8:** Sensitive layers of the ILD muon system, taken from [21].

of  $10^{-3}$  at 500 GeV centre-of-mass energy. The beam calorimeter (BeamCAL) extend the forward coverage, which are hit by many beamstrahlung pairs after each bunch crossing. BeamCAL estimates a bunch-by-bunch luminosity. An additional hadron calorimeter, LHCAL, at the forward region extends the angular coverage of the HCAL to that of the LumiCAL. Electron tagging is possible with the very forward calorimeters [30], which aids event reconstruction at high centre-of-mass energy.



**Figure 3.9:** The forward calorimeters of the ILD, taken from [21]. LumiCAL, BeamCAL, and LHCAL are the luminosity calorimeter, beam calorimeter, and forward hadronic calorimeter, respectively.

### 3.7 The CLIC versus the ILC

The two main differences between the CLIC and the ILC are the high centre-of-mass energy and the high bunch charge density leading to significant beam related backgrounds. Within a bunch train, there is 0.5 ns between bunch crossings at the CLIC. There are

two main sources of beam induced background: incoherent electron pairs from photon (real or virtual) interactions with individual particles of the other beam, and interactions of two photons from the colliding beams. These differences leads to a modification in the detector design and the reconstruction software for the CLIC.

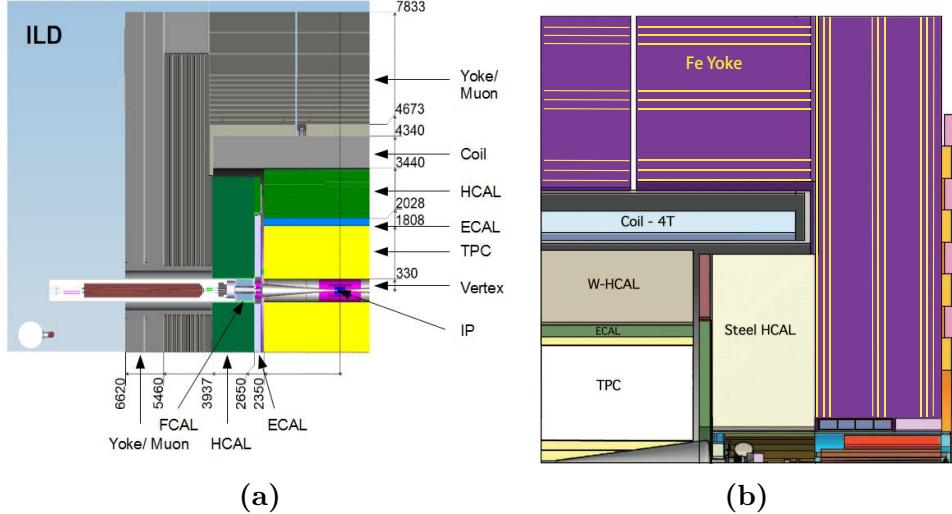
### 3.7.1 The CLIC\_ILD versus the ILD

There are two detector concepts studied in the CLIC conceptual design report [20], the CLIC\_ILD and the CLIC\_SiD. The CLIC\_ILD detector concept is based on the ILD design. They share similarities due to similar physics motivations. Only the differences are highlighted here. A comparison of the CLIC\_ILD and the ILD longitudinal cross sections can be seen in figure 3.10. A comparison of key parameters of the ILD and the CLIC\_ILD detector concepts is shown in table 3.1.

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

**Table 3.1:** A comparison of key parameters of the ILD and CLIC\_ILD detector concepts. ECAL  $r_{min}$  is the smallest distance from the calorimeter to the main detector axis. HCAL Absorber B / E indicates the absorber material for the barrel (B) and the endcap (E). The table is adapted from [20].

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



**Figure 3.10:** Figure 3.10a and Figure 3.10a shows longitudinal cross section of top quadrant of the ILD and the CLIC\_ILD detector concepts, taken from [21] and [20] respectively. From interaction point (IP) outwards, there is a tracking system comprising a large time projection chamber (TPC) augmented with silicon tungsten layer, highly granular electromagnetic calorimeters (ECAL) and hadronic calorimeters (HCAL), muon chambers, forward calorimeters (FCAL), magnetic coils and iron yokes. The numbers are in units of mm.

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

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

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

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

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

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

**Table 3.2:** Comparison of the LumiCAL and the BeamCAL at the ILD and the CLIC\_ILD. The table is adapted from [20].



# Chapter 4

## Simulation, reconstruction and analysis software

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

— William Shakespeare, 1564 - 1616

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

In previous chapters, overviews of the theory and the future linear collider experiments have been described. In this chapter, the simulation and event reconstruction chain are discussed, followed by discussion on common analyses software to be used in subsequent chapters. Simulation and reconstruction of events of the future Linear Colliders, ILC and CLIC, share a common software framework. Therefore, the shared reconstruction is discussed first, and the CLIC specific issues are highlighted afterwards. The event reconstruction is focused on the PandoraPFA event reconstruction, which is the framework for the photon reconstruction algorithms in chapter 4. Lastly analysis software and

multivariate analysis are presented. The multivariate analysis is discussed in a separate section because of its complexity.

## 4.1 Monte Carlo event generation

Monte Carlo (MC) event generation is the first step for the simulated study. Most events, electron-positron interaction, are generated with the WHIZARD software [31, 32], with no polarisation of the electron and positrons. Some simple events are generated manually with HEPEVT format. The PYTHIA [33] is used to describe parton showering, hadronisation and fragmentation. The parameters for the PYTHIA are tuned to OPAL data from the LEP [34]. The TAUOLA [35] debrides the tau lepton decay with correct spin correlations of the day products. The Initial State Radiation (ISR) is simulated in WHIZARD, with the ISR photons being collinear with the beam direction. The Final State Radiation (FSR) is simulated with default parameters in PYTHIA.

## 4.2 Event Simulation

For all the simulated events used in this thesis, the simulation software used is the GEANT4 [36]. The detector geometry description is provided by the MOKKA [37]. The QGSP\_BERT physics list is used to describe the hadronic shower decay in the detector.

## 4.3 Event Reconstruction

With simulated events (or real data in the future) as inputs, the next step is to reconstruct these events. The Reconstruction software runs in the Marlin framework [38], as a part of the iLCSoft. The event reconstruction contains following steps: digitisation of simulated calorimeter hits, reconstruction of tracks in the tracking system (using pattern recognition algorithms), and particle flow objects (PFOs) reconstruction with PandoraPFA [23, 24]. Details of the reconstruction can be found in [19, 20]. Here particle flow reconstruction via PandoraPFA will be discussed in details, as it provides the software framework for the photon reconstructions in PandoraPFA in chapter 5 and used in chapter 6 and chapter 7.

## 4.4 PandoraPFA event reconstruction

The tradition energy flow approach to calorimetry is unable to meet the mass and energy resolution requirements for future linear colliders. The particle flow approach with PandoraPFA has a proof-of-principle demonstration of its capability to reach the required resolution. The particle flow approach also put stringent requirements on the detector design, which is described in section 3.4.1. By associating calorimeter hits to the tracks, around 60% of the jet energy from charged particles is measured by the tracker, which has a much better resolution than the calorimeter. Small cell sizes of the calorimeters are required to identify hits from different particles. The traditional sum of calorimeter cell energies is replaced by particle flow reconstruction algorithms - a complex pattern recognition problem. The PandoraPFA algorithm has been developed and used in the ILC and CLIC simulation studies. There are over 60 linear collider specific reconstruction algorithms. Each aims to address a particular topological issue in the reconstruction.

Developed with the ILD detector concept, PandoraPFA has been adapted to the CLIC condition and shows its ability to deliver required energy resolutions [20]. In the recent development, the core base codes for basic object and memory managements are factorised in the Pandora C++ Software Development Kit [39]. In the subsequent sections, the main steps in the PandoraPFA reconstructions are described. The details of the reconstruction can be found in [23, 24, 39]. The inputs of PandoraPFA are digitised calorimeter hits and reconstructed tracks, with some detector geometry information to aid the reconstruction. The output are reconstructed particles with four-momenta, also known as the Particle Flow Objects (PFOs).

### 4.4.1 Track selection

Tracks from the inner tracking detectors are important inputs of the PandoraPFA reconstruction. These tracks are selected based on their topological properties, how likely they are from physical processes, and whether they are consistent with the tracker resolution. Only tracks passing the selection are used for the subsequent reconstruction.

Special topologies of tracks are identified, such as when a neutral particle decays or converts into a pair of charged tracks, leaving tracks of a “V0” shape. This is identified by searching for a pair of tracks originated from a single point. Another topology is the “kinks” when a charged particle decays to a single charged particles with neutral

particles. The last special topology are the “prongs” when a charged particle decays to multiple charged particles. This information is stored and passed on to the subsequent reconstruction, along side with helical track fit (using last 50 reconstructed tracker hits) and the track projection to the front of the ECAL.

#### 4.4.2 Calorimeter selection

The other important inputs of the PandoraPFA reconstruction are the calorimeter hits from calorimeters. The properties of a calorimeter hit include the position, its layer in the calorimeter, and its energy response from the calorimeter digitiser.

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

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

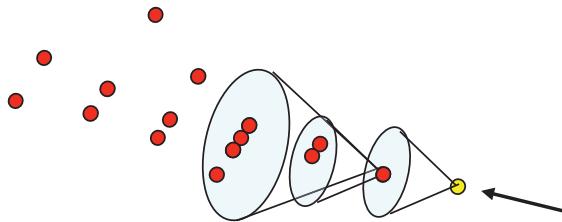
Isolated hits, often originating from low energy neutrons in a hadronic shower, are difficult to associate to the correct hadronic shower. They are identified and not used during the clustering stage. However, these isolated hits participate in the reconstruction during the last step, particle flow object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

#### 4.4.3 Particle Identification

To improve the reconstruction of the charged particles, where calorimeter hits are associated to tracks, dedicated particle identification algorithms identify muons and photons and remove identified particles from the reconstruction. By removing the hits from muons and photons, the reconstruction of charged particles is improved as fewer hits are left to be reconstructed. Hence the pattern recognition problem is reduced. Identified muons and photons do not participate in the later clustering and re-clustering stages, but re-enter the reconstruction at the fragment removal stage (see section 4.4.8). The details of the photon reconstruction algorithms and photon related algorithms are described in chapter 5.

#### 4.4.4 Clustering

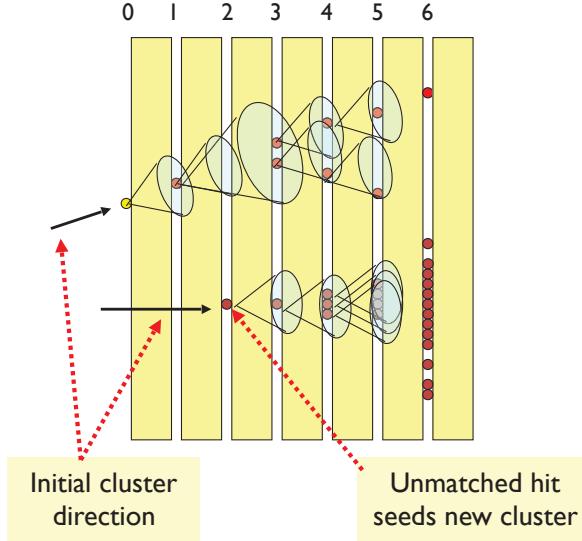
A cone clustering algorithm is used to group calorimeter hits from innermost to outmost psuedo-layer. The output custers are further processed, merged or split based on their topological properties. Since the cone clustering algorithm widely used in many reconstruction algorithms in the PandoraPFA, it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



**Figure 4.1:** Illustration of the cone based clustering, taken from [40]

In general there are two main types of clustering algorithms: cone based algorithms and sequential combination algorithms(see section 4.6.2). The main clustering scheme used in the PandoraPFA is the cone based algorithms. The clustering algorithm forms clusters to group calorimeter hits. Illustrated in figure 4.1, cone clustering algorithm identifies a seed first, shown as the yellow dot. The algorithm then forms a cone to include hits that are within a specified opening angle to the seed hit. The cone based clustering algorithm is preferred because the direction of particle flow is largely unchanged from the originated particle, whether the particle flow is an electromagnetic shower, QCD radiation, or hadronisation.

Figure 4.2 shows the clustering algorithm used in the PandoraPFA. The seed for the cone clustering is typically the projection of a energetic track to the front of the ECAL. A high energy calorimeter hit can also be used as a seed. The initial cluster direction is taken as the direction of the seed. Afterwards, a cone with a specified opening angle and depth will be formed around the seed. The four-momentum of calorimeter hits are therefore summed to obtain the cone's four-momentum. The building of the cone is iterated from the inner layer of the ECAL to the outer layer. At each layer, possible associations with hits in previous layers and the same layer are checked. If a hit is not associated with the cone, the hit is used to seed a new cluster. The clustering algorithm produces basic working objects, custers.



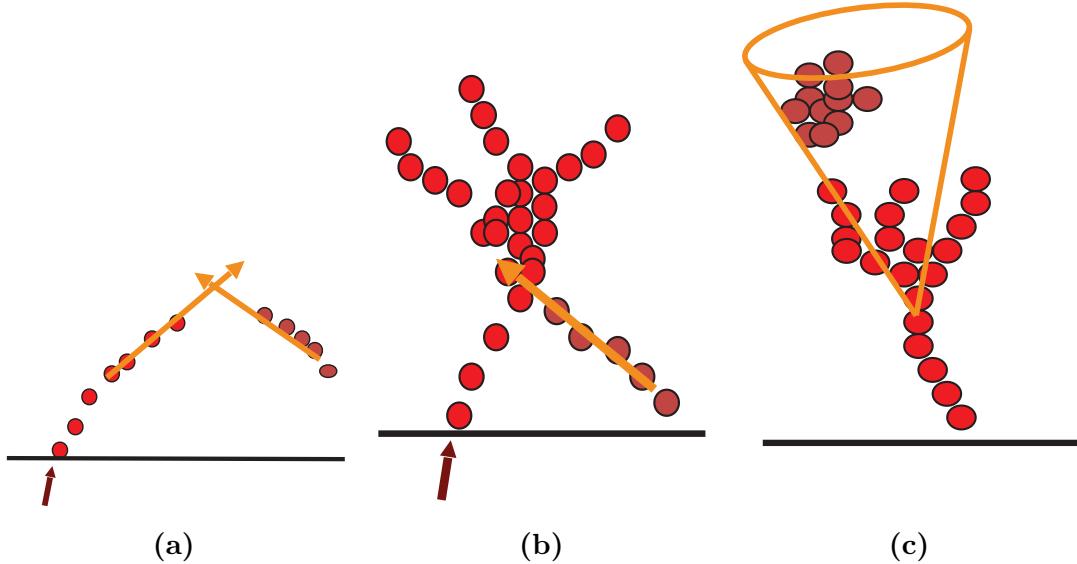
**Figure 4.2:** Illustration of the clustering algorithm used in the PandoraPFA, taken from [40]

#### 4.4.5 Topological cluster association

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

#### 4.4.6 Track-cluster association

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



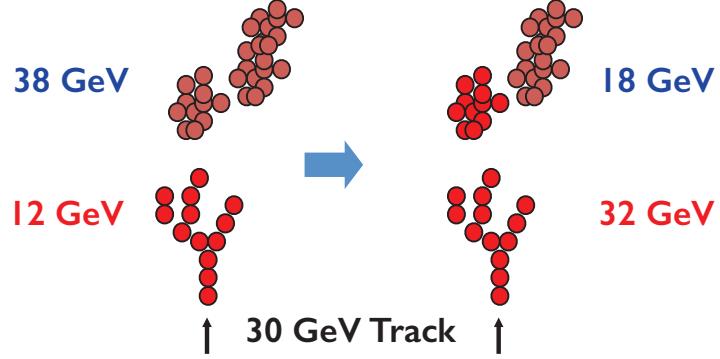
**Figure 4.3:** Examples of topological association in the PandoraPFA. Figure 4.3a, figure 4.3b, and figure 4.3c show rules for looping track segments, back-scattered tracks from hadronic showers, and cone association, respectively. In each case, the arrow indicates the tracks. The dark red dots represent the calorimeter hits in the associated cluster. The slightly fainter red dots represent the calorimeter hits in the neutral cluster. The black line represents the front of the ECAL. The figures are taken from [40].

#### 4.4.7 Re-clustering

The cluster association scheme described in the previous section works well for low energy (less than 50 GeV) jet. For a high energy jet, particles and the subsequent hadronic showers are more boosted and more likely to overlap each other. Therefore, it is important to re-cluster based on the compatibility of the cluster energy and the associated track momentum.

The re-clustering is performed on a statistical basis. If the cluster energy and the associated track momentum do not match, the cluster will be re-clustered using the same clustering algorithm with different parameters, or different cluster algorithms. Afterwards, out of many clusters obtained with the re-clustering step, the cluster that matches the best with the track momentum is chosen, and the cluster will be associated with that track. A schematic diagram of the re-clustering stage is shown in figure 4.4. In the figure, the black upright arrows indicate the tracks. The dark red dots represent the calorimeter hits in the associated cluster. The slightly fainter red dots represent the calorimeter hits in the neutral cluster. In this example, the cluster energy is less than the associated track momentum. The topological association algorithms did not add the natural cluster,

as it would have formed a cluster with too much energy. The re-clustering scheme tries different cone clustering by splitting the neutral cluster so that the topological association could make correct association.



**Figure 4.4:** Illustration of the re-clustering algorithm in PandoraPFA, taken from [40]. The arrow indicates the tracks. The dark red dots represent the calorimeter hits in the associated cluster. The slightly fainter red dots represent the calorimeter hits in the neutral cluster. The cluster energy is less than the associated track momentum. The topological association algorithms did not add the natural cluster, as it would have formed a cluster with too much energy. The re-clustering algorithm tries different cone clustering to split the neutral cluster so that the topological association could make correct association.

#### 4.4.8 Fragment removal

This stage of the reconstruction will focus on merging low energy clusters. These neutral clusters are likely to be fragments of other clusters. The merging criterion are mostly based on the proximity and the energy comparison. Algorithms dealing with photon fragment merging and photon splitting are described in details in chapter 5.

#### 4.4.9 Particle Flow Object Creation

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

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

## 4.5 The CLIC specific simulation and reconstruction

There are a few simulation and reconstruction issues specific to the CLIC, which affect the analysis in chapter 7. One issue is that the luminosity spectrum for interactions with photon from Beamstrahlung is different to electron-positron interactions. Therefore a solution is presented in section 4.5.1 to correct for the differences. Another issue is that there is a large amount of beam induced background in CLIC, which needs to be suppressed before physics analyses. The background suppression is described in section 4.5.2. Lastly simulated masses of particles are given in table 4.2, which will be used in section 7.4.1.

### 4.5.1 Luminosity spectrum

The electron-photon interaction, where the photon is produced from initial state radiation via Beamstrahlung , has a different instantaneous luminosity than the electron-positron interaction. Hence, for the same time-frame, the total integrated luminosity of the electron-photon interaction is different to that of the electron-positron interaction. To correct for the difference in the luminosity, a simulated study [41] was performed with the GUINEAFIG [42] and was simulated in the WHIZARD, to identify the ratio of the integrated luminosity of the electron-photon interaction to the electron-positron interaction. The results are summarised in table 4.1. For the physics analysis in chapter 7, event number for processes with initial-state photons from Beamstrahlung are corrected with the ratios in table 4.1.

### 4.5.2 Beam induced backgrounds

The other issue considered when using the CLIC\_ILD detector concept is the beam induced background. At high  $\sqrt{s}$ , the background becomes important for the event reconstruction. Therefore, the beam induced background are considered in the simulation.

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

**Table 4.1:** Luminosity ratio for processes with initial-state photons from Beamstrahlung at the CLIC, at  $\sqrt{s} = 1.4 \text{ TeV}$  and  $3 \text{ TeV}$ . The table summarises results from [41].

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

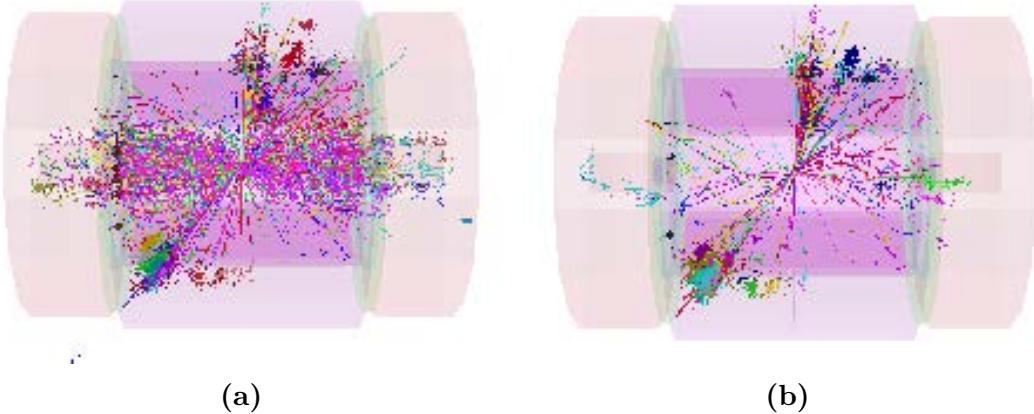
These  $\gamma\gamma \rightarrow \text{hadrons}$  background events are hadronised with the PYTHIA, and superimposed on the physics process simulations to save computational resources. The choice of 60 bunch crossings is a conservative estimate of the amount of background in the experiment condition [43, 43].

The beam induced background deposits significant amounts of energies in the detector. It needs to be suppressed for physics analyses, like the one in chapter 7. Two software have been developed to suppress these background: a track selector and a PFO selector [24].

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

The PFO selector discards PFOs that are originated from the beam induced background, based on the transverse momentum ( $p_T$ ) and time information of the PFOs. The PFOs from  $\gamma\gamma \rightarrow \text{hadrons}$  often have low  $p_T$  and have a range of time-of-arrivals. In contrast, the PFOs from physics processes have a range of  $p_T$ , and the time-of-arrivals are close to the bunch crossing time. By utilising the high spatial resolution from the high granular calorimeter, individual PFOs can be tracked and reconstructed. The PFO selector uses different  $p_T$  and time cuts for the central part of the detector and for the forward part of the detector. There are different cuts for photons, neutral PFOs, and charged PFOs. Three configurations of these cuts are developed, namely “loose”, “normal”, and “tight” selections. As the name suggested, “loose” selection corresponds

to a looser cut of  $p_T$  and time-of-arrival, allows a larger value of  $p_T$  and time-of-arrival. The optimal configuration depends on the  $\sqrt{s}$  of the collision, and the physics process to study. Figure 4.5 shows the effect of the suppression of the background with the tight PFO selection. Reconstructed particles for a time window of 10 ns (100 ns in HCAL barrel) in a simulated  $e^+e^- \rightarrow HH \rightarrow t\bar{b}t\bar{b}$  event in the CLIC\_ILD detector, with 60 bunch crossings of  $\gamma\gamma \rightarrow$  hadrons background overlaid in figure 4.5a. The effect of applying tight PFO section cuts is shown in figure 4.5b.



**Figure 4.5:** Reconstructed particles for a time window of 10 ns (100 ns in HCAL barrel) in a simulated  $e^+e^- \rightarrow HH \rightarrow t\bar{b}t\bar{b}$  event in the CLIC\_ILD detector, with 60 bunch crossings of  $\gamma\gamma \rightarrow$  hadrons background overlaid in figure 4.5a. The effect of applying tight PFO section cuts is shown in figure 4.5b. The figures are taken from [24].

### 4.5.3 CLIC simulated particle masses

Another information used in the analysis (section 7.4.1) with the CLIC detector concept is the simulated mass and width of quarks and bosons. These values are listed in table 4.2 and are used for generating Standard Model samples.

## 4.6 Analysis software

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

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

**Table 4.2:** The masses and widths of quarks and bosons used for generating Standard Model samples. The H mass is specified for individual samples. The table is taken from [20].

#### 4.6.1 Monte Carlo truth linker

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

#### 4.6.2 Jet algorithms

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

A jet is typically a visually obvious structure in an event display. The momentum and the direction of a jet tend to resemble the original particle. Despite the relative simplicity of identifying jets visually, it is a challenge for a pattern recognition program to identify jets effectively and efficiently. Early work on jet finding started in 1977 [44], where later development can be found in reviews [45–47]. This section is based on these reviews.

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

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

The chosen jet algorithm implementation for this thesis is the FastJet C++ software package [48, 49], providing a wide range of jet finding algorithms. The implementation in the Marlin software package is called the MarlinFastJet. The symbols in the subsequent discussion follow the convention in [48].

## $k_t$ algorithm

Longitudinally-invariant  $k_t$  algorithm [50, 51] is one of the common sequential combination algorithms for  $p\bar{p}$  collider experiments. There are two variants: 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

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

$$d_{iB} = p_{Ti}^2, \quad (4.2)$$

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

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

The exclusive variant is similar to the inclusive variant. First difference is that when  $d_{iB} < d_{ij}$ , particle  $i$  is discarded and part of the beam jet. The beam jet contains particles that are considered from the beam induced background. The beam jet is not used as a

output. The second difference is that when both  $d_{ij}$  and  $d_{iB}$  are above some threshold,  $d_{cut}$ , the clustering will stop. In practise, the exclusive mode allows a specified number of jets to be found, which will automatically choose the  $d_{cut}$ . The inclusive mode would find as many jets as the algorithm allows. An example usage of the exclusive  $k_t$  algorithm can be found in section 7.4.1.

## Durham algorithm

The Durham algorithm [52], also known as  $e^+e^- k_t$  algorithm, is commonly used for the  $e^+e^-$  collider experiments. It has a single distance metric:

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

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

Compared to the  $k_t$  algorithm, it uses energy instead of  $p_T$  in the distance metric, and it does not have a beam jet. This is because that for the  $e^+e^-$  collider at low centre-of-mass energies, the beam induced background is not severe and the collisions energies are known.

## Jet algorithm for the CLIC

Although CLIC is a  $e^+e^-$  collider, the significant beam-induced background adds a large amount of energy. Therefore, traditional  $e^+e^-$  jet algorithms, like the Durham algorithm, is not suitable for the CLIC collision environment. Studies have shown that jet algorithms for the  $p\bar{p}$  colliders give better performances for the CLIC [20, 53].

### 4.6.3 Event shape variables

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

The classical event shape thrust [60], is defined as

$$T = \max_{\hat{t}} \frac{\sum_i |\hat{t} \cdot \vec{p}_i|}{\sum_i |\vec{p}_i|} \quad (4.4)$$

where  $\vec{p}_i$  is the momentum vector of the particle  $i$ . Summation is over all particles in the event. Thrust axis,  $\hat{t}$ , is a unit vector. (Principle) Thrust value,  $T$ , is 1 for a perfect pencil-like back-to-back two-jet event, and 0.5 for a perfect spherical event. The thrust value is useful in picking out back-to-back two-jet event. Thrust axis is useful to separate each jet in a back-to-back two-jet event. Aplanarity is another useful event shape variable that distinguishes spherical symmetrical events from planar and linear events. The definition is

$$S = \frac{3}{2}(\lambda_1), \quad (4.5)$$

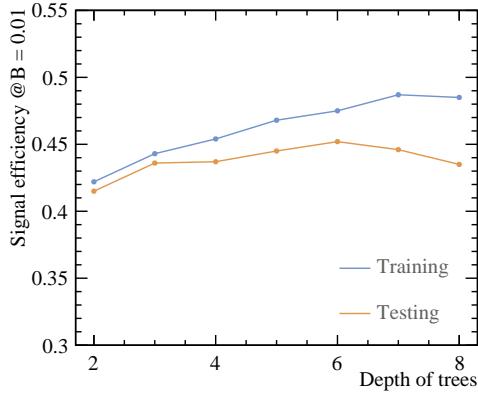
where  $\lambda_1$  is the largest eigenvalue in the diagonalised sphericity tensor,  $S^{\alpha\beta}$ .

## 4.7 Multivariate Analysis

Multivariate analysis (MVA) has become increasingly common in high energy physics. MVA is typically used as the last step of the physics analysis to select signal from background. It can be viewed as an advanced tool for regression or classification. Comparing to the traditional cut based method, modern machine learning technique offers much improvement in data analysis. Software package for MVA used throughout this document is TMVA [59].

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

Strictly, there should be three statistically independent samples for the MVA. One sample is for the training. Another sample for the validation, including optimisation and checking for overfitting. The last sample is for testing. However, due to technical reason



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

(TMVA only natively supports two samples), sometimes the same sample is used for the validation and the testing, which is acceptable with large statistics.

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

### 4.7.1 Optimisation and overfitting

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

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

Figure 4.6 shows a typical overfitting plot. Overfitting is defined when the efficiency of signal selection in the training samples increases, but the efficiency in the testing sample decreases. The example in figure 4.6 is chosen from double Higgs analysis at  $\sqrt{s}$

$= 3 \text{ TeV}$ , using Boosted Decision Tree model. The efficiency of signal selection is defined as the signal fraction when background fraction is 1%, report by the TMVA training process. In figure 4.6 , the depth of the tree, or the number of layers in the tree, reflects the complexity of the model. From tree depth 2 to 5, the efficiency for both testing and training samples increases. From tree depth 6 onwards, the overfitting occurs. In this particular example, one should choose a tree depth fewer than 7 to avoid overfitting.

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

### 4.7.2 Choice of models

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

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

#### Rectangular Cut

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

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

Because its simplicity, the cut method is often performed manually, much more often in the time pre-date the wide spread of machine learning methods. It is still commonly used for the pre-selection step before the MVA (see section 7.7), and other simple cases. Unless specified, the optimal cuts proposed in this document for various physics analysis are found using the rectangular cut method manually.

## Projective Likelihood

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

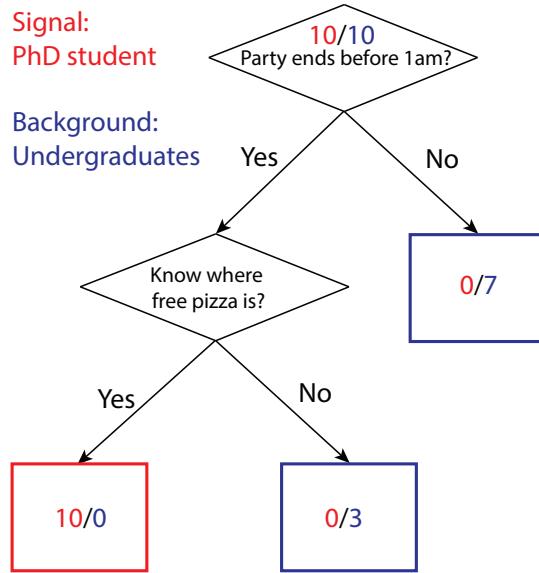
Likelihood classifier implemented in the TMVA calculates the probability density for each discriminative variable, for signal and background (hence PDE approach). The overall signal and background likelihood are defined as products of the individual probability density. The likelihood ratio,  $R$ , is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

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

## Decision tree

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

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



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

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

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

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

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

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

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

where free pizza is”,  $p = \frac{10}{15}$ . Gini index is  $2p(1-p) \simeq 0.44$ . Therefore, the first cut is “Party ends before 1am”.

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

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

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

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

### Boosted decision tree

Boosted decision tree (BDT) contains a forest of decision trees , where each tree is iterated many times using a technique called boosting. By overcoming the instability of a single decision tree, BDT is often regarded as the best out-of-the-box machine learning method. There are two common boosting methods: adaptive boosting and gradient

boosting. First introduced in [62], the adaptive boosting is discussed in further details, as it is simpler to understand than gradient boosting.

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

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

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

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

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

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

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

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

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

The most important parameter is the depth of a tree, which determines how many end nodes a tree has, or the degrees of freedom of a tree. The related parameter is the number of trees. Experience shows that using many small trees yields the best result. The performance as a function of the depth is shown in figure 4.6.

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

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

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

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

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

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

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

For the end node, it is determined as either signal-like or background-like, based on the majority for the training event in the end node. Numerically, it corresponds to 1/0. However, the end node could also use signal purity as the output, resulting in a continues

spectrum of [0,1]. The adaptive boosting algorithm is modified for the output value continues spectrum.

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

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

#### 4.7.3 Multiple classes

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

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

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

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

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



# Chapter 5

## Photon Reconstruction in PandoraPFA

*“When it is obvious that the goals cannot be reached, don’t adjust the goals, adjust the action steps.”*

— Confucius, 551 BC-479 BC

Correctly reconstructing photon results in an excellent single photon energy resolution. Such a good energy resolution is necessary to reconstruct heavy particles using decay channels involving photons, such as tau lepton decay or  $\pi^0$  decay.

Another aspect of the photon reconstruction is the photon separation resolution, where spatially close-by photons should be resolved whenever possible. Hence the photon resolution should provide reconstructed photons with high complexity and purity, making the photon number a good physics quantity for photon-counting experiment. The most recent photon-counting example benefited from this photon reconstruction is the  $H \rightarrow \gamma\gamma$  simulation study at  $\sqrt{s} = 3$  TeV at the CLIC [64].

An efficient photon reconstruction in a dense jet environment improves the overall reconstruction. As the particle flow approach to the calorimetry aims to reconstruct each individual particle, by assigning correct calorimeter hits to photons, assigning the remaining hits to the tracks for charged particles becomes an easier problem. Since the essential part of the particle flow reconstruction is the track-cluster association, a high-performance photon reconstruction thus leads to a reduced-density environment for charged-particle reconstruction, which in return improves the jet energy resolution.

Part of this chapter has been published in a conference proceeding [65].

## 5.1 Overview of photon reconstruction in PandoraPFA

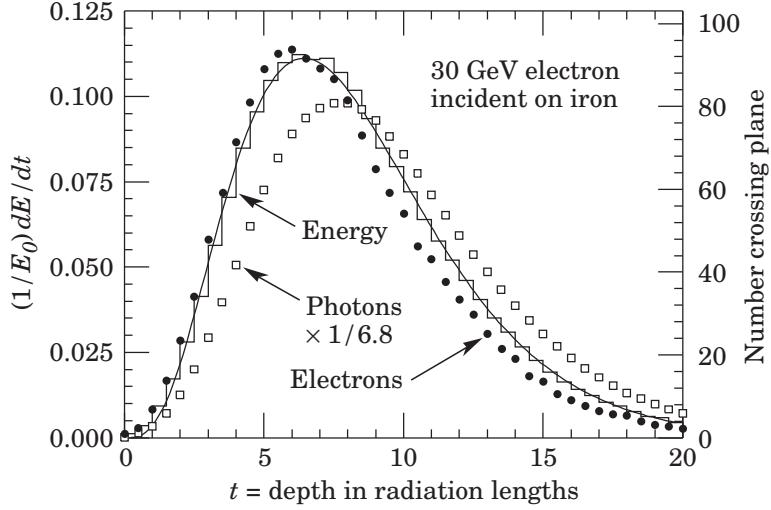
Photon reconstruction consists of various algorithms. Each deals with a particular aspect of the photon reconstruction. The stand-alone PHOTON RECONSTRUCTION algorithm is the main powerhouse. It carefully reconstruct photons from calorimeter hits in the ECAL, with special treatments for photons close to charged particles. This algorithm is implemented at the early stage of the reconstruction (see section 4.4.3), aiming to provide a clean environment for the subsequent charged-particle reconstruction.

Two photon fragment removal algorithms for fragments in the ECAL, and one algorithm for fragments in the HCAL, improve the single photon energy resolution. The photon splitting algorithm separate accidentally merged photons, which helps photon separation resolution.

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

## 5.2 Electromagnetic shower

An electromagnetic (EM) shower refers to the pair production and bremsstrahlung when a high energy photon or electron passing though a thick absorber. The interaction generates many low energy photons and electrons, producing a shower-like structure. Two suitable length scales to describe the EM shower are radiation length and Molière radius. The radiation length of a material is used to describe the longitudinal shower profile, defined as the mean distance travelled where an electron loses its energy by a factor of  $1/e$  via bremsstrahlung. It is also defined as the mean free path for pair production by a high energy photon [66]. The Molière radius of a material is a suitable length describe the transverse shower profile.



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

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

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

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

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

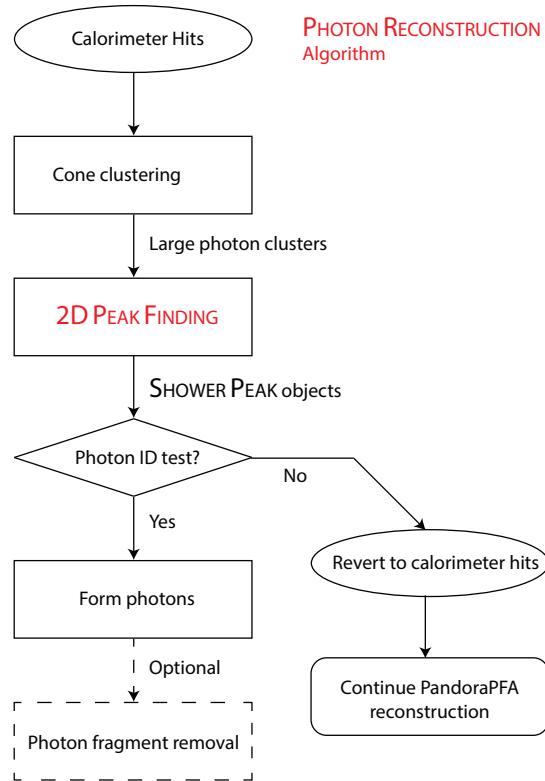
where  $E_c$  is the critical energy. This parametrisation should only be used to describe an average behaviour of the EM shower, as the fluctuation is important. For the photon identification, the longitudinal shower profile is compared with equation 5.1.

The transverse shower profile can be described as a narrow cone widening as the shower develops. 90% of the energy is contained in a fiducial cylinder with a radius of one Molière radius. Transverse profile is often represented by a sum of two Gaussian

function. The Gaussian nature of the transverse profile allows the separation of two EM showers using a two dimensional peak finding algorithm.

### 5.3 Photon Reconstruction algorithm

The PHOTON RECONSTRUCTION algorithm is a photon reconstruction and identification algorithm at the early stage of the reconstruction. It corresponds to “Particle ID” stage in section 4.4.3 of the PandoraPFA reconstruction chain. Main steps of PHOTON RECONSTRUCTION are: coarsely forming photon clusters, finding photon candidates, photon ID test, and optional fragment removals. Finding photon candidates uses the transverse shower profile information, which requires a two dimensional peak finding algorithm, further explained section 5.4. The photon ID test involves a multi dimensional likelihood classifier, which is described in section 5.5. The optional fragment removal algorithm shares a common base case class with another photon fragment removal algorithm. Hence they are discussed together in section 5.6 . A flow diagram of the PHOTON RECONSTRUCTION algorithm is shown in figure 5.2.



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

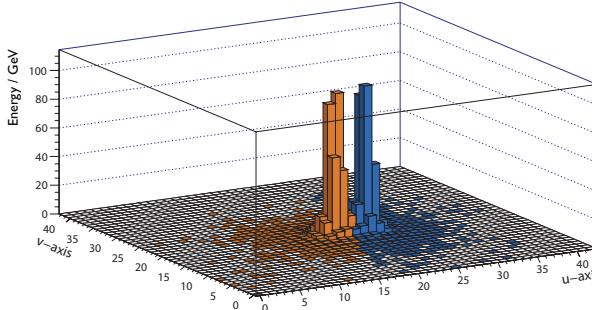
### 5.3.1 Form photon clusters

The input of the PHOTON RECONSTRUCTION algorithm is a collection of calorimeter hits. Muon clusters have been removed prior to this step. This step finds large potential photon clusters in the ECAL, which may contain multiple photons. For simplicity, the algorithm opts to reuse the cone clustering algorithm (see section 4.4.4) provided inside PandoraPFA to find large clusters. Since the target for reconstruction is neutral electromagnetic shower in the ECAL, the cone cluster algorithm is set to use high energy calorimeter hits as initial seeds. The parameters for the cone clustering are such that forming large clusters is preferred.

### 5.3.2 Find photon candidates

The next stage is to refine the large photon clusters into smaller photon candidates. Each photon candidate should contain calorimeter hits from one photon only. The aim for this step is to split a three dimensional cluster into several smaller clusters. The three dimensional splitting problem is hard. Therefore, a translation is needed to map the three dimensional problem to a more manageable two dimensional problem. The translation relies on the transverse distribution of characteristic EM showers, which is characterised by a narrow cone, widening while the shower develops. Therefore, along the direction of the photon, an EM shower can be modelled as a dense shower core with peripheral hits. When the energy deposition is projected on to a plane, the EM shower core would appear as a peak in the plane. By identifying a peak in the two dimensional plane, an EM shower core is identified. Hence three dimensional cluster splitting problem is mapped to a two dimensional peak finding problem with the projection translation. Figure 5.3 shows an example of a large photon cluster projected on to a two dimensional plane, where two EM showers are identified.

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



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

### 5.3.3 Photon ID test

SHOWER PEAK objects from the 2D PEAK FINDING algorithm are tested for photon tagging. A set of discriminating variables that exploit features of electromagnetic showers are calculated. A multidimensional likelihood classifier is implemented, which needs to be trained before applying. The response from the classifier determines if a SHOWER PEAK object is a photon. If it is a photon, it would be tagged and separated from the event. If it is not a photon, the calorimeter hits of the SHOWER PEAK object will be passed on to the next stage of the reconstruction (see figure 5.2). Because the classifier is complicated, it is discussed in section 5.5.

### 5.3.4 Photon Fragment removal

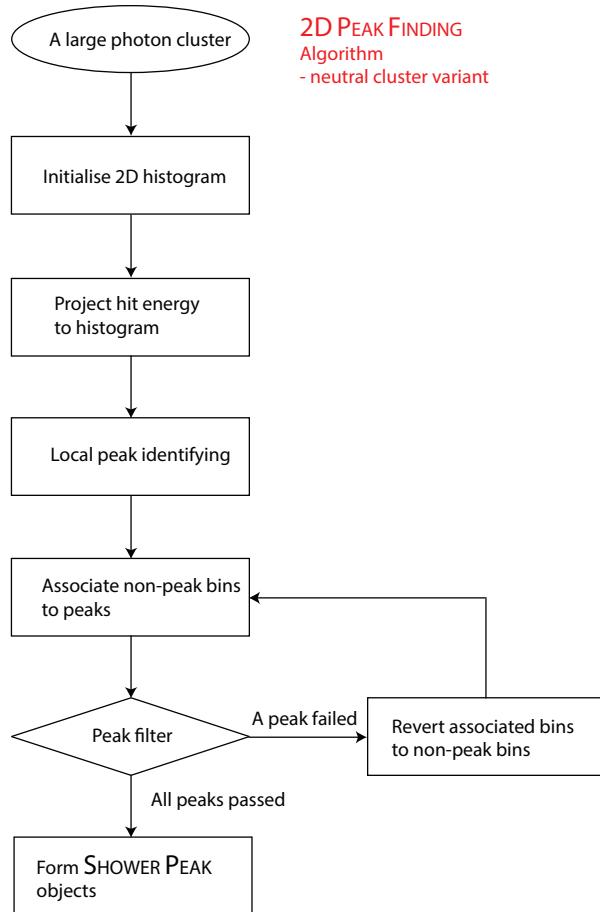
The optional photon fragment removal algorithm aims to merge small photon fragment to identified photons. Since this step shares the same logic as the fragment removal algorithm in section 5.6, this step is described in details in section 5.6.

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

## 5.4 Two dimensional peak finding algorithm for photon candidate

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

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



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

### 5.4.1 Initialise 2D histogram

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

### 5.4.2 Project hit energy to histogram

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

$$\vec{s}_i = \frac{\vec{a}_i - \langle \vec{a} \rangle}{d_{cell}}, \quad (5.3)$$

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

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

### 5.4.3 Local peak identifying

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

#### 5.4.4 Associate non-peak bins to peaks

Having tagged all local peaks, this step associates non-peak bins to peaks based on the distance to the peak and the energy of the peak. The energy dependence is needed as the transverse EM shower width increases with the increase of the energy. After all local peak bins are found, non-peak bins are associated to a peak bin, by choosing the peak bin that minimise the metric

$$\frac{d}{\sqrt{E_{peak}}} \quad (5.4)$$

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

#### 5.4.5 Peak filtering

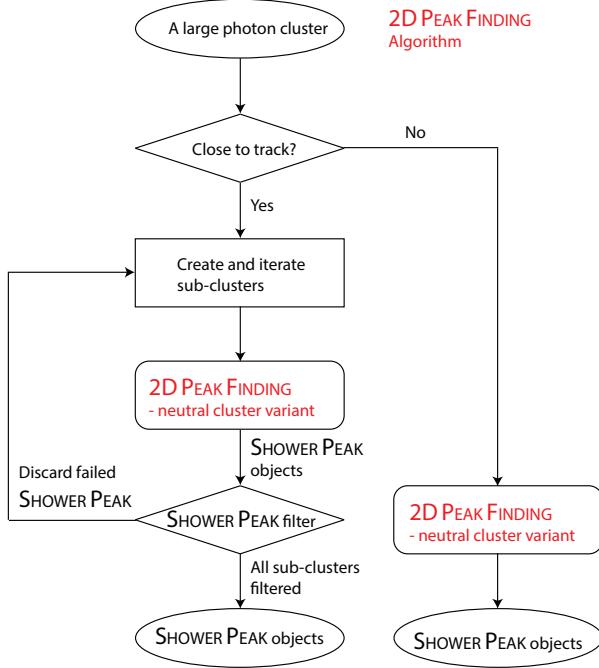
The performance of the two dimensional peaking finding algorithm is improved by clever programming and physics arguments. For a given two dimensional histogram, such as the one in figure 5.3, major peaks most likely correspond to physical photons, while the minor peaks more likely come from fluctuations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

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

This marks the end of the PHOTON RECONSTRUCTION algorithm with neutral clusters, outlined in figure 5.4.

### 5.4.6 Candidate close to track projection

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



**Figure 5.5:** Flow chart for 2D PEAK FINDING algorithm.

Figure 5.5 shows the flow chart for the full 2D PEAK FINDING algorithm, including the treatment to clusters close to tracks. If a cluster is less than 3 mm from the closest track projection, it is treated as a potentially charged cluster. The ECAL is sliced longitudinally to help identify photon candidates. For example, the default three slices will result in three ECAL fiducial spaces, which cover spaces from the front of the ECAL to a third, two thirds and the back of the ECAL, respectively. The peaking finding algorithm is repeated for sliced clusters contained in each fiducial space. A peak and corresponding SHOWER PEAK object are only preserved as a photon candidate if the peak exists for every sliced cluster, and if its position is shifted by no more than one

neighbouring bin between sliced clusters. Furthermore, if a peak bin is within the eight neighbouring bins of the track projection, the peak and its associated bins are flagged as charged particles (non-photon).

### 5.4.7 Inclusive mode

The two dimensional histogram is iterated many times during the algorithm. The time complexity is  $O(n^2)$  for a  $n \times n$  histogram (Default  $n = 41$ ). Therefore, for the purpose of speed, it is undesirable to have a very large histogram. Having a small finite histogram speeds up the calculation. However, because of the finite size, only energy deposition projected onto the histogram would be considered for peak finding. Calorimeter hits outside the histogram would be lost when SHOWER PEAK objects are constructed. This behaviour is suitable for PHOTON RECONSTRUCTION algorithm (section 5.3) and for photon fragment removal (section 5.6). However, for photon splitting (section 5.8), there should be no calorimeter hits loss from splitting a photon. Hence inclusive mode of the peak finding algorithm is developed, and it allows energy deposition projected outside the histogram to be associated with identified peaks.

## 5.5 Likelihood classifier for photon ID

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

### 5.5.1 Projective Likelihood in PandoraPFA

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

Two variables exploit the longitudinal EM shower distribution.  $t_0$  is the start layer from the longitudinal shower profile (see section 5.2), shown in figure 5.6a.  $\delta l$  is fractional

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

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

difference from the expected shower profile [23]:

$$\delta l = \frac{1}{E_0} \sum_i |\Delta E_{obs}^i - \Delta E_{EM}^i|. \quad (5.5)$$

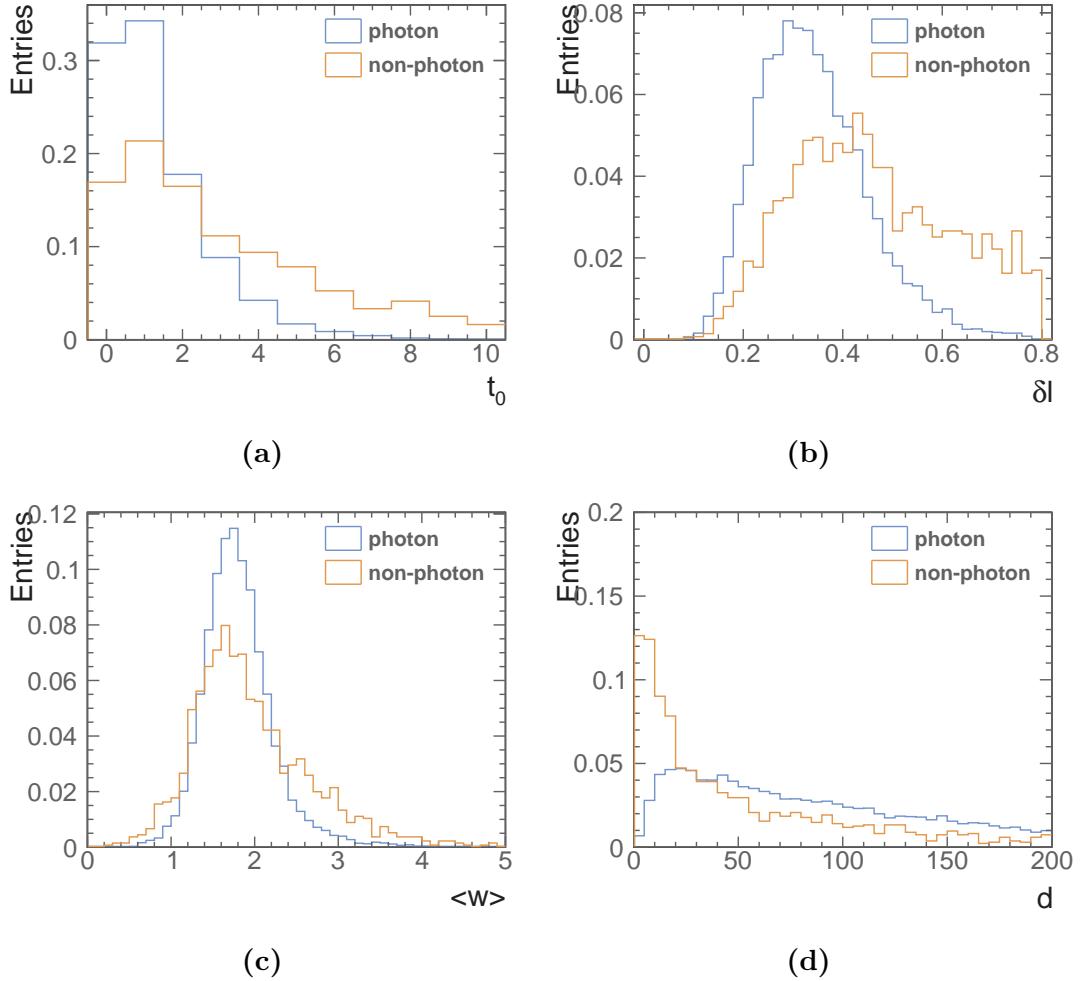
$\delta l$  is minimised as a function of the  $t_0$ . The  $\delta l$  distribution is shown in figure 5.6b. For a photon,  $t_0$  and  $\delta l$  are expected to be close to 0.

Three variables use the transverse shower information.  $\langle w \rangle$  is the energy weighted r.m.s. distance of a bin to its peak, shown in figure 5.6c. This is a measure of the transverse shower size.  $\delta \langle w_{UV} \rangle$  is the smallest ratio of the two r.m.s. distances in each U, V axis direction, a measure of the circularity of the transverse shower. Last variable,  $\delta E_{cluster}$ , is the ratio between the photon candidate energy to the big cluster energy. This is a measure of the dominance of a photon in a large cluster.

Last variable for the classifier,  $d$ , is the distance between the photon candidate and the closest track projection. It is less likely to be a photon if it is close to a track. Its distribution is shown in figure 5.6d.

The distributions of these variables are normalised to probability distribution, stored in binned histograms. The classifier is improved by realising the variable distributions varies with photon energies. Thus these distributions are divided by bins of photon candidate energy. The default energy bins edges are 0.2, 0.5, 1, 1.5, 2.5, 5, 10, 20 GeV, which covers a good range of photon energies. The classifier training typically uses simulated 250 GeV jet events, with sample  $Z' \rightarrow u\bar{u}/d\bar{d}/s\bar{s}$ . High energy jets allow the training for high energy photon candidates.

Projective likelihood model with probability density estimators is used in PandoraPFA for the photon ID due to its simplicity and low requirement on computing resources. The details are described in section 4.7.2.



**Figure 5.6:** Variables for photon ID test, obtained from simulated 500 GeV di-jet  $Z' \rightarrow u\bar{u}/d\bar{d}/s\bar{s}$  events. Photon ID in plots are determined using MC information. Figure 5.6a, Figure ??, , Figure 5.6c, and Figure 5.6d show distributions for start positions of longitudinal shower profiles, fractional difference from the expected longitudinal shower profile, energy weighted r.m.s. distance of a bin to its peak, and distance between the photon candidate and the closest track projection.

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

$$pid = \frac{N \prod_i P_i}{N \prod_i P_i + N' \prod_i P'_i} \quad (5.6)$$

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

During classification, a candidate passes the photon ID test if

$$\begin{cases} pid > 0.6, & \text{if } 0.2 < E < 0.5 \text{ GeV} \\ pid > 0.4, & \text{if } E \geq 0.5 \text{ GeV} \end{cases} \quad (5.7)$$

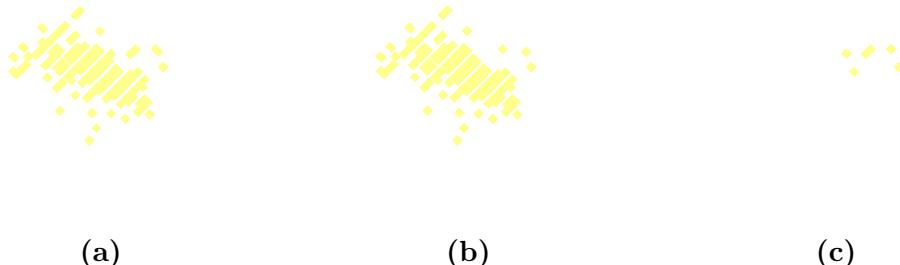
where  $E$  is the candidate energy. Two values of the  $pid$  cuts reflect the confidence of the photon ID test with different candidate energy. The test is more cautious with low energy candidate.

## 5.6 Photon fragment removal algorithm in the ECAL

During the reconstruction, it is possible that a core of the photon electromagnetic shower is identified as a photon (the main photon). The outer part of the shower is reconstructed as a separate particle, and wrongly identified as a photon or a neural hadron. Figure 5.7 shows a typical creation of such a photon fragment. A fragment does not have the electromagnetic shower structure, and typically it has much lower energy than a proper photon. If a photon-fragment pair is merged, the pair should be consistent with an one-particle profile.

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

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



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

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

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

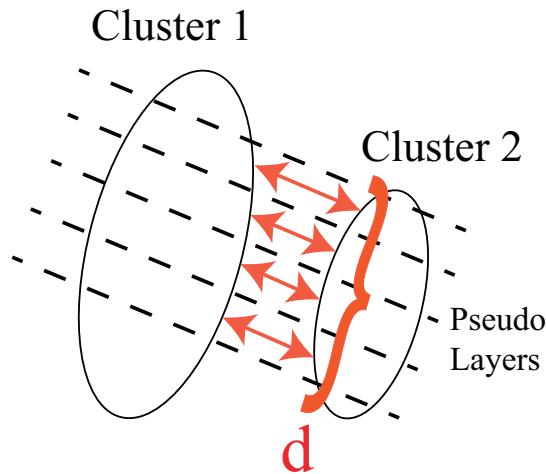
Table 5.2 lists cuts for merging photon-photon-fragment pairs and photon-neutral-hadron-fragment pairs for both low energy and high energy fragments.  $d$ ,  $d_c$  and  $d_h$  are mean energy weighted intra-layer distance within the pair, distance between two centroids, and minimum distance between calorimeter hits of each PFO in the pair, respectively. Three distance measurements have subtle difference.  $d_c$  gives the distance between centroids of each PFO in the pair, which is a computationally quick but crude measurement.  $d_h$  is the minimum distance between calorimeter hits of each PFO in the pair. For a true photon-fragment,  $d_h$  should be close to zero as the pair should be spatially close.  $d$  is the mean energy weighted intra-layer distance between each PFO in the pair (see figure 5.8):

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

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

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

where  $i$  indicates  $i^{th}$  pseudo-layer of the ECAL.  $d_{l,i}$  is the minimum distance between calorimeter hits of the pair in the  $i^{th}$  pseudo-layer.  $E_{f,i}$  is the energy of the fragment in the  $i^{th}$  pseudo-layer.  $d$  is a better measurement of the closeness of the pair. All three distance metrics should be small to merge a photon-fragment pair.



**Figure 5.8:** Illustration of distance metric,  $d$ .

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

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

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

Comparing cuts with low  $E_f$  and high  $E_f$ , the cuts are similar. Cuts for high  $E_f$  allow higher energy fragments to be merged.

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

are more likely to be low energy, whilst high energy neutral fragments are more likely to be photon fragments.

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

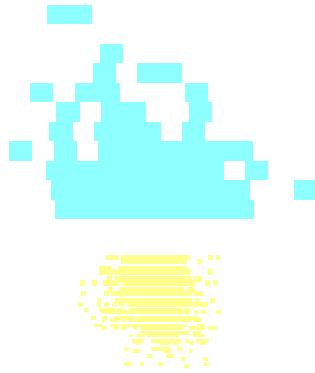
## 5.7 Photon fragment recovery algorithm in the HCAL

Section 5.6 described an effective algorithm to remove photon fragments that are peripheral to the main photon, the electromagnetic shower core. An example of such fragment is shown in figure 5.7. There is another type of fragments originated from the leakage effect of the ECAL. When a high energy EM shower is not fully contained in the ECAL, shower deposits energy in the HCAL, which often forms a neutral hadron in the HCAL. Previous algorithms only consider calorimeter hits in the ECAL. Therefore no attempts have been made to recover photon fragments in the HCAL. This section presents an algorithm to merge photon fragments in the HCAL. This photon fragment recovery algorithm is important when reconstructing high energy photons.

An example of a 500 GeV photon reconstructed into a main photon in the ECAL (yellow) and a neutral hadron fragment in the HCAL (blue) is shown in figure 5.9. For the ILD detector, this ECAL leakage effect appears when the photon energy is above 50 GeV.

Shown in figure 5.9, photon fragments in the HCAL is spatially close to the main photon. A fitted cone from the main photon, if extended to the HCAL, covers most of the fragment. These features allow a set of cuts developed to merge fragments in the HCAL, listed in table 5.3

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



**Figure 5.9:** An event display of a typical 500 GeV photon, reconstructed into a main photon in the ECAL (yellow) and a neutral hadron fragment in the HCAL (blue).

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

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

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

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

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

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

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

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

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

## 5.8 Photon splitting algorithm

Algorithms described above deal with forming photons from calorimeter hits in the ECAL, merging photon fragments in the ECAL and the HCAL. Another aspect in photon reconstruction is splitting accidentally merged photons. During the particle reconstruction, it is possible that photons are accidentally merged if they are spatially

close. Hence another algorithm at the end of the particle reconstruction addresses this issue and tries to split merged photons. This algorithm focuses on energetic photons.

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

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

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

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

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

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

## 5.9 Characterise the performance

Motivations and implementations of four different photon related algorithms have been described above. The main photon reconstruction algorithm in section 5.3 improves the

photon completeness and the photon pair resolution, due to the improved two dimensional peak finding algorithm in section 5.4. The fragment removal algorithms in section 5.6 and section 5.7 further reduce the photon fragments in the ECAL and the HCAL. The photon splitting algorithm in section 5.8 exploits the peak finding algorithms to separate photons using transverse shower information, which improves the photon separation resolution. Photon reconstruction improves single photon resolutions. It also improves jet energy resolution at high  $\sqrt{s}$  because of the high photon reconstruction completeness.

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

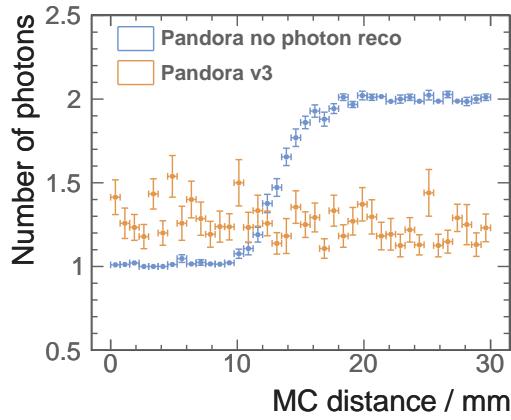
- No stand-alone photon reconstruction algorithms
- With a stand-alone photon reconstruction algorithm from PandoraPFA version 1.
- With full photon related algorithms described above, incorporated in PandoraPFA version 3.

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

## 5.10 Compare with no photon reconstruction

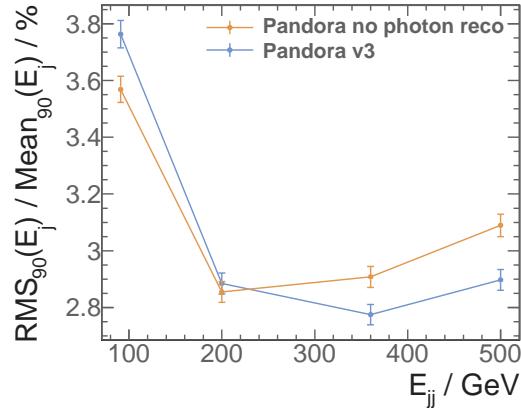
This section compares the performance with and without photon related algorithms using photon pair and jet samples. The nominal ILD detector model is used. The single and two photon events were generated with a uniform distribution in the solid angle for a range of the opening angles between the pair. Events are selected such that there is no early photon conversion and the Monte Carlo photon deposits energies in the calorimeter. The events are further restricted to photon decaying in barrel and end cap region only, to minimise the detector effect.

Figure 5.10 shows the photon reconstruction for two spatially close photons reconstructed without and with photon algorithms. Without the photon related algorithms, fragments are produced. The number of photons between 0 and 10 mm separation is around 1.5. The true photon number for that region should be 1, as it is extremely challenging to separate photons less than two cells apart. Without the photon related algorithms, photon separation resolution is much worse, as the number of photon appears



**Figure 5.10:** Average number of photons using two photons of 500 and 50 GeV per event sample reconstructed without and with photon algorithms.

to be flat between 0 to 30 mm separation. With the photon related algorithms, two photons start to be separated at 10 mm and fully separated at 20 mm separation.



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

The improvement in completeness and resolution in photon reconstruction, leads to a considerable improvement in the jet energy resolution at high energy. Jet energy resolution is defined as the root mean squared divided by the mean for the smallest width of distribution that contains 90% of entries, using  $Z' \rightarrow u\bar{u}/d\bar{d}/s\bar{s}$  sample at barrel region. The angular cut is to avoid the barrel/endcap overlap region. The light quark decay of the Z is used as PandoraPFA does not attempt to recover missing momentum from semi-leptonic decay of heavy quarks. Using 90% of the entries is robust and focus on the Gaussian part of the distribution. The total jet energy is sampled at 91, 200,

360 and 500 GeV. Shown in figure 5.11, the jet energy resolutions are much better at 360 and 500 GeV with improved photon reconstruction. By identifying photons before reconstructing charged particles in a dense jet environment, the reconstruction task is easier and less likely to make mistake. However, at low energy, the reconstruction is worse with photon algorithms, because photon algorithms are optimised for high energies.

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

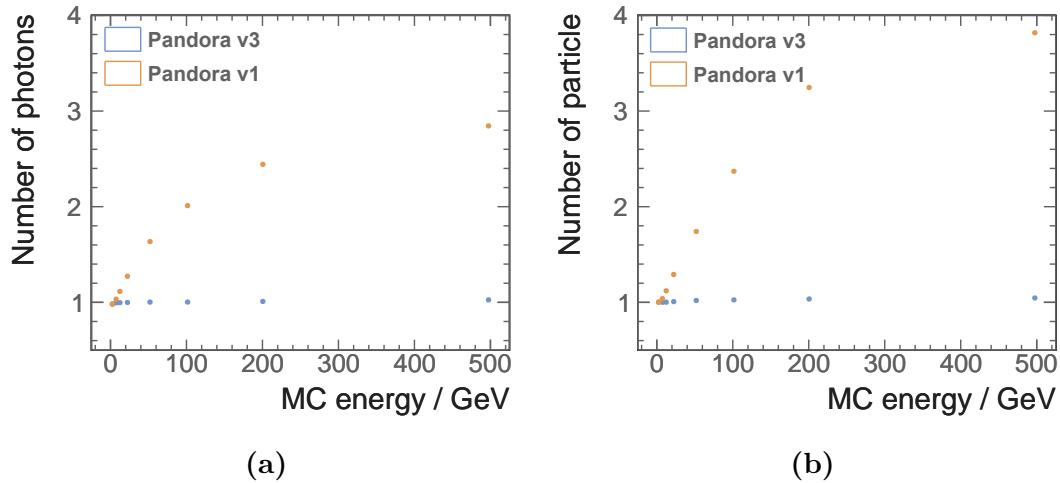
Photon confusion	$\sqrt{s} = 91$ GeV	200 GeV	360 GeV	500 GeV
PandoraPFA without photon algorithms / %	0.7	0.9	1.3	1.4
PandoraPFA with photon algorithms / %	1.4	0.9	0.9	0.9

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

## 5.11 Compare with photon reconstruction in PandoraPFA version 1

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

Figure 5.12a shows the reduction in fragments identified as photons, using a single photon per event sample. Indicating as the blue dots on the plots, average number of photon stays below 1.05 even at 500 GeV (true value 1). For a 100 GeV photon, the

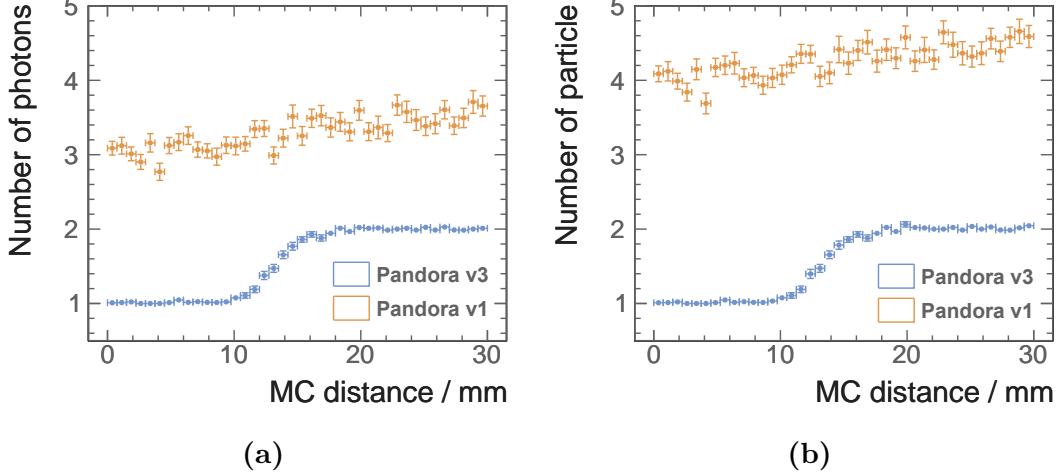


**Figure 5.12:** Figure 5.12a and figure 5.12b shows the average number of reconstructed photons and reconstructed particles, as a function of their true energy using a single photon per event sample. The top orange and bottom blue dots are reconstructed with PandoraPFA version 1 and version 3. The photon reconstruction is changed in PandoraPFA version 2.

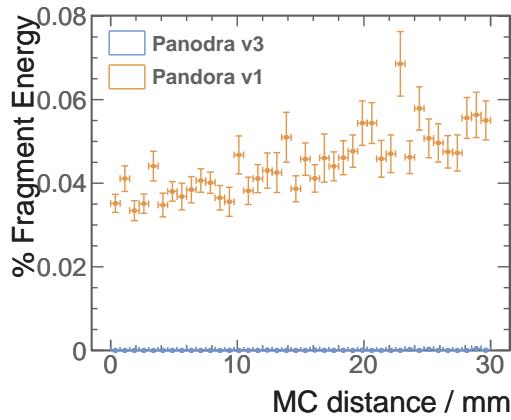
average number of photons is reduced to 1 from 2. For a 500 GeV photon, the number is reduced to 1.05 from 2.8.

A similar trend is shown in figure 5.12b, where the extra fragments identified as neutral hadrons have also taken into account. For a 100 GeV photon, the average number of particles is reduced to 1.24. For a 500 GeV photon, the number is reduced to 1.05 from 3.8.

Figure 5.13 illustrates a reduction in the photon fragments and the neutral hadron fragments using two photons of 500 and 50 GeV per event sample. The figure shows the MC distance separation from 0 to 30 mm, which corresponds to approximately 6 ECAL square cells of the default ILD detector model. This is a difficult test for fragment removal as high energy photons are more likely to create fragments and the imbalance in the two photon energies makes it more difficult to separate correctly. In both figure 5.13a and figure 5.13b, the average numbers of photon and particle are below 2.05 at 30 mm apart, which is significantly better than the reconstruction in PandoraPFA version 1. Two photons start to be resolved at 10 mm apart, and fully resolved at 20 mm apart. The resolution is better than reconstruction in PandoraPFA version 1, which is difficult to extract due to excess fragments.

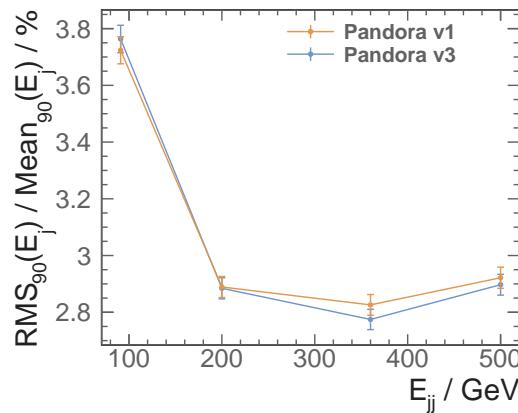


**Figure 5.13:** Figure 5.13a and figure 5.13b shows the average number of reconstructed photons and reconstructed particles, as a function of the MC distance separation in the calorimeter, using two photons of 500 and 50 GeV per event sample. The top orange and bottom blue dots are reconstructed with PandoraPFA version 1 and version 3. The photon reconstruction is changed in PandoraPFA version 2.



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

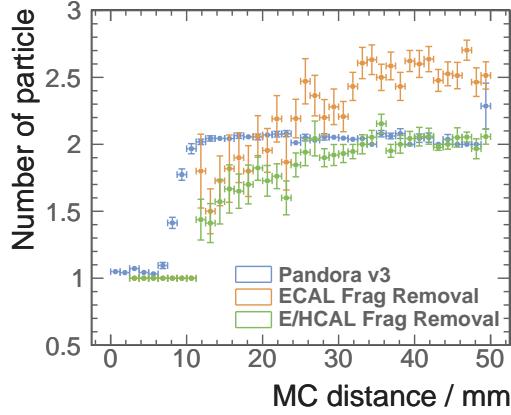
Another metric to reflect the improvement in photon reconstruction is the fraction of the fragment energy to the total energy as function of the distance separation. Shown in figure 5.14, using two photons of 500 and 50 GeV per event sample, a reduction in fragment energy can be seen clearly. With improved reconstruction, the average fragment energy fraction is below 0.1% up to 30 mm apart, whilst around 5% energy would be reconstructed in fragments with PandoraPFA version 1.



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

The improvement in completeness and resolution in photon reconstruction, as shown in single photon and double photon reconstruction, leads to a small improvement in the jet energy resolution at high energy. The total jet energy is sampled at 91, 200, 360 and 500 GeV. Shown in figure 5.15, the jet energy resolutions are better at 360 and 500 GeV with improved photon reconstruction. This is due to more aggressive photon reconstruction especially nearby tracks, which is more useful at a high-energy dense jet environment.

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



**Figure 5.16:** Figure shows the average number of photons, as a function of the MC distance separation in the calorimeter, using two photons of 500 and 500 GeV per event sample. The blue, orange, and green dots are reconstructed with PandoraPFA version 3, PandoraPFA version 1 with fragment removal in the ECAL, and PandoraPFA version 1 with fragment removal in the ECAL and the HCAL. The photon reconstruction is changed in PandoraPFA version 2.

## 5.12 Understand photon reconstruction improvement

As stated before, photon reconstruction algorithm in section 5.3 and photon splitting algorithm in section 5.8 improves the photon completeness and the photon pair resolution. The fragment removal algorithm in section 5.6 removes fragments in the ECAL. High energy fragment removal algorithm in section 5.7 removes fragments in the HCAL. To show the incremental improvement, the average number of particle for a high energy photon pair, 500 - 500 GeV is shown in figure 5.16. Blue, orange, and green dots represent full photon reconstruction, only fragment removal algorithms in the ECAL, and fragment removal algorithms in the ECAL and the HCAL, respectively.

With fragment removal algorithm in the ECAL, the number of fragment is reduced significantly (comparing with figure 5.13b). With the additional fragment removal in the HCAL, the number of fragments are reduced further. At 40 mm apart, with both fragment removal algorithms (green dots), there is less than 0.05 fragment per photon pair.

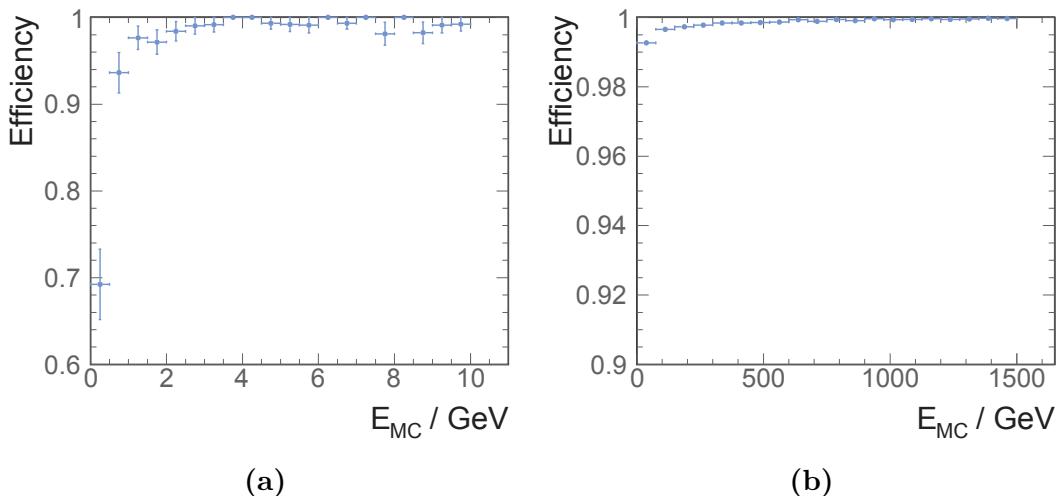
The introduction of the revised photon reconstruction and photon splitting improves the photon separation resolution. Photons pair starts to be resolved at 5 mm apart for 500 - 500 GeV pair and fully resolved at 15 mm apart. With previous photon reconstruction

in PandoraPFA version 1, the same photon pair starts to be resolved at 10 mm apart and fully resolved at around 40 mm apart.

## 5.13 Current photon reconstruction performance

In section 5.11, the improved performance of the photon reconstruction is demonstrated with different metrics, using single photon, double photons and jet samples. In this section, the performance of the photon reconstruction as a function of photon energies will be described.

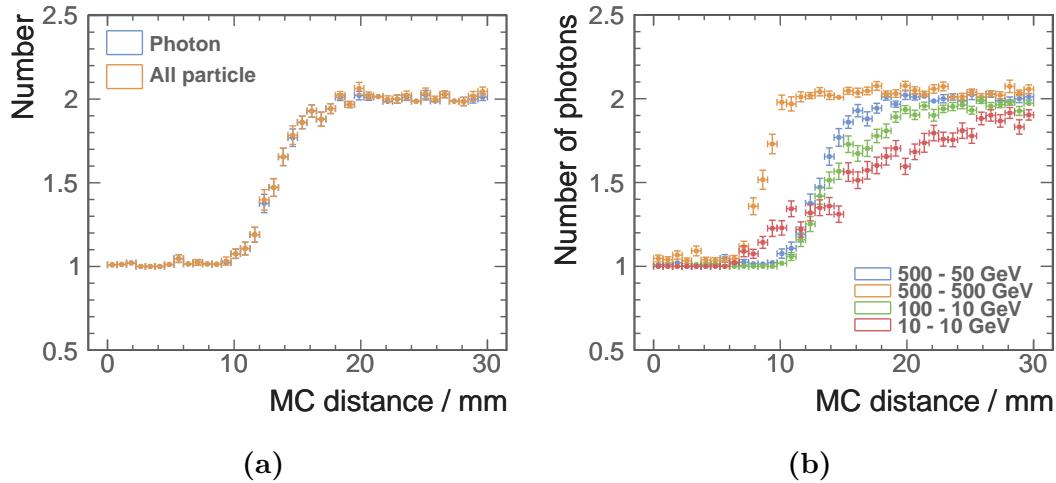
Single particle reconstruction and ID efficiency is demonstrated in figure 5.17. Using single photon samples, an event can have an efficiency of 1, if the photon reconstructed corresponds to the truth photon, or an efficiency of 0. The low efficiency in the first bin, from 0 to 0.25 GeV is because photon reconstruction does not attempt to reconstruct photons below 0.2 GeV. The single photon reconstruction efficiency is above 98% for photons above 2 GeV and above 99.5% for photons above 100 GeV.



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

For simple samples such as two photons per event, there are very few fragments. Shown in figure 5.18a for 500 and 50 GeV photons pair sample, the average number of photons and particles beyond 20 mm apart is less than 0.05 above the true value, 2.

The resolving power of a photon pair depends on energies of two photons. Figure 5.18b is an example of average number of photon reconstructed for different photon pairs.



**Figure 5.18:** Figure 5.18a shows the average numbers of photon and particle using two photons of 500 and 50 GeV per event sample. Figure 5.18b shows the average numbers of photon for four different photon pairs: 500 - 50, 500 - 500, 100 - 10 and 10 - 10 GeV.

When the energies of two photons are similar, the resolving power is greater. This is because that the two photon showers have similar sizes, and the 2D PEAK FINDING algorithm can exploit the symmetry in the size of the EM showers. For example, 500 - 500 GeV photon pair and 10 - 10 GeV photon pair start to be resolved at 6 mm apart, which is about 1 ECAL cell. The asymmetrical photon pair, 500 - 50 GeV and 100 - 10 GeV pair, starts to be resolved at 10 mm apart, which is about 2 ECAL cells.

For an energetic photon, it is more difficult to remove fragments, but it is easier to identify the photon. The electromagnetic shower core is more dominant than the peripherals. Therefore separating two energetic photons is easier than separating two low energy photons. This can be seen in figure 5.18b. At 20 mm apart, two photons in 500 - 500 GeV pair are fully resolved, where approximately only 60% of two photons in 10 - 10 GeV pair are resolved.

# Chapter 6

## Tau Lepton Decay Modes Classification

“MVA: Turn numbers into gold.”

— TMVA

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

The ability to identify tau decay mode is also a benchmark for detector performance. Since tau lepton has a very short life time [70], only its decay products can be detected via the tracking detectors and calorimeters. Therefore the performance of the calorimetric and track systems determines the ability to separate different tau decay modes.

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

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

## 6.1 Overview of the analysis

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

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

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

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

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

## 6.2 Decay modes

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

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

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

## 6.3 Simulation and reconstruction

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

Two million events were reconstructed with iLCSoft version v01-17-07 [71] and PandoraPFA version 3 [39], where the photon reconstruction is described in chapter 5.

## 6.4 Event pre-selection

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

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

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

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

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

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

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

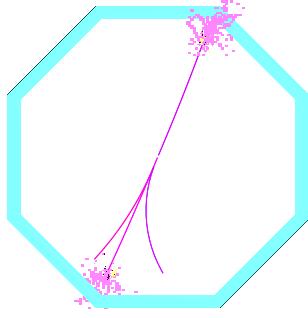
final states with more than one particles typically have more visible energies. The polar angle acceptance efficiencies depend on the final states, as light final states are boosted and more likely in the forward region.

## 6.5 Select single tau decay

The  $e^+e^- \rightarrow \tau^+\tau^-$  channel contains two tau leptons travelling in opposite directions. Since the tau decay final state separation is applicable to single tau, PFOs in one event are divided into two collections, each corresponding to one tau. Separation of PFOs uses the principle thrust axis vector, explained in the section 4.6.3, which is the axis that most PFOs aligned to. Two collections are obtained based on the sign of the scalar product between the thrust axis and the PFO momentum vector. An example of the simulated  $e^+e^- \rightarrow \tau^+\tau^-$  event is shown in figure 6.1.

## 6.6 Discriminative variables

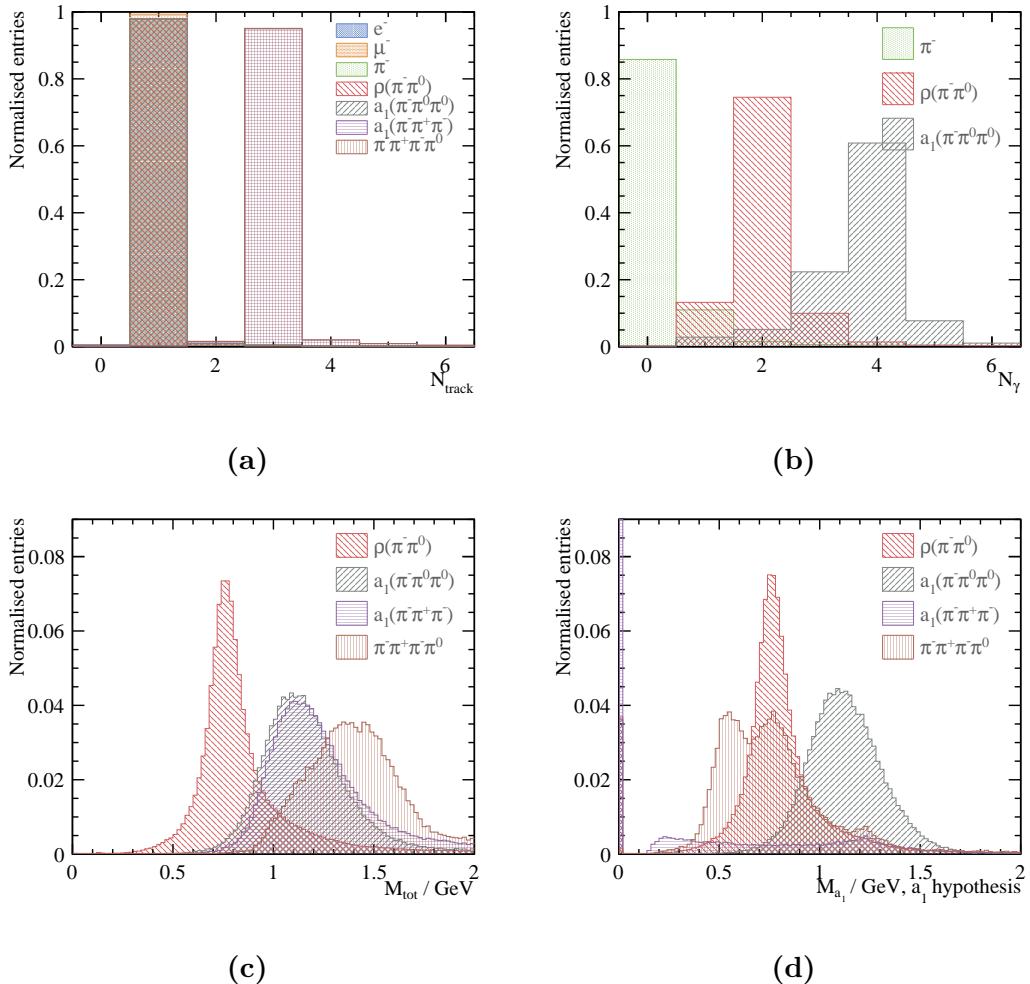
Having pre-selected events, a set of discriminating variables are carefully developed for the multivariate analysis (MVA). The full list of variables are shown in table 6.4.



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

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

**Table 6.4:** Variables used in the MVA



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

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

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

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

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

Extra variables are used to differentiate specific final states.

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

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

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

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

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

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

$$\chi^2 = \left( \frac{m_{tot} - m_{a_1}^{MC}}{\sigma_{a_1}^{MC}} \right)^2 + \left( \frac{m_{\gamma_1\gamma_2} - m_{\pi^0}^{MC}}{\sigma_{\pi^0}^{MC}} \right)^2 + \left( \frac{m_{\gamma_3\gamma_4} - m_{\pi^0}^{MC}}{\sigma_{\pi^0}^{MC}} \right)^2, \quad (6.2)$$

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

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

### 6.6.2 Separate $e^-$ from $\pi^-$

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

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

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

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

## 6.7 Multivariate Analysis

For the multivariate analysis, the multiclass class of the TMVA package [72] was used to perform a multiclass classification, which trains the seven final states simultaneously. The multiclass class is an extension of the standard two-class signal-background classifier. The discussion on multivariate analysis can be found in section 4.7. The multiclass classifier is discussed in section 4.7.3.

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

## 6.8 Classification Efficiency

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

For the  $e^-$  decay mode, a high correct classification is achieved, due to using particle ID from PandoraPFA and calorimetric information. Similarly for  $\mu^-$  decay mode, 99.5%

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

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

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

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

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

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

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

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

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

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

## 6.9 Electromagnetic calorimeter optimisation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

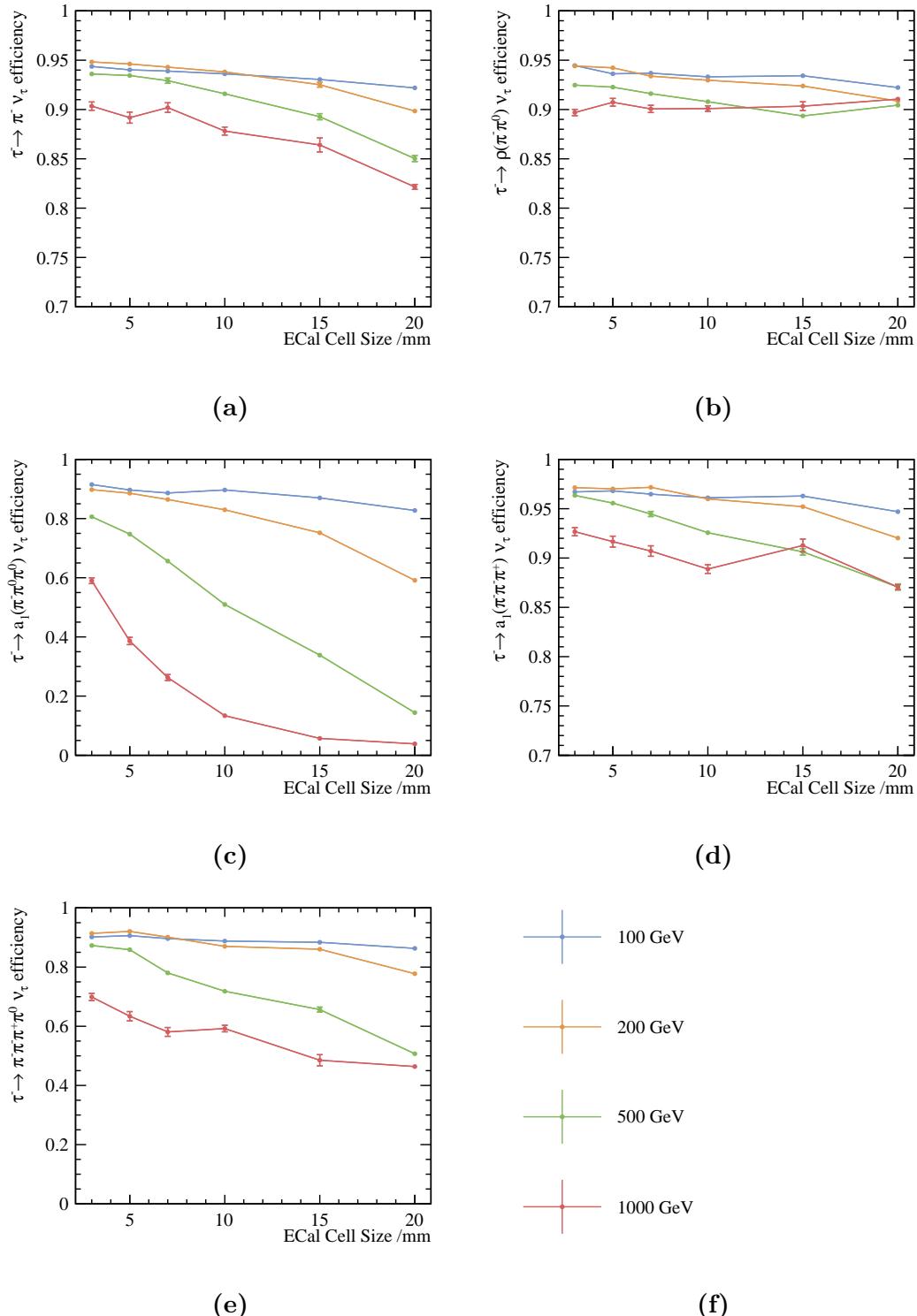
### 6.9.1 Tau hadronic decay correct classification efficiency

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

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

$$\varepsilon_{had} = \frac{\sum_i^5 Br_i \varepsilon_i}{\sum_i^5 Br_i}, \quad (6.3)$$

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



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

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

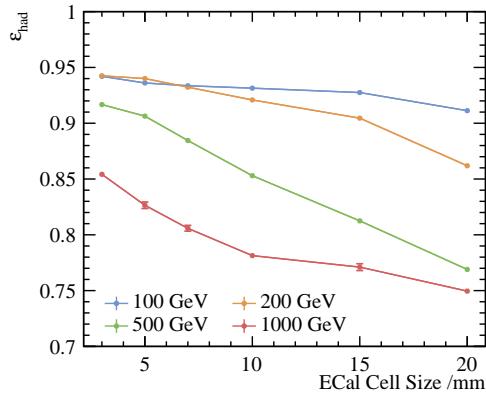
For  $\varepsilon_{had}$  at  $\sqrt{s} = 100$  GeV, the efficiency decreases from 94% at 3 mm cell size, to 91% at 20 mm cell size. The decrease is approximately proportional to the increase in the cell size.

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

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

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

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



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

## 6.10 Separate H from Z with tau pair decay

H can be separated from Z using tau pair decay channel. The difference in the spin of the boson reflects in the different polarisation correlation of the tau pair. By extracting the polarisation correlation, parent bosons can be separated. A proof-of-principle analysis is presented to demonstrate the physics usage of the tau decay mode classification, motivated by theoretical studies such as in [16]. The theoretical motivation is described in section 2.10. The subsequent sections discuss the ability to reconstruct the polarisation correlation of the Z using tau pair decay, and to match with the truth information.

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

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

### 6.10.1 Event pre-selection

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

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

### 6.10.2 Identify tau pairs

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

## Direct tau search

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

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

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

**Table 6.8:** Optimised parameters of modified ISOLATEDTAUIDENTIFER.

## Jet clustering

Tau hadronically decay can be also identified as a small jet. Durham algorithm [52] was used to jet clustering, also known as  $e^+e^- k_t$  algorithm. (see section 4.6.2). The jet algorithm runs in the exclusive mode to find four jets.

## Select tau candidates

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

the minimisation function is

$$\chi^2 = \frac{(m_{qq} - m_Z)^2}{\sigma_{m_{qq}}^2} + \frac{(E_{qq} - \frac{\sqrt{s}}{2})^2}{\sigma_{E_{qq}}^2}, \quad (6.4)$$

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

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

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

the pair with smallest  $\chi^2$  is chosen.

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

## Boost to Z decay rest frame

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

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

$$p_{\tau\tau}^\mu = \begin{pmatrix} \sqrt{s} \\ \sqrt{s} \times \sin(\theta_{beam}) \\ 0 \\ 0 \end{pmatrix} - \sum_i^{\text{non-}\tau} p_i^\mu, \quad (6.6)$$

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

$$p_{\tau\tau,\text{correct}}^\mu \equiv p_{\tau\tau}^\mu \times \frac{\frac{1}{2}\sqrt{s}}{E_{\tau\tau}}. \quad (6.7)$$

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

### 6.10.3 Variables

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

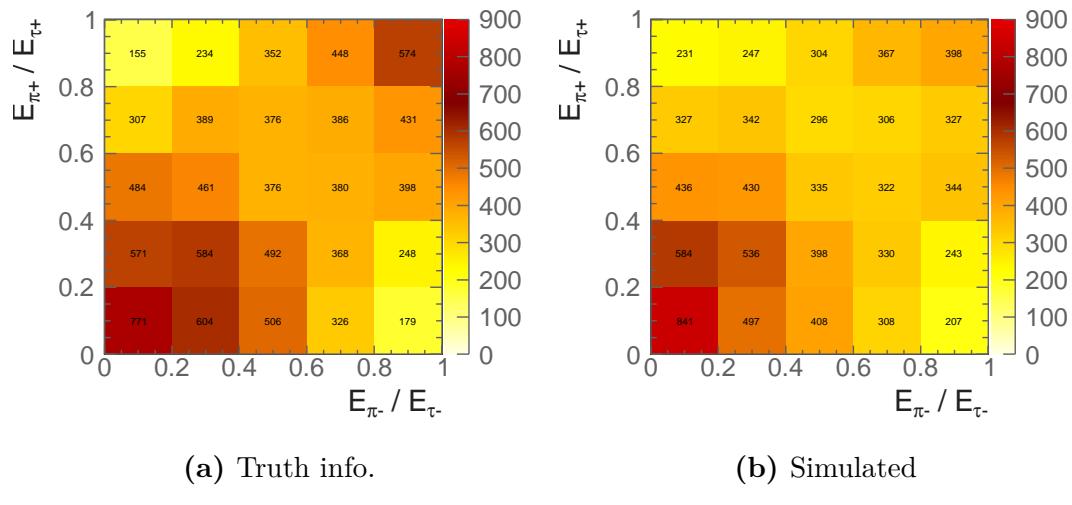
### 6.10.4 Multivariate analysis

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

### 6.10.5 Result

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

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



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



# Chapter 7

## Double Higgs Boson Production Analysis

*“The supreme art of war is to subdue the enemy without fighting.”*

— Sun Tzu, 544 BC-496 BC

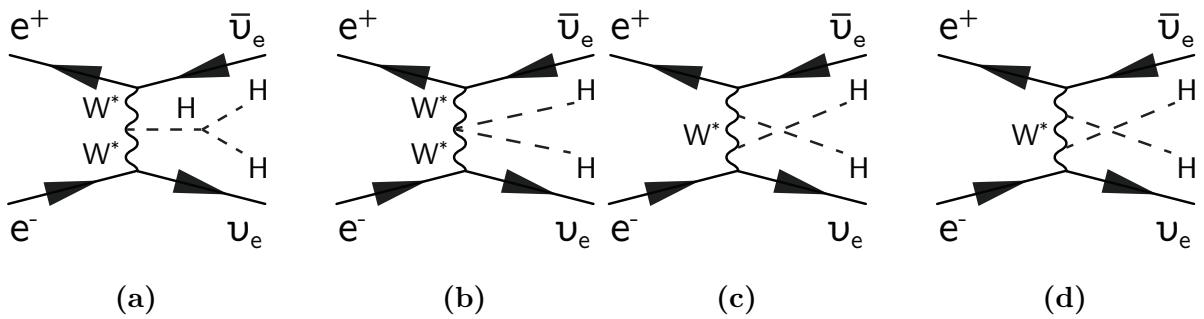
Since the Higgs was discovered at the LHC in 2012 [1, 18], for the particle physicists, it is crucial to understand the interaction between the Higgs and other particles. Another issue to investigate is whether the discovered Higgs is a Standard Model Higgs. A number of Higgs theories beyond the Standard Model may be tested via the double Higgs production in an electron-positron collider. Studying the double Higgs production allow the measurement of the Higgs trilinear self coupling,  $g_{\text{HHH}}$ , and the quartic coupling,  $g_{\text{WWHH}}$ . Monte Carlo studies have shown that the precision of the  $g_{\text{HHH}}$  reached by a multi-TeV linear collider, such as the Compact Linear Collider (CLIC), is superior to the LHC and the HL-LHC [13].

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

In this chapter, a full CLIC\_ILD detector simulation study has been performed for the double Higgs production channel,  $e^+e^- \rightarrow HH\nu_e\bar{\nu}_e$ , via  $W^+W^-$  fusion. Event generation and simulation will be discussed first. An overview of the analysis, including lepton finding and jet reconstruction, is presented, followed by an optimised multivariate analysis to distinguish signal from background processes. The optimised event selection is used to derive an estimate of the uncertainty on  $g_{HHH}$  and  $g_{WWHH}$  measurements at the CLIC. The results of this analysis have been published in [15].

## 7.1 Analysis Straggly Overview

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



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

Double Higgs production can be also produced via  $e^+e^- \rightarrow ZHH$ , where  $Z$  decays to  $\nu\bar{\nu}$ . This  $ZHH$  channel has also been used for study at future  $e^+e^-$  colliders, for example, the ILC at  $\sqrt{s} = 500$  GeV [21]. However, for the relevant CLIC energies of  $\sqrt{s} = 1.4$  TeV and 3 TeV, its contribution to the  $HH\nu\bar{\nu}$  final state is small compared to that of the  $W^+W^-$  fusion, and can be neglected.

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

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

The process,  $\text{HH}\nu_e\bar{\nu}_e \rightarrow b\bar{b}W^+W^-\nu_e\bar{\nu}_e \rightarrow b\bar{b}qqqq\nu_e\bar{\nu}_e$ , results in a six quark final state with missing momentum. The high number of quarks requires an efficient jet reconstruction and a jet pairing algorithm to select the signal events. The two  $b$  quarks in the final state can be identified statistically with  $b$  jet tagging.

A fast simulation study of the double Higgs production has already been performed using the CLIC\_ILD detector model for  $\sqrt{s} = 1.4 \text{ TeV}$  and  $3 \text{ TeV}$  [20]. In this chapter a full CLIC\_ILD detector simulation study is presented. Firstly, suitable signal and background channels are identified. In order to select the signal events, events with identified isolated lepton are vetoed. Vertex information is used to identify  $b$  quark jets. The PFOs in an event are then clustered into jets depending on the number of quarks in the final states. The jets are then used as inputs for pre-selection and multivariate analysis. The event reconstruction was performed using the Marlin framework and reconstruction package in iLCSoft v01-16. More details on the reconstruction software can be found in chapter 4.

## 7.2 Monte Carlo sample generation

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

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

and “quasi-real” photons are generated separately. For the “quasi-real” photon initiated processes, the Equivalent Photon Approximation (EPA) has been used [73].

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

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

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

The cross section of the signal,  $HH \rightarrow b\bar{b}W^+W^-$ , is scaled according to values listed in [74], as the default Higgs branching ratios in the generator software are less accurate than the values listed in [74].

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

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

## 7.3 Lepton identification

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

Channel	$\sigma(\sqrt{s} = 1.4 \text{ TeV}) / \text{fb}$
$e^+e^- \rightarrow HH\nu_e\bar{\nu}_e$	0.149
$e^+e^- \rightarrow HH\nu_e\bar{\nu}_e \rightarrow b\bar{b}W^+W^-$ , hadronic	0.018
$e^+e^- \rightarrow HH\nu_e\bar{\nu}_e \rightarrow b\bar{b}b\bar{b}\nu_e\bar{\nu}_e$	0.047
$e^+e^- \rightarrow HH \rightarrow \text{others}$	0.085
$e^+e^- \rightarrow q\bar{q}H\nu\bar{\nu}$	0.86
$e^+e^- \rightarrow c\bar{c}H\nu\bar{\nu}$	0.36
$e^+e^- \rightarrow b\bar{b}H\nu\bar{\nu}$	0.31
$e^+e^- \rightarrow qqqq$	1245.1
$e^+e^- \rightarrow qqqq\ell\ell$	62.1*
$e^+e^- \rightarrow qqqq\ell\nu$	110.4*
$e^+e^- \rightarrow qqqq\nu\bar{\nu}$	23.2*
$e^+e^- \rightarrow qq$	4009.5
$e^+e^- \rightarrow qq\ell\nu$	4309.7
$e^+e^- \rightarrow qq\ell\ell$	2725.8
$e^+e^- \rightarrow qq\nu\nu$	787.7
$e^\pm\gamma(\text{BS}) \rightarrow e^\pm qqqq$	2317
$e^\pm\gamma(\text{EPA}) \rightarrow e^\pm qqqq$	574
$e^\pm\gamma(\text{BS}) \rightarrow \nu qqqq$	159.1†
$e^\pm\gamma(\text{EPA}) \rightarrow \nu qqqq$	34.7†
$e^\pm\gamma(\text{BS}) \rightarrow qqH\nu$	31.5*
$e^\pm\gamma(\text{EPA}) \rightarrow qqH\nu$	6.78*
$\gamma(\text{BS})\gamma(\text{BS}) \rightarrow qqqq$	21406.2*
$\gamma(\text{BS})\gamma(\text{EPA}) \rightarrow qqqq$	4018.7*
$\gamma(\text{EPA})\gamma(\text{BS}) \rightarrow qqqq$	4034.8*
$\gamma(\text{EPA})\gamma(\text{EPA}) \rightarrow qqqq$	753.0*

**Table 7.1:** List of signal and background samples used in the double Higgs analysis with the corresponding cross sections at  $\sqrt{s} = 1.4 \text{ TeV}$ .  $q$  can be u, d, s, b or t. Unless specified,  $q$ ,  $\ell$  and  $\nu$  represent either particles or the corresponding anti-particles.  $\gamma$  (BS) represents a real photon from beamstrahlung (BS).  $\gamma$  (EPA) represents a “quasi-real” photon, simulated with the Equivalent Photon Approximation. For processes labeled with \* and †, events are generated with the invariant mass of the total momenta of all quarks above 50 and 120 GeV, respectively.

step in the event selection. Primary leptons deposit energies in the tracking detector. The start of the energy deposition is typically very close to the interaction point. The primary leptons often have energies above 10 GeV and are isolated from other particles. Whilst electrons and muons are stable enough to deposit energies in the calorimeters, tau leptons are very short lived. With a typical decay lifetime of 290 fs [2], it decays before reaching the vertex detector. Therefore, only decay products of the tau leptons can be reconstructed.

### 7.3.1 Electron and muon identification

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

#### IsolatedLeptonFinder

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

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

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

**Table 7.2:** Optimised parameters of ISOLATEDLEPTONFINDER

## IsolatedLeptonIdentifier

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

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

Table 7.3 lists the selection cuts for ISOLATEDLEPTONIDENTIFIER. Same variables used in the ISOLATEDLEPTONFINDER and the ISOLATEDLEPTONIDENTIFIER are defined in the same way.  $p_T$  is the transverse momentum in GeV.  $E_{cone1}$  and  $E_{cone2}$  are the total energy of PFOs within a cone of an opening angle of  $\cos^{-1}(0.995)$  and  $\cos^{-1}(0.99)$  respectively around the lepton in GeV.

### 7.3.2 Tau lepton identification

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

The existing TAUFINDER [75] reconstruction package has been optimised. In addition, a package, ISOLATEDTAUIDENTIFIER, is developed to provide additional tau lepton identification.

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

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

## TauFinder

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

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

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

TAUFINDER	Selection
Veto $\gamma\gamma \rightarrow \text{hadrons}$	$p_T < 1 \text{ GeV},  \cos(\theta_Z)  > 1.1 \text{ rad}$
Seed particle	$p_T > 10$
Tau candidate cone opening angle	0.03 rad
Tau candidate rejection	$N_{X^+} > 3, N_{tau} > 10, m_{tau} > 2 \text{ rad}$
Isolation	$E_{cone} < 3 \text{ GeV}$

**Table 7.4:** Optimised parameters of TAUFINDER.

### IsolatedTauIdentifier

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

The iterative particle addition procedure stops when the cone opening angle is larger than a threshold. If multiple temporary tau candidates of the same tau seed pass the selection, the one with smallest opening angle is chosen to form the final tau candidate. To reduce fake tau decay products from  $\gamma\gamma \rightarrow \text{hadrons}$  background, particles with energies less than 1 GeV do not participate.

Table 7.4 lists the optimised parameters for ISOLATEDTAUIDENTIFER. Variables are defined in the same way as those in previous sections:  $\theta_S$  is the opening angle of the search cone in rad;  $cone1$  and  $cone2$  are defined as a cone of an opening angle of  $\cos^{-1}(0.95)$ , and  $\cos^{-1}(0.99)$  respectively around the tau seed;  $r_0$  is the Euclidean distance in mm of the tau seed associated track starting point to the interaction point in  $x$ - $y$ - $z$  three dimensional space.

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

**Table 7.5:** Optimised parameters of ISOLATEDTAUIDENTIFIER

Compared to the TAUFINDER algorithm, the main difference is that the ISOLATED-TAUIDENTIFIER has an iterative approach to build up a tau candidate, which allows a dynamic tau search cone size. The ISOLATEDTAUIDENTIFIER also has smaller cut values on minimum  $p_T$  and invariant mass, but stricter isolation criterions.

### 7.3.3 Very forward electron identification

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

It is challenging to identify electrons in these forward calorimeters. Most particles in these forward calorimeters are from the beam induced background. Nevertheless, sufficiently high energy electrons can be efficiently identified in the BeamCAL and LumiCAL [30].

The reconstruction of electrons in the forward calorimeters uses a parametrisation approach. By parameterising the particle ID efficiency using MC particles, the equivalent performance to the full detector simulation approach can be achieved [76, 77].

For the BeamCAL, an existing electron tagging algorithm was developed using  $\sqrt{s} = 3 \text{ TeV}$  collision environment by comparing the simulated electron and background energy distributions [76]. An indicative performance plot of 500 GeV electron tagging efficiency as a function of polar angle is shown in figure 7.2a. An electron is tagged if the energy is significantly larger than the expected background energy distributions integrated over 40 bunch crossing.

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

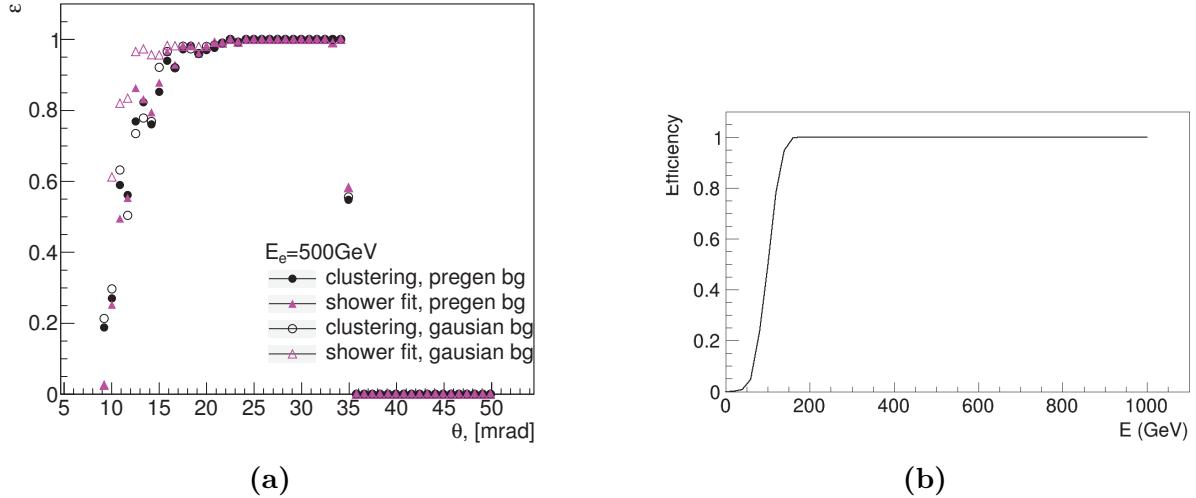
The input of the BeamCAL electron tagging algorithm is the four momenta of the MC electron. Since the algorithm assumes collisions at  $\sqrt{s} = 3 \text{ TeV}$ , for the  $\sqrt{s} = 1.4 \text{ TeV}$  user case, the momenta of the MC electron is scaled down by a factor of  $\frac{3}{1.4}$ .

For the LumiCAL, figure 7.2 shows the LumiCAL electron tagging efficiency as a function of the electron energy for polar angle  $\theta = 50 \text{ mrad}$  where events are overlaid with background energy deposition integrated over 100 bunch crossings [77]. In this analysis, the LumiCAL electron tagging efficiency,  $\varepsilon$ , is parameterised as

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

where  $E$  is the energy of the electron and  $\text{erf}$  is the error function. For each MC electron in the LumiCAL, a random number between 0 and 1 is generated. If the random number is less than  $\varepsilon$  the MC electron is tagged.

Due to lack of tracking ability in the forward region, energy depositions from electrons or photons can not be differentiated in forward calorimeters. Therefore, both photons and electrons are tagged by the algorithms above.



**Figure 7.2:** Figure 7.2a shows 500 GeV electron tagging efficiency in the BeamCAL as a function of polar angle, with different methods to model backgrounds: pregenerated and Gaussian, and two methods to identify electrons: clustering algorithm and shower fitting algorithm, taken from [76]. Figure 7.2b shows the electron tagging efficiency in the LumiCAL as a function of the electron energy, for polar angle  $\theta = 50\text{ mrad}$ , taken from [77].

### 7.3.4 Lepton identification performance

The performances of the different lepton finding processors for signal events and the selected background processes are shown in table 7.6. Numbers in the table represent the fractions of events where no leptons are identified by the individual lepton finder. ISOLATEDLEPTONIDENTIFIER and ISOLATEDTAUIDENTIFIER 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  $e^+e^- \rightarrow qqqq\ell\nu$  events survive after rejecting events where leptons are identified.

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

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

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

**Table 7.6:** The performances of lepton finders on the signal events and selected background events at  $\sqrt{s} = 1.4$  TeV. Numbers represent the fractions of events where no leptons are identified by the individual lepton finder.

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

**Table 7.7:** The performances of lepton finders on the signal events and selected background events at  $\sqrt{s} = 3$  TeV. Numbers represent the fractions of events where no leptons are identified by the individual lepton finder.

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

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

## 7.4 Jet reconstruction

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

### 7.4.1 Jet reconstruction optimisation

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

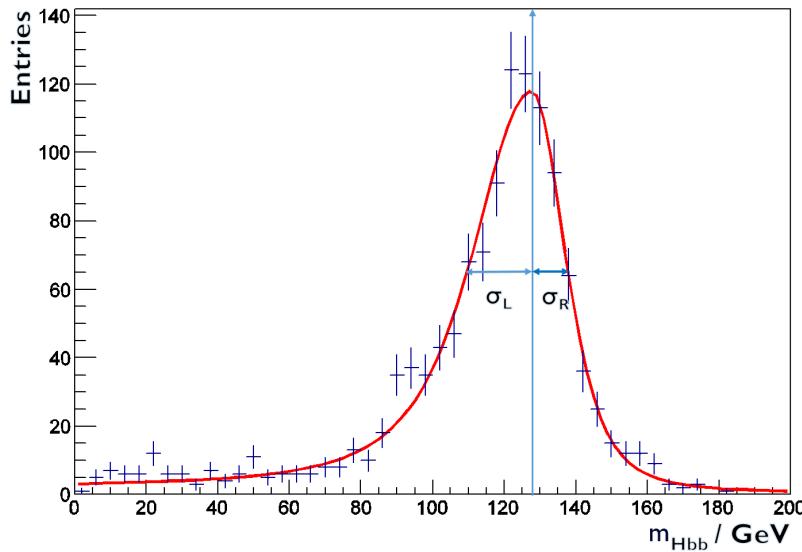
The value of the  $R$  parameter and the PFO collection are chosen to optimise the invariant mass and mass resolution of H and W. To choose the optimal parameters,  $\text{HH} \rightarrow b\bar{b}W^+W^- \rightarrow b\bar{b}qqqq$  events are processed through  $k_t$  jet algorithm in the 6-jet exclusive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

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

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

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

where:  $\mu$  is the fitted mass peak position;  $\sigma_L$  and  $\sigma_R$  allow an asymmetrical width of the distribution;  $\alpha_L$  and  $\alpha_R$  fit the tail of the distribution; and  $A$  is the normalisation factor.



**Figure 7.3:** A typical example of  $m_{H_{bb}}$  mass distribution with the superimposed fitting function in red. The arrow shows the fitted mean peak position.

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

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

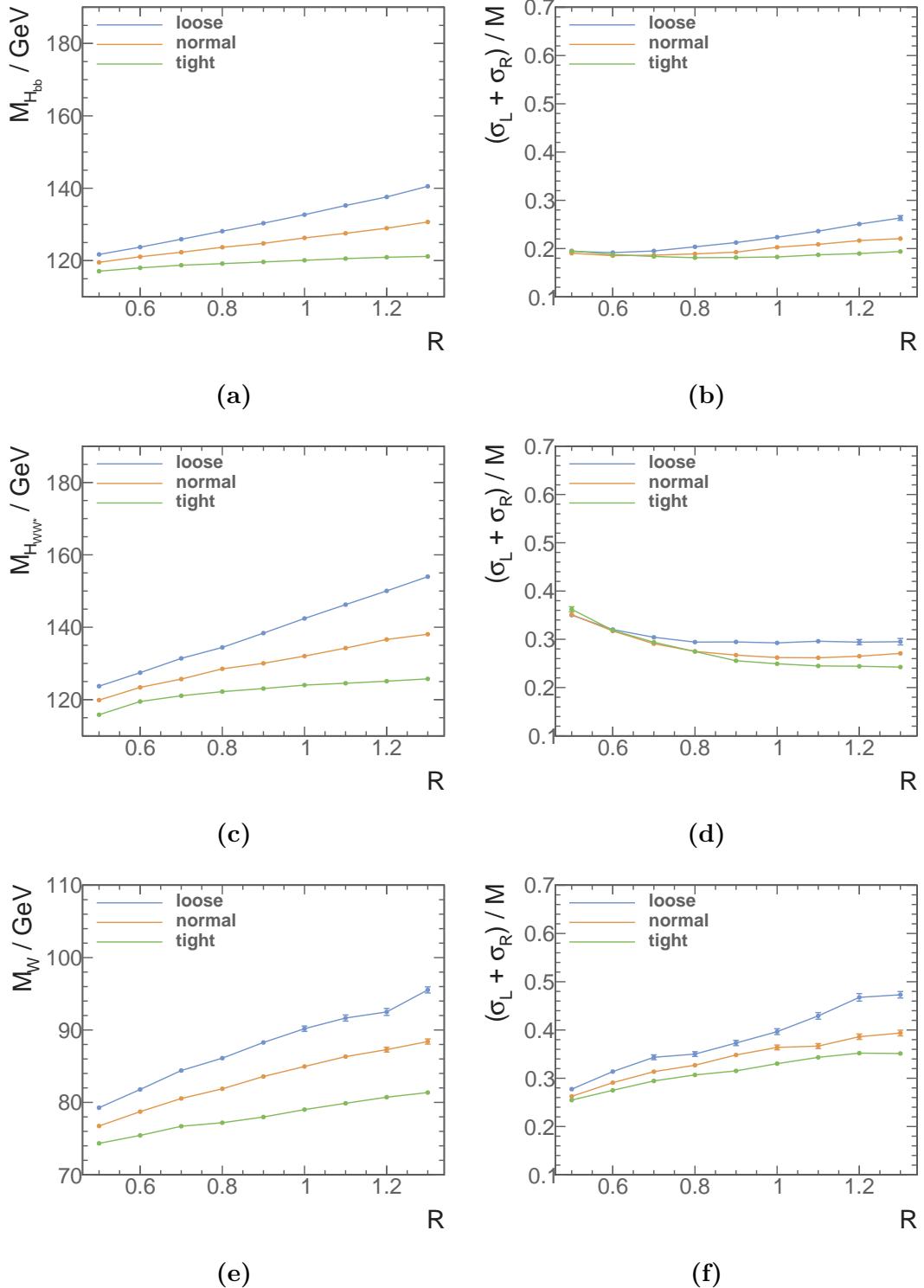
The choice of PFO collection impacts number of PFOs in the event. The loose PFO selection has the most PFOs in the event and, therefore, the largest invariant mass and worst mass resolution.

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

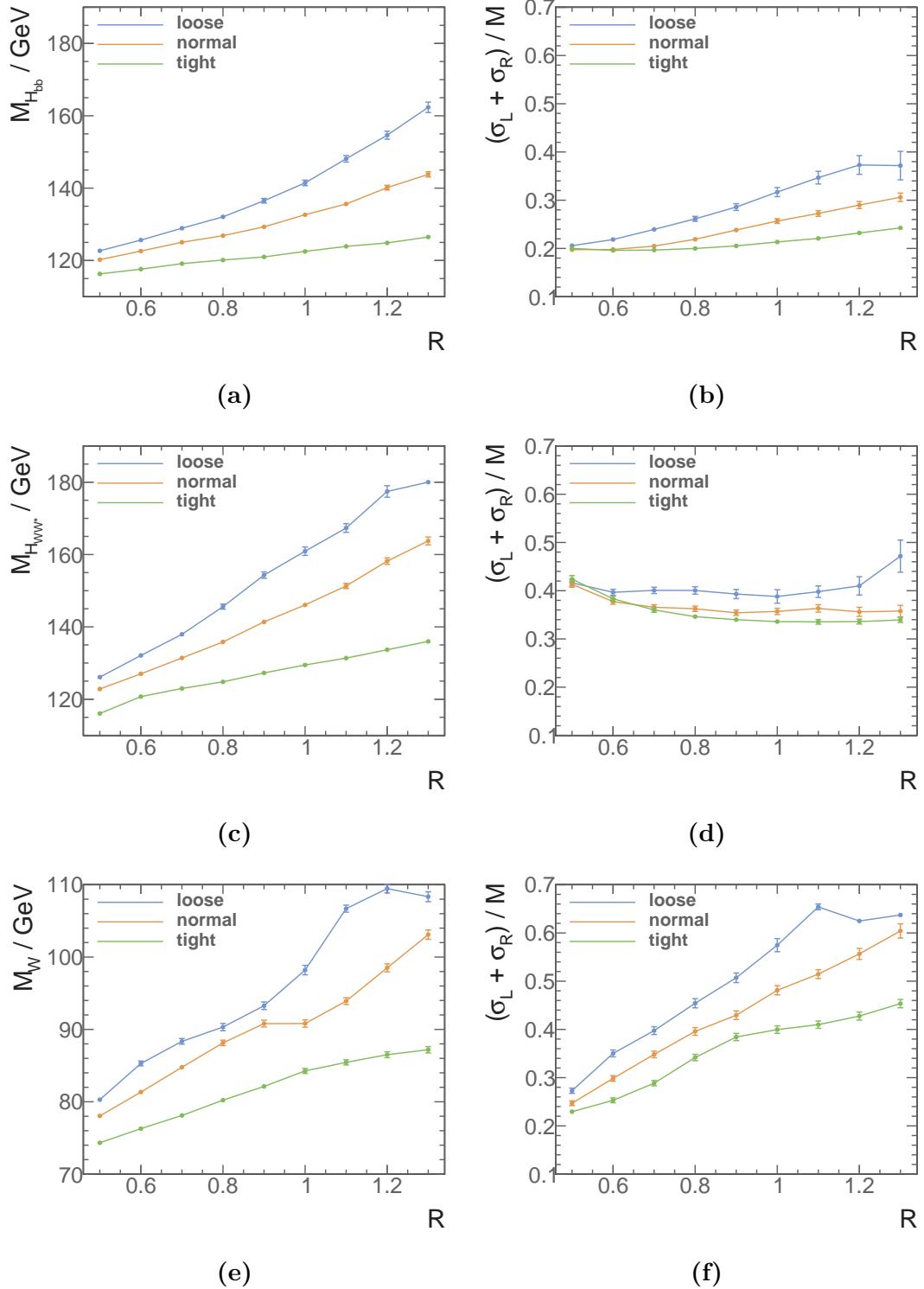
Fitted jet parameters	$\sqrt{s} = 1.4 \text{ TeV}$	$\sqrt{s} = 3 \text{ TeV}$
$\mu_{H_{bb}}$	$122.3 \pm 0.2$	$119.1 \pm 0.3$
$\sigma_{L,H_{bb}}$	$15.2 \pm 0.2$	$15.0 \pm 0.3$
$\sigma_{R,H_{bb}}$	$7.55 \pm 0.16$	$8.4 \pm 0.2$
$\mu_{H_{WW^*}}$	$125.7 \pm 0.2$	$123.0 \pm 0.3$
$\sigma_{L,H_{WW^*}}$	$29.4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

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

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



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



**Figure 7.5:** Distributions of a) fitted mass peak positions of  $H_{bb}$ , b) relative mass peak width of  $H_{bb}$ , c) fitted mass peak positions of  $H_{WW^*}$ , d) relative mass peak width of  $H_{WW^*}$ , e) fitted mass peak positions of  $W$ , and f) relative mass peak width 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  $\sqrt{s} = 3$  TeV.

## 7.5 Jet flavour tagging

As the signal channel,  $\text{HH} \rightarrow b\bar{b}W^+W^- \rightarrow b\bar{b}qqqq$ , has two b quarks in the final state, identifying jets originated from b quarks is an important part of the event selection. To establish the likelihood of a b jet (b-jet tag value), LCFIPlus [55] software package is used.

The flavour tagging processor, LCFIPlus [55] is based on the LCFIVertex package [56], which was used in the simulation studies for the ILC Letter of Intent [57, 58] and the CLIC Concept Design Report [20]. LCFIPlus is modular and can be used in any order. However here it will be described in the order that is used in this analysis.

The inputs are PFOs, after jet clustering using  $R = 0.7$  and the selected PFO collection. The vertex finding algorithms in the LCFIPlus perform vertex fitting and identify primary and secondary vertices. There is a “V0” particle rejection step. A “V0” particle is a neutral particle that decays into pairs of charged particles. The topology of the “V0” particles can be similar to the decay of b or c quarks. Hence it is important to remove the “V0” particles to improve the b-quark and c-quark flavour tagging (see section 4.4.1 for a similar V0 rejection).

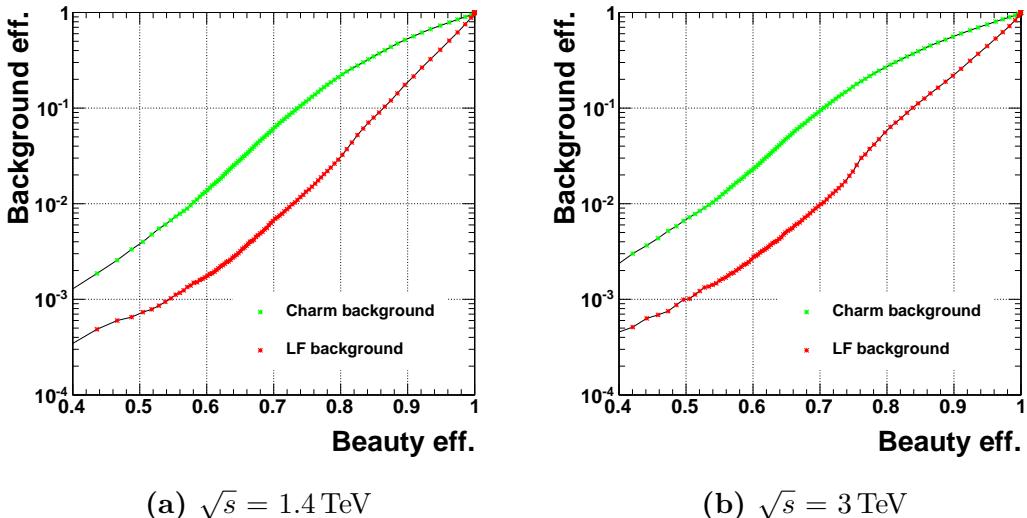
Once the primary and secondary vertices are found, PFOs are re-clustered into jets. This jet re-clustering scheme ensures that the secondary vertices and the muons which is identified from semi-leptonic decay of the quarks, fall into the same jet. Therefore, the topology of the jet is consistent with the hadronic decay of heavy quarks. The jet re-clustering scheme are Durham and Durham modified algorithms (see section 4.6.2).

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

The last step is to gather the information about vertices and jets, and deploy a multivariate analysis. The multivariate classifier used, Boosted Decision Tree, is implemented in the TMVA software package [59]. A number of flavour sensitive variables are calculated. The jets are divided into four subsets: jets with zero, one, two properly reconstructed vertices, or a single-track pseudo-vertex. For each subset, a jet can either be classified to a b jet, a c jet, or a light flavour quark jet (u, d or s). The multiclass classifier’s response is normalised across different subsets.

The events to train the multiclass classifier are  $e^+e^- \rightarrow Z\bar{\nu}\nu$  event at  $\sqrt{s} = 1.4$  TeV, where Z decays to  $b\bar{b}$ ,  $c\bar{c}$ , or  $u\bar{u}/d\bar{d}/s\bar{s}$ . The selection efficiency of b jets and c jets with the training samples is shown in figure 7.6b. The normalised distribution of the highest b-jet tag value for the signal events is shown in figure 7.7.

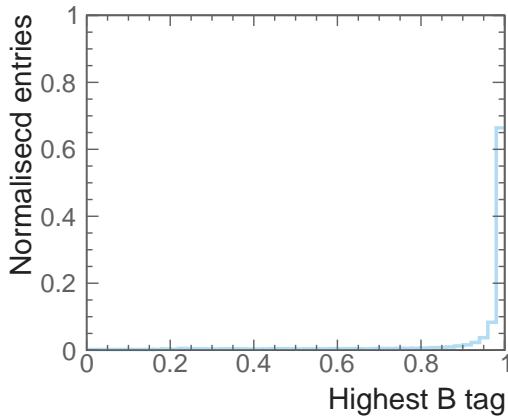
For the  $\sqrt{s} = 3$  TeV, the flavour tagging processor is re-trained in the same way as in the  $\sqrt{s} = 1.4$  TeV analysis. The training events are from the same channel but at  $\sqrt{s} = 3$  TeV. The performance of the flavour tagging with training samples is shown in figure 7.6a. Compared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



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

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

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



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

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

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

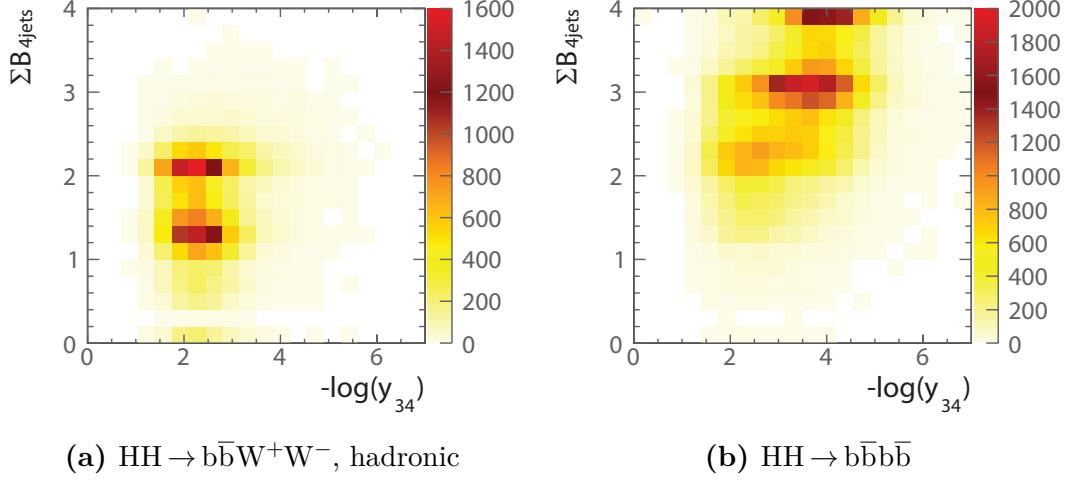
Figure 7.8 shows the sum of b-jet tag values, when the event is clustered into four jets, as a function of  $-\log(y_{34})$  for the hadronic  $W^+W^-$  decay in  $HH \rightarrow b\bar{b}W^+W^-$  and  $HH \rightarrow b\bar{b}b\bar{b}$ .  $y$  parameter is a measure of number of jets. As expected, two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

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

$$\varepsilon = \frac{N_{HH \rightarrow b\bar{b}W^+W^-, \text{ hadronic}} \in S}{N_{HH \rightarrow b\bar{b}W^+W^-, \text{ hadronic}}} \times \frac{N_{HH \rightarrow b\bar{b}b\bar{b}} \in \neg S}{N_{HH \rightarrow b\bar{b}b\bar{b}}}, \quad (7.4)$$

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

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



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

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

## 7.6 Jet pairing

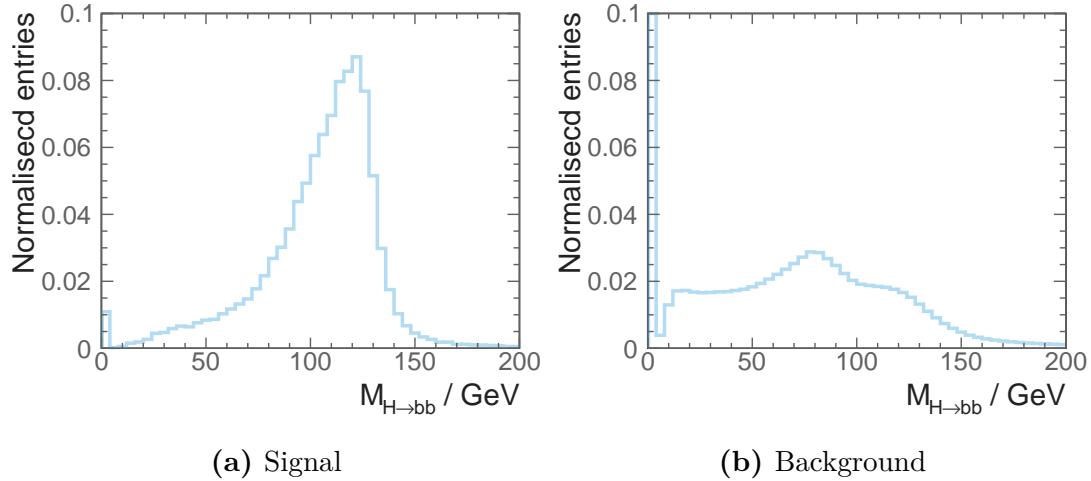
Having optimised the jet reconstruction, and obtained the six jets from the jet re-clustering step in the LCFIPlus processor, the next step is to group jets according to signal event topology. Jets are paired up such that there are two jets for  $H \rightarrow b\bar{b}$ , two jets for hadronic decay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

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

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

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

where the indices  $i$  to  $l$  indicate the one of the six jets,  $\mu$  and  $\sigma$  are the fitted invariant mass and the fitted width from table 7.8. The asymmetrical structure of the fitting function is reflected in the jet pairing  $\chi^2$ . The jet pairing is only considered when at least one of the jets associated to the  $H_{bb}$  decay with a b-jet tag  $> 0.2$ . Of these combinations of jets, the jet pairing giving smallest  $\chi^2$  is selected. The normalised distribution of  $m_{H_{bb}}$  after jet pairing, for the signal channel,  $HH \rightarrow b\bar{b}W^+W^-$  and the sum of all background channels can be seen in figure 7.9. For the signal channel, the distribution peaks around the expected mass of  $m_{H_{bb}}$ . It can also be seen from the figures that around 1% of the signal events have no solutions for the jet pairing, as no jet has a b-jet tag  $> 0.2$ . The full list of fraction of events after jet pairing selection can be found in table 7.9.



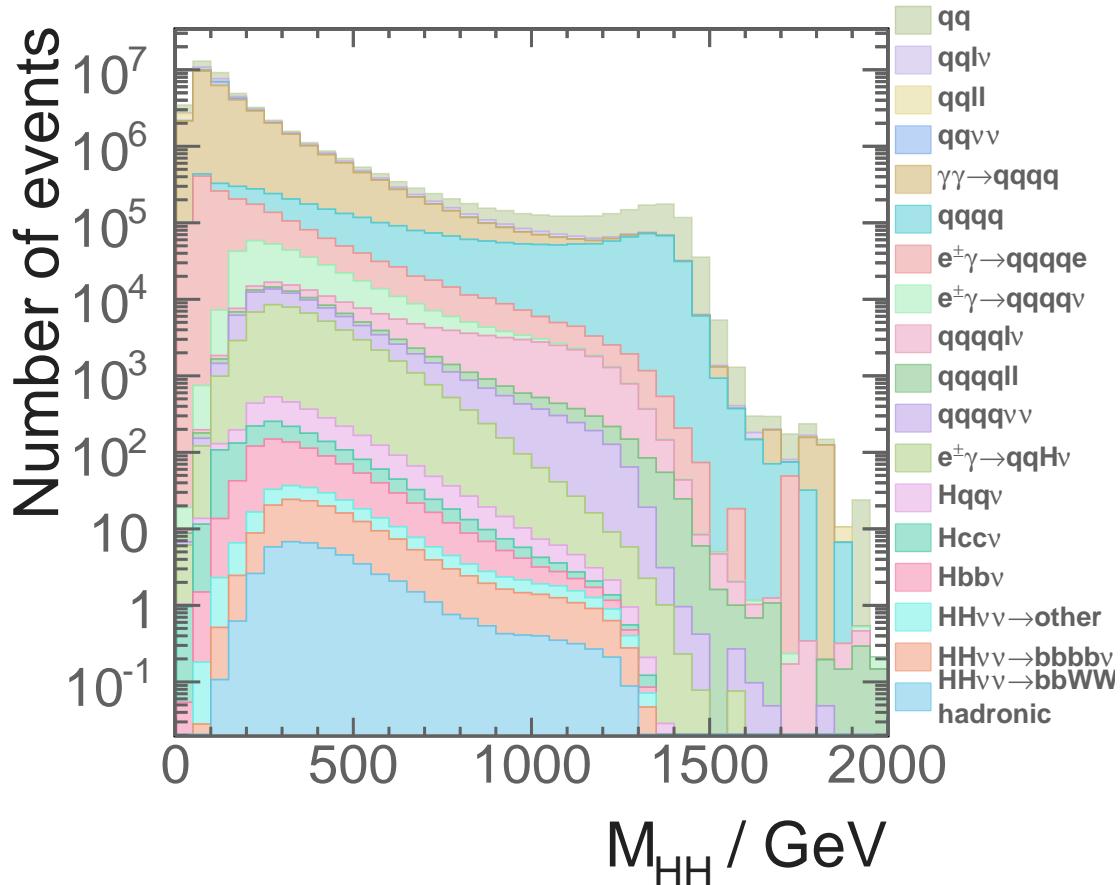
**Figure 7.9:** The normalised distribution of  $m_{H_{bb}}$  after jet pairing, for a) signal channel, hadronic  $W^+W^-$  decay of  $HH \rightarrow b\bar{b}W^+W^-$ , b) sum of all background channels. All plots are shown for  $\sqrt{s} = 1.4 \text{ TeV}$ .

## 7.7 Pre-selection

After the association of jets to candidate bosons are made under hypothesis that an event is signal, kinematic and topological variables can be calculated. A set of pre-selection

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

**Table 7.9:** Pre-selection cut efficiencies for signal and background samples at  $\sqrt{s} = 1.4$  TeV, assuming an integrated luminosity of  $1500 fb^{-1}$ . The selection efficiencies are presented in a “flow” fashion. Every selection cut contains all the cuts to the left of it.



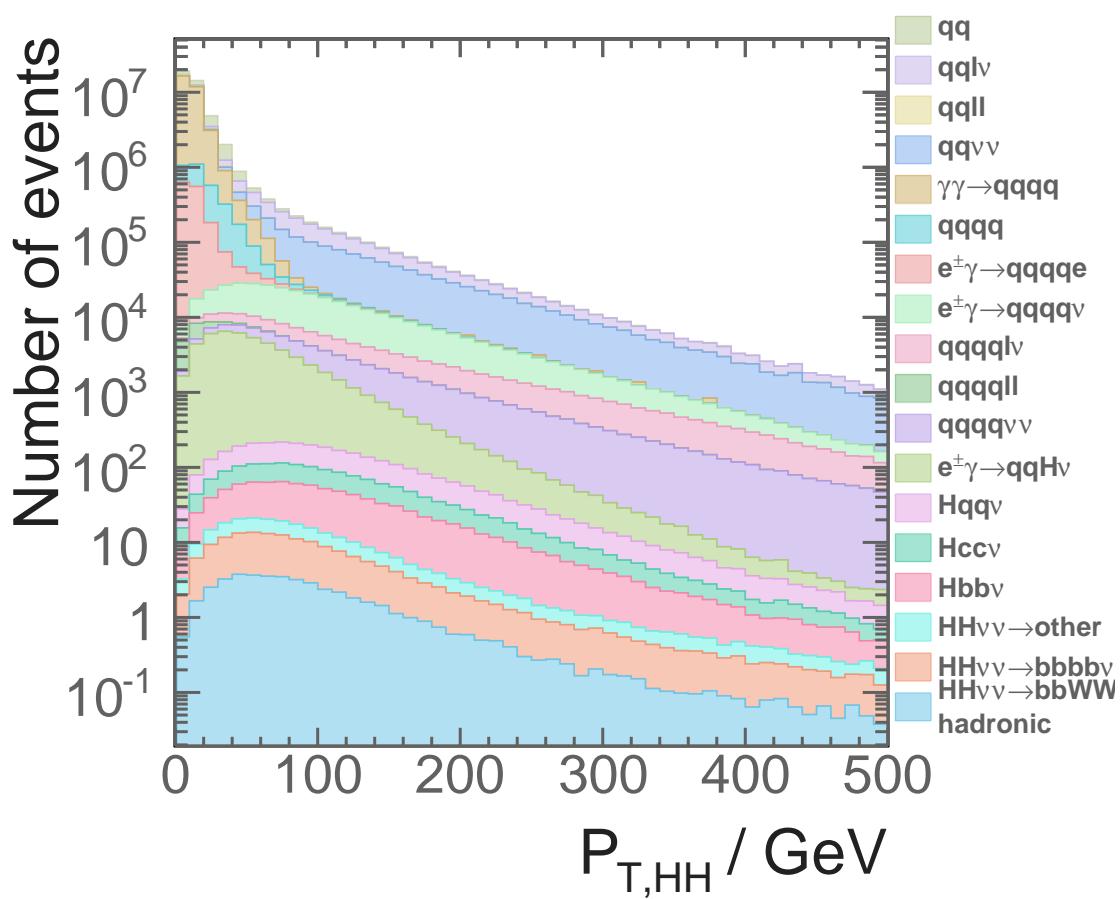
**Figure 7.10:** Distributions of the invariant mass of the two Higgs system for  $\sqrt{s} = 1.4 \text{ TeV}$ , assuming an intergraded luminosity of  $1500 \text{ fb}^{-1}$ .

cuts are placed to discard the phase space dominated by background events. Cuts on  $p_T$ , b-jet tag, and invariant mass of the double Higgs system are used.

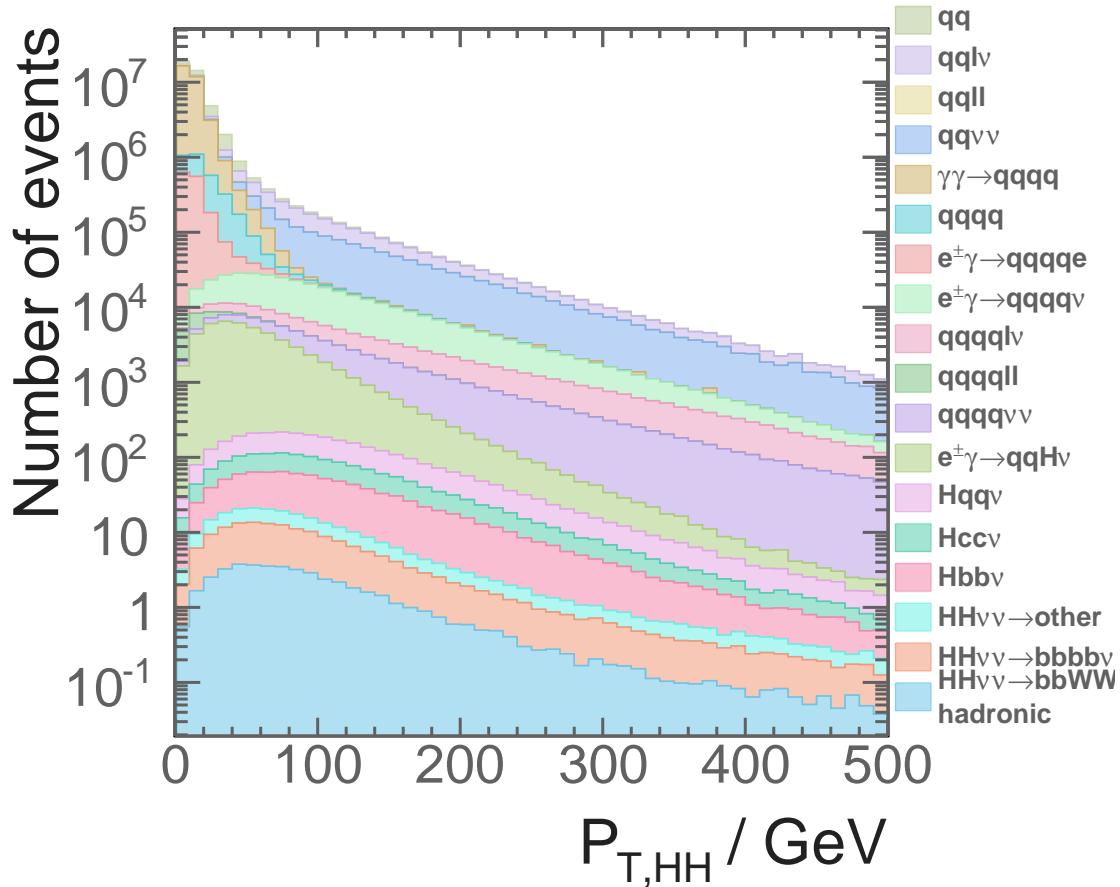
Since both Higgs bosons are on mass shell, the invariant mass of the double Higgs system is large. Consequently, a cut on  $m_{\text{HH}} > 150 \text{ GeV}$ , as shown in Figure 7.10, removes a small amount of signal events, but discard lots of background events, especially  $\gamma\gamma \rightarrow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

Many background events do not have b-quark jets in the final state. Therefore, by requiring the second highest b-jet tag value greater than 0.2, as shown in Figure 7.12, background events with no b-jets in final states are removed.

The signal final states have neutrinos and hence missing momentum in the events. Therefore, the transverse momentum of the two Higgs system is non zero. A cut of



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



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

$p_T > 30 \text{ GeV}$ , as shown in figure 7.12, is extremely effective against background channels with no neutrinos in the final state.

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

## 7.8 MVA variables

Having extracting information about leptons, b-jets, and jet pairing, a number of variables are used to differentiate the signal and the background events. These variables are the

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

**Table 7.10:** List of signal and background samples after pre-selection cuts at  $\sqrt{s} = 1.4\text{ TeV}$ . The selection efficiencies are presented in a “flow” fashion. Every selection cut contains all the cuts to the left of it and all the cuts in table 7.9 .

basis of the subsequent MVA event selection. The variables used are listed in table 7.11. The distributions of the four most power discriminators are show in figure 7.13.

### 7.8.1 Invariant mass variables

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

After the jet pairing under the hypothesis of the signal events, the distributions of the invariant mass of the physical bosons of the signal events have peaks around the expect masses, where the distributions of the background events do not have such peak structure. Shown in the figure 7.13a (also similar plots can be found in jet paring section), the distributions of the invariant mass of the  $H_{bb}$  is different to the distributions of the background events. Similarly, the distributions of the invariant mass of the  $H_{WW^*}$ , shown in figure 7.13b have different peak position to the distributions of the background events. The invariant mass of the double Higgs system is large due to the presence of two on-mass-shell Higgs bosons in the signal final state.

### 7.8.2 Energy and momentum variables

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

For the off-mass-shell  $W$ , its energy variable instead of the invariant mass variable is used, as invariant mass distribution of  $W^*$  does not have a resonance peak. The energy of the missing momenta is powerful against background events with no neutrinos in the final state. The missing momenta is calculated by assuming the collision at  $\sqrt{s}$  and a beam crossing angle of 20 mrad. Other momentum variables correspond to the same physical bosons or the double Higgs system used for the invariant mass variables, for the same reason that the distributions of these momentum variables are different for the signal events and the background events.

### 7.8.3 Lab-frame angle variables

Four lab-frame angle variables are in the MVA event selection: the pseudorapidity of the missing momenta ( $\eta_{mis}$ ), 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 angle is because that there are more phase space for small angles for the pseudorapidity. The pseudorapidity of the missing momenta is defined as

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

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

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

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

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

### 7.8.4 Boosted-frame angle variables

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

These variables are some of the most powerful variables. For example,  $\cos(\theta_{H_{bb}}^*)$  for the signal events has a uniform distribution, shown in figure 7.13d, as it is equally likely

to decay in any open angle. For the background events, by pairing jets under the signal event hypothesis,  $\cos(\theta_{H_{bb}}^*)$  does not have a flat distribution. The  $\cos(\theta_{H_{bb}}^*)$  for the background events peaks at 1.

### 7.8.5 Event shape variables

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

The sphericity,  $\mathbf{S}$ , is a measurement of the spherically symmetry of the event, which will be different for the signal and background events. The sphericity is derived from the sphericity tensor [61]. The sphericity tensor is defined as

$$\mathbf{S}^{\alpha\beta} = \frac{\sum_i p_i^\alpha p_i^\beta}{\sum_i |\vec{p}_i|^2}, \quad (7.9)$$

where  $\vec{p}_i$  is the momentum vector of the particle  $i$ . Summation is over all particles in the event.  $\alpha$  and  $\beta$  refer to the x, y, z coordinate axis. Eigenvalues of tensor  $\mathbf{S}$  can be found, or in this case diagonalisation of the matrix  $\mathbf{S}$ , denoted with  $\lambda_1, \lambda_2, \lambda_3$ . The normalisation condition requires  $\lambda_1 \geq \lambda_2 \geq \lambda_3$  and  $\lambda_1 + \lambda_2 + \lambda_3 = 1$ . Sphericity,  $S$ , is defined in terms of  $\lambda$ ,

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

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

The  $y$  parameter is a measure of the number of jets in an event. It describes the transition of the exclusive jet algorithm going from  $N$  clustered jets to  $N+1$  clustered jets. For example,  $y_{23}$  would be the  $d_{cut}$  value for a exclusive jet algorithm, above which the jet algorithm returns 2 jets, below which the jet algorithm returns 3 jets. Numerically the  $y$  parameter is often much smaller than one. A typically way to convert the small number to a human acceptable range is to take the negative logarithm of the number.

Category	Variable
Invariant mass	$m_{H_{bb}}, m_{H_{WW^*}}, m_W, m_{HH}$
Energy and momentum	$E_{W^*}, E_{mis}, p_{TH_{bb}}, p_{TH_{WW^*}}, p_{TW}, p_{THH}$
Lab-frame angles	$\eta_{mis}, A_{H_{bb}}, A_W, A_{HH}$
Boosted-frame angles	$\cos(\theta_{H_{bb}}^*), \cos(\theta_{H_{WW^*}}^*), \cos(\theta_W^*), \cos(\theta_{W^*}^*), \cos(\theta_{HH}^*)$
Event shape	$ \mathbf{S} , -\ln(y_{23}), -\ln(y_{34}), -\ln(y_{45}), -\ln(y_{56})$
b and c tag	$B_{1,H_{bb}}, B_{2,H_{bb}}, B_{1,W}, B_{1,W^*}, C_{1,H_{bb}}, C_{1,W}$
PFOs number	$N_{H_{bb}}, N_{H_{WW^*}}, N_W, N_{W^*}$

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

### 7.8.6 b and c tag variables

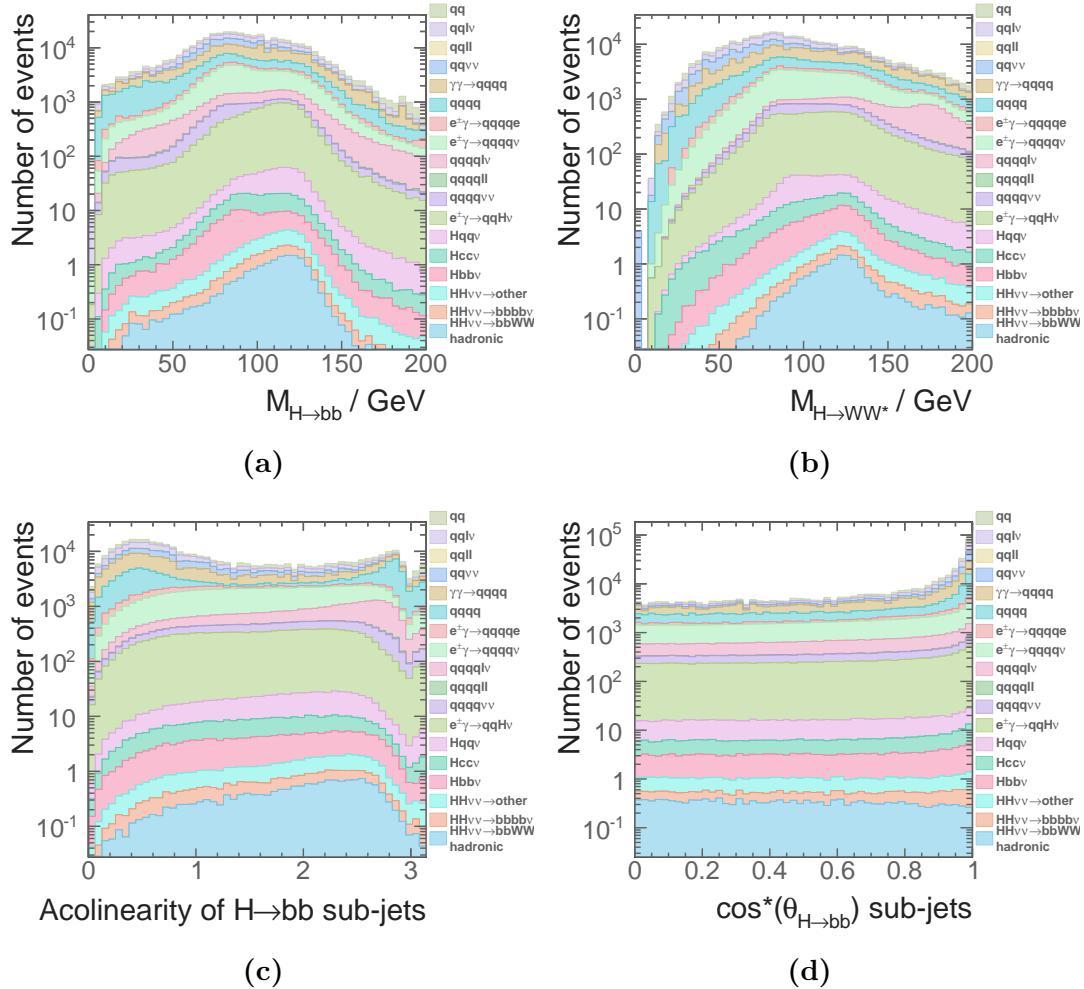
Six b-jet and c-jet tag variables are used in the MVA event selection: the highest b-jet tag value of the two jets associated with  $H_{bb}$  ( $B_{1,H_{bb}}$ ), the lowest b-jet tag value of the two jets associated with  $H_{bb}$  ( $B_{2,H_{bb}}$ ), the highest b-jet tag value of the two jets associated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

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

### 7.8.7 PFOs number variables

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

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



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

### 7.8.8 Cuts to aid the MVA

A set of cuts reduce the range of invariant masses variables in order to increase the effectiveness of the MVA event selection. (See section 4.7.2). Occasional large values of the invariant masses variables skew the distribution of these variables. Therefore by limiting the range of the variables, the MVA classifier could focus on the phase spaces with high 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  $< 1400 \text{ GeV}$ .

## 7.9 Multivariate analysis

After gathering information and applying pre-selection cuts, signal events are selected using the multivariate analysis (MVA) with Boosted Decision Tree classifier (BDT), as implemented in the TMVA [59].. The parameters for boosted decision tree were optimised and checked for overtraining, following the strategy outlined in section 4.7. Half of the events were used for training, and the other half used for testing and classifier optimisation. The optimised parameters are listed in table 7.12.

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

## 7.10 Signal selection results

Number of events passed the MVA event selection at  $\sqrt{s} = 1.4 \text{ TeV}$ , assuming a luminosity of  $1500 fb^{-1}$  are listed in table 7.13 for individual channels. A few background channels have non-zero events after the MVA event selection.  $e^+e^- \rightarrow q\bar{q}H\nu\bar{\nu}$  events are difficult to discard because its topology, one single Higgs plus neutrino, is very similar to the signal event topology. Similarly,  $e^+e^- \rightarrow qqqq\ell\nu$  events can be confused as the signal

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

**Table 7.12:** Optimised parameters for the boosted decision tree classifier. See section 4.7.2 for detailed explanation of variables.

events when the lepton is undetected in the forward region, or the energy of the lepton is too low to be tagged.  $e^+e^- \rightarrow qqqq\nu\bar{\nu}$  events can also have a similar topology to the signal events. Other background channels that are not discarded after the MVA are the electron-photon and photon interactions with the same final states as the channels above.

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

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

The  $e^+e^- \rightarrow HH\nu_e\bar{\nu}_e \rightarrow b\bar{b}W^+W^-\nu\bar{\nu}$  hadronic decay at  $\sqrt{s} = 3$  TeV analysis follows the same strategy as the analysis at  $\sqrt{s} = 1.4$  TeV. A brief discussion of each step and the results are provided and differences are highlighted. Cross sections of used samples are listed in table 7.14.

The pre-selection cuts at  $\sqrt{s} = 3$  TeV are largely the same as ones for  $\sqrt{s} = 1.4$  TeV analysis. The cuts are listed in table 7.16. The reason for using a different b-jet tag cut is because the performance of flavour tagging is worse at high  $\sqrt{s}$ . Figure 7.14b shows the

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

**Table 7.13:** List of signal and background events with selection efficiency and number of events at  $\sqrt{s} = 1.4$  TeV, assuming a luminosity of  $1500 fb^{-1}$ . The number of events ( $N$ ), the selection efficiencies of pre-selection cuts ( $\varepsilon_{presel}$ ), the selection efficiencies of MVA after pre-selection cuts ( $\varepsilon_{MVA}$ ), and the number of events after MVA ( $N_{MVA}$ ) are shown. - represents a number less than 0.01.

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

**Table 7.14:** List of signal and background samples with the corresponding cross sections at  $\sqrt{s} = 3 \text{ TeV}$ .  $q$  can be u, d, s, b or t. Unless specified,  $q$ ,  $\ell$  and  $\nu$  represent particles and their corresponding anti-particles.  $\gamma$  (BS) represents a real photon from beamsstrahlung (BS).  $\gamma$  (EPA) represents a “quasi-real” photon simulated with the Equivalent Photon Approximation. For processes involving Higgs production explicitly, simulated Higgs mass is 126 GeV. Otherwise, Higgs mass is set to 14 TeV. For processes labelled with \*, the generator level cut requires invariant mass of quarks greater than 50 GeV.

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

**Table 7.15:** The extracted fitted parameters of optimal jet reconstructions for tight selected PFO collection with  $R = 0.7$  at  $\sqrt{s} = 3 \text{ TeV}$ .

distribution of the highest b-jet tag, where the cut above 0.7 helps to reduce background events with no b-jet in final states. Figure 7.14a shows the distribution of the invariant mass of the two Higgs system, where the cut above 150 GeV is effective against samples with two quark final states.

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

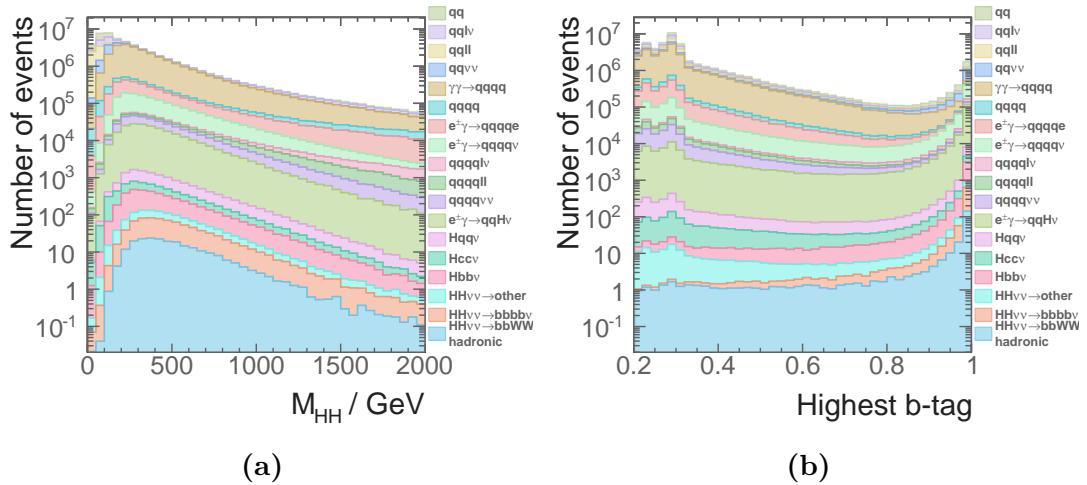
**Table 7.16:** Pre-selection cuts at  $\sqrt{s} = 3 \text{ TeV}$ .

The cuts to aid the MVA at  $\sqrt{s} = 3 \text{ TeV}$  are largely the same as the ones at  $\sqrt{s} = 3 \text{ TeV}$ , apart from the difference on the cut of the invariant mass of HH due to higher  $\sqrt{s}$ . The selection efficiency of the lepton veto and the pre-selection is shown in table 7.17.

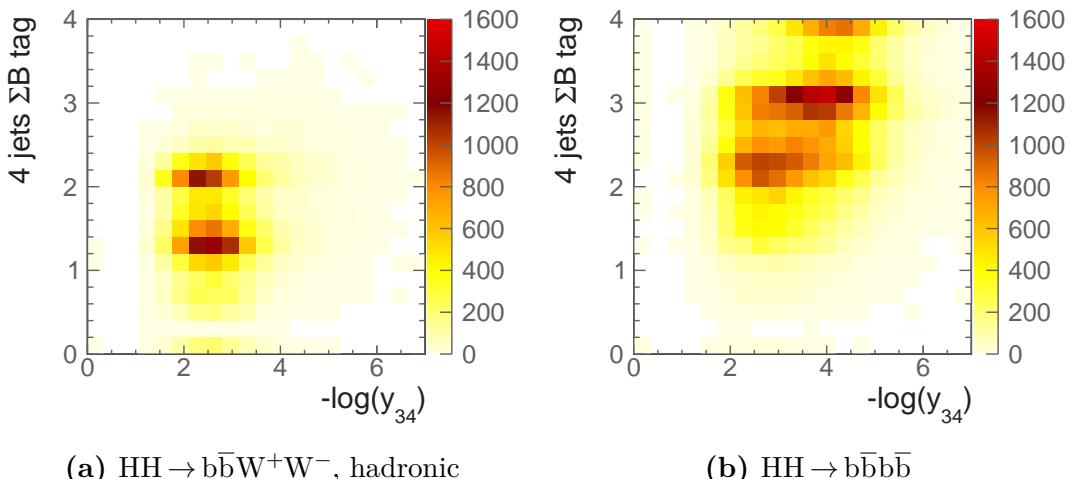
The mutually exclusive cuts divide samples - both signal and background - into two mutually exclusive sets for the parallel analyses of two subchannels;  $HH \rightarrow b\bar{b}W^+W^- \rightarrow b\bar{b}q\bar{q}q\bar{q}$  and  $HH \rightarrow b\bar{b}b\bar{b}$ . The cuts are obtained using the same strategy in section 7.5.1. The cuts and efficiencies are listed in table 7.16. The two dimensional spaces for two sub-channels

Channel / Efficiency $\sqrt{s} = 3 \text{ TeV}$	Expected number of events	Lepton ID and jet pairing	$m_{HH} > 150 \text{ GeV}$	$B_1 > 0.7$
$e^+e^- \rightarrow HH\nu_e\bar{\nu}_e \rightarrow b\bar{b}W^+W^-\nu_e\bar{\nu}_e$ , hadronic	146.0	80.2%	79.9%	69.7%
$e^+e^- \rightarrow HH\nu_e\bar{\nu}_e \rightarrow b\bar{b}b\bar{b}\nu_e\bar{\nu}_e$	355.0	83.4%	82.9%	81.2%
$e^+e^- \rightarrow HH\nu_e\bar{\nu}_e \rightarrow \text{other}$	675.0	36.7%	35.8%	25.2%
$e^+e^- \rightarrow qqH\nu\bar{\nu}$	6120	59.5%	58.5%	40.4%
$e^+e^- \rightarrow c\bar{c}H\nu\bar{\nu}$	2300	64.8%	58.4%	39.3%
$e^+e^- \rightarrow b\bar{b}H\nu\bar{\nu}$	3560	69.7%	68.4%	64.2%
$e^+e^- \rightarrow qqqq$	1093000	48.5%	39.7%	3.0%
$e^+e^- \rightarrow qqqq\ell\ell$	338600	14.7%	14.2%	0.7%
$e^+e^- \rightarrow qqqq\ell\nu$	213200	19.7%	19.4%	10.0%
$e^+e^- \rightarrow qqqq\nu\bar{\nu}$	143000	58.4%	57.3%	11.9%
$e^+e^- \rightarrow qq$	5897800	62.8%	13.2%	2.7%
$e^+e^- \rightarrow qq\ell\nu$	11121800	28.3%	11.9%	0.3%
$e^+e^- \rightarrow qq\ell\ell$	6639200	38.3%	2.9%	0.7%
$e^+e^- \rightarrow qq\nu\nu$	2635000	71.4%	24.1%	5.3%
$e^\pm\gamma(BS) \rightarrow e^\pm qqqq$	4007354	23.3%	21.5%	0.8%
$e^\pm\gamma(EPA) \rightarrow e^\pm qqqq$	1151200	12.0%	11.0%	0.5%
$e^\pm\gamma(BS) \rightarrow \nu qqqq$	829184	61.5%	59.3%	19.9%
$e^\pm\gamma(EPA) \rightarrow \nu qqqq$	216800	30.8%	29.8%	9.4%
$e^\pm\gamma(BS) \rightarrow qqH\nu$	185018.0	58.2%	56.1%	37.2%
$e^\pm\gamma(EPA) \rightarrow qqH\nu$	46800	29.9%	28.9%	19.1%
$\gamma(BS)\gamma(BS) \rightarrow qqqq$	18009414	54.2%	49.2%	1.9%
$\gamma(BS)\gamma(EPA) \rightarrow qqqq$	3824548	33.5%	30.2%	1.2%
$\gamma(EPA)\gamma(BS) \rightarrow qqqq$	3828498	33.7%	30.3%	1.2%
$\gamma(EPA)\gamma(EPA) \rightarrow qqqq$	805400	22.0%	19.8%	0.8%

**Table 7.17:** List of signal and background samples with selection efficiencies and event numbers after the pre-selection cuts at  $\sqrt{s} = 3 \text{ TeV}$ , assuming a luminosity of  $2000 \text{ fb}^{-1}$ . The selection efficiencies are presented in a “flow” fashion. Every selection cut contains all the cuts to the left of it.



**Figure 7.14:** Distributions of variables used in discriminative pre-selection cuts at  $\sqrt{s} = 3 \text{ TeV}$ , after rejecting events with identified leptons and jet pairing.



**Figure 7.15:** Sum of b-jet tag as a function of  $y_{34}$  at  $\sqrt{s} = 3$  TeV, shown for hadronic decay of  $\text{HH} \rightarrow b\bar{b}W^+W^- \rightarrow b\bar{b}qqqq$  and  $\text{HH} \rightarrow b\bar{b}W^+W^-$  sub-channels.

are shown in figure 7.15. The selection efficiencies after the mutually exclusive cuts are shown in table 7.18.

The same set of variables for the MVA are used at  $\sqrt{s} = 3$  TeV as in the analysis at  $\sqrt{s} = 1.4$  TeV. The Boosted Decision Tree classifier is optimised at  $\sqrt{s} = 3$  TeV. The efficiencies of the pre-selection cuts and the efficiencies of the MVA selections are listed in table 7.19, alongside the number of events after the MVA selection. Background channels that are dominant after the MVA selection are almost identical to those at  $\sqrt{s} = 1.4$  TeV. Hence see section 7.10 for discussion.

Channel	Previous cuts and loose cuts	Mutually exclusive
$e^+e^- \rightarrow HH\nu_e\bar{\nu}_e \rightarrow b\bar{b}W^+W^-\nu_e\bar{\nu}_e$ , hadronic	69.5%	61.7%
$e^+e^- \rightarrow HH\nu_e\bar{\nu}_e \rightarrow b\bar{b}b\bar{b}\nu_e\bar{\nu}_e$	81.1%	18.8%
$e^+e^- \rightarrow HH\nu_e\bar{\nu}_e \rightarrow$ other	25.1%	20.0%
$e^+e^- \rightarrow q\bar{q}H\nu\bar{\nu}$	40.3%	35.9%
$e^+e^- \rightarrow c\bar{c}H\nu\bar{\nu}$	39.2%	26.2%
$e^+e^- \rightarrow b\bar{b}H\nu\bar{\nu}$	64.2%	25.9%
$e^+e^- \rightarrow qqqq$	2.5%	1.4%
$e^+e^- \rightarrow qqqq\ell\ell$	0.7%	0.6%
$e^+e^- \rightarrow qqqq\ell\nu$	9.2%	7.2%
$e^+e^- \rightarrow qqqq\nu\bar{\nu}$	11.8%	9.0%
$e^+e^- \rightarrow qq$	2.5%	1.4%
$e^+e^- \rightarrow qq\ell\nu$	0.3%	0.1%
$e^+e^- \rightarrow qq\ell\ell$	0.7%	0.4%
$e^+e^- \rightarrow qq\nu\nu$	5.3%	3.1%
$e^\pm\gamma(BS) \rightarrow e^\pm qqqq$	0.8%	0.7%
$e^\pm\gamma(EPA) \rightarrow e^\pm qqqq$	0.4%	0.4%
$e^\pm\gamma(BS) \rightarrow \nu qqqq$	19.8%	16.4%
$e^\pm\gamma(EPA) \rightarrow \nu qqqq$	9.2%	7.6%
$e^\pm\gamma(BS) \rightarrow qqH\nu$	37.2%	30.2%
$e^-\gamma(EPA) \rightarrow qqH\nu$	18.7%	15.5%
$\gamma(BS)\gamma(BS) \rightarrow qqqq$	1.9%	1.7%
$\gamma(BS)\gamma(EPA) \rightarrow qqqq$	1.1%	1.0%
$\gamma(EPA)\gamma(BS) \rightarrow qqqq$	1.1%	1.0%
$\gamma(EPA)\gamma(EPA) \rightarrow qqqq$	0.7%	0.6%

**Table 7.18:** List of signal and background samples with the selection efficiencies after loose cuts and mutually exclusive cuts at  $\sqrt{s} = 3$  TeV. The selection efficiencies are presented in a “flow” fashion. Every selection cut contains all the cuts to the left of it.

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

**Table 7.19:** List of signal and background samples selection efficiencies and event numbers after MVA at  $\sqrt{s} = 3 \text{ TeV}$ , for a luminosity of  $2000 \text{ fb}^{-1}$ . The number of events, selection efficiency of pre-selection, selection efficiency of MVA after pre-selection, number of events after MVA are shown. - represents a number less than 0.01.

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

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

The strategy of the analysis is very similar to the hadronic decay analysis. The main difference are that there is one lepton in the final state and the final state has four quarks instead of six.  $H_{bb}$  and  $W$  can not be reconstructed due to the leptonic decay of one of the  $W$ . Hence, events are selected when there is one identified lepton using the same lepton finding processors. The jet reconstruction parameters are the same as hadronic decay analysis at the  $\sqrt{s} = 3$  TeV. The pre-selection cuts are the same hadronic decay analysis at the  $\sqrt{s} = 3$  TeV, with the performance listed in table 7.20. There are no mutually exclusive cuts since there is no semi-leptonic analysis in the parallel analysis. Variables used in the MVA classifier are listed in table 7.21. The efficiencies of the pre-selection cuts, the efficiencies of the MVA selections are listed in table 7.22, alongside the number of events after the MVA selection. Since there are three neutrinos in the final state, reconstructing the correct signal event topology is more difficult. The MVA performance is worse and almost all background channels are non-zero the MVA. Nevertheless, dominant background channels are almost identical to those at  $\sqrt{s} = 1.4$  TeV. Discussion of the results are provided in section 7.10.

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

**Table 7.20:** Pre-selection cuts at  $\sqrt{s} = 3$  TeV for semi-leptonic analysis.

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

**Table 7.21:** Variables used in MVA for semi-leptonic analysis at  $\sqrt{s} = 3 \text{ TeV}$ .

## 7.13 Result interpretation

The results of analyses at the  $\sqrt{s} = 1.4 \text{ TeV}$  and  $\sqrt{s} = 3 \text{ TeV}$  are summarised in table 7.23. The expected precisions on the cross sections, which is roughly  $\sqrt{N_S + N_B}/N_S$  at  $\sqrt{s} = 1.4 \text{ TeV}$  and  $3 \text{ TeV}$ , are:

$$\frac{\Delta [\sigma(HH\nu_e\bar{\nu}_e)]}{\sigma(HH\nu_e\bar{\nu}_e)} = \begin{cases} 179\%, & \text{at } \sqrt{s} = 1.4 \text{ TeV}, \\ 92\%, & \text{at } \sqrt{s} = 3 \text{ TeV}, \end{cases} \quad (7.11)$$

where  $N_S$  is all  $e^+e^- \rightarrow HH\nu_e\bar{\nu}_e$  events passed the MVA. The result at  $\sqrt{s} = 3 \text{ TeV}$  combines hadronic and semi-leptonic decay sub-channels.

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

$$\frac{\Delta \lambda}{\lambda} \approx \kappa \cdot \frac{\Delta [\sigma(HH\nu_e\bar{\nu}_e)]}{\sigma(HH\nu_e\bar{\nu}_e)}. \quad (7.12)$$

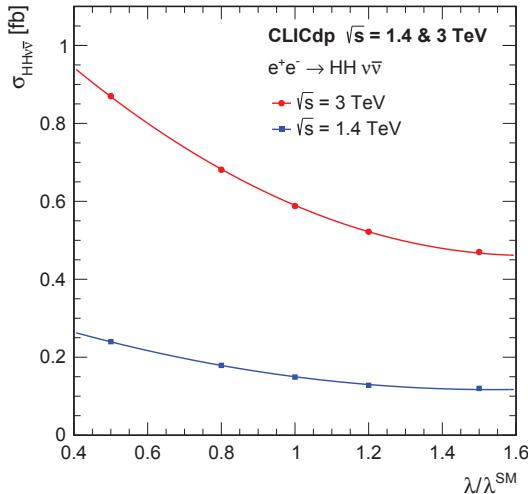
$\kappa$  can be extracted by varying the  $\lambda$  and parameterising the cross section change at a general level. Figure 7.16 shows the cross section as a function of the coupling at generator level for  $\sqrt{s} = 1.4 \text{ TeV}$  and  $\sqrt{s} = 3 \text{ TeV}$ . The negative gradient indicates that the dependence on  $\lambda$  experiences the destructive interference with other SM Feynman diagrams. At the SM  $\lambda$  value, the  $\kappa$  is 1.22 and 1.47 at  $\sqrt{s} = 1.4 \text{ TeV}$  and  $3 \text{ TeV}$  respectively. Since  $\kappa$  is extracted from the relation at generator level, the fully simulated

Channel / Efficiency $\sqrt{s} = 1.4 \text{ TeV}$	N	$\varepsilon_{\text{presel}}$	$\varepsilon_{\text{MVA}}$	$N_{\text{MVA}}$
$e^+e^- \rightarrow HH\nu_e\bar{\nu}_e \rightarrow b\bar{b}W^+W^-\nu_e\bar{\nu}_e$ , semi-leptonic	96.8	44.6%	21.9%	13.11
$e^+e^- \rightarrow HH\nu_e\bar{\nu}_e \rightarrow b\bar{b}b\bar{b}\nu_e\bar{\nu}_e$	355.0	13.3%	10.9%	5.38
$e^+e^- \rightarrow HH\nu_e\bar{\nu}_e \rightarrow \text{other}$	724.2	13.1%	13.6%	12.75
$e^+e^- \rightarrow q\bar{q}H\nu\bar{\nu}$	6120	7.4%	13.7%	62.63
$e^+e^- \rightarrow c\bar{c}H\nu\bar{\nu}$	2300	6.3%	12.1%	17.10
$e^+e^- \rightarrow b\bar{b}H\nu\bar{\nu}$	3560	15.9%	5.1%	18.03
$e^+e^- \rightarrow qqqq$	1093000	0.6%	0.2%	15.04
$e^+e^- \rightarrow qqqq\ell\ell$	338600	1.0%	0.06%	1.85
$e^+e^- \rightarrow qqqq\ell\nu$	213200	27.6%	0.5%	270.33
$e^+e^- \rightarrow qqqq\nu\bar{\nu}$	143000	1.9%	1.6%	43.78
$e^+e^- \rightarrow qq$	5897800	0.4%	0.3%	60.82
$e^+e^- \rightarrow qq\ell\nu$	11121800	0.3%	0.08%	21.24
$e^+e^- \rightarrow qq\ell\ell$	6639200	0.6%	0.2%	84.14
$e^+e^- \rightarrow qq\nu\nu$	2635000	0.4%	0.9%	92.55
$e^\pm\gamma(BS) \rightarrow e^\pm qqqq$	4007354	1.2%	-	-
$e^\pm\gamma(EPA) \rightarrow e^\pm qqqq$	1151200	1.1%	-	-
$e^\pm\gamma(BS) \rightarrow \nu qqqq$	829184	3.6%	1.5%	452.45
$e^\pm\gamma(EPA) \rightarrow \nu qqqq$	216800	11.0%	0.9%	200.65
$e^\pm\gamma(BS) \rightarrow qqH\nu$	185018	7.9%	10.4%	1521.93
$e^-\gamma(EPA) \rightarrow qqH\nu$	46800	22.8%	7.1%	750.85
$\gamma(BS)\gamma(BS) \rightarrow qqqq$	18009414	0.4%	-	-
$\gamma(BS)\gamma(EPA) \rightarrow qqqq$	3824548	1.0%	-	-
$\gamma(EPA)\gamma(BS) \rightarrow qqqq$	3828498	1.0%	0.08%	28.85
$\gamma(EPA)\gamma(EPA) \rightarrow qqqq$	805400	1.1%	-	-

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

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

**Table 7.23:** Number of signal and background events, and significance after MVA for all  $HH \rightarrow b\bar{b}W^+W^-$  analyses.



**Figure 7.16:** Cross section for the  $e^+e^- \rightarrow HH\nu_e\bar{\nu}_e$  process as a function of the ratio  $\lambda/\lambda_{SM}$  at  $\sqrt{s} = 1.4 \text{ TeV}$  and  $3 \text{ TeV}$ , taken from [15].

reconstruction selection may favour certain Feynman diagrams, therefore affecting the sensitivity to  $\lambda$ .

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

$$\frac{\Delta\lambda}{\lambda} = \begin{cases} 218\%, & \text{at } \sqrt{s} = 1.4 \text{ TeV}, \\ 135\%, & \text{at } \sqrt{s} = 3 \text{ TeV}. \end{cases} \quad (7.13)$$

Since the Feynman diagrams for the double Higgs boson productions include t-channel WW-fusion, the cross section can be enhanced by using polarised electron beam. For  $P(e^-) = 80\%$ , the uncertainty of  $\lambda$  becomes:

$$\frac{\Delta\lambda}{\lambda} = \begin{cases} 163\%, & \text{at } \sqrt{s} = 1.4 \text{ TeV}, \\ 97\%, & \text{at } \sqrt{s} = 3 \text{ TeV}. \end{cases} \quad (7.14)$$

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

## 7.14 Combined results

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

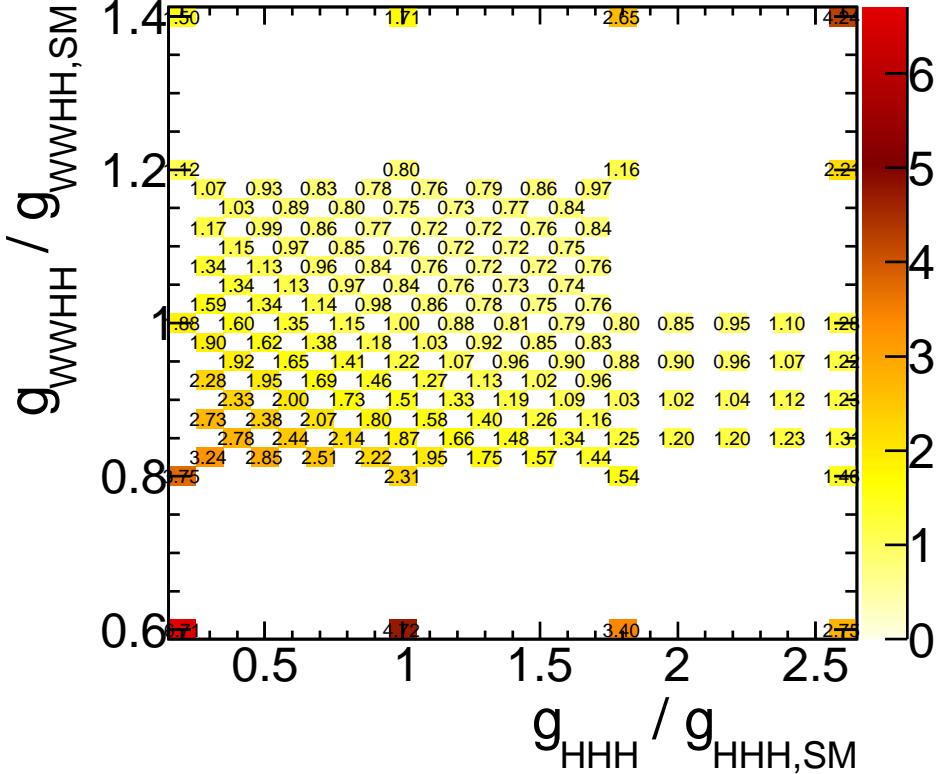
$$\frac{\Delta [\sigma(HH\nu_e\bar{\nu}_e)]}{\sigma(HH\nu_e\bar{\nu}_e)} = \begin{cases} 44\%, & \text{at } \sqrt{s} = 1.4 \text{ TeV}, \\ 20\%, & \text{at } \sqrt{s} = 3 \text{ TeV}, \end{cases} \quad (7.15)$$

This translates to the uncertainty on the Higgs triple self coupling  $\lambda$ , without electron polarisation:

$$\frac{\Delta\lambda}{\lambda} = \begin{cases} 54\%, & \text{at } \sqrt{s} = 1.4 \text{ TeV}, \\ 29\%, & \text{at } \sqrt{s} = 3 \text{ TeV}. \end{cases} \quad (7.16)$$

## 7.15 Simultaneous couplings extraction

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

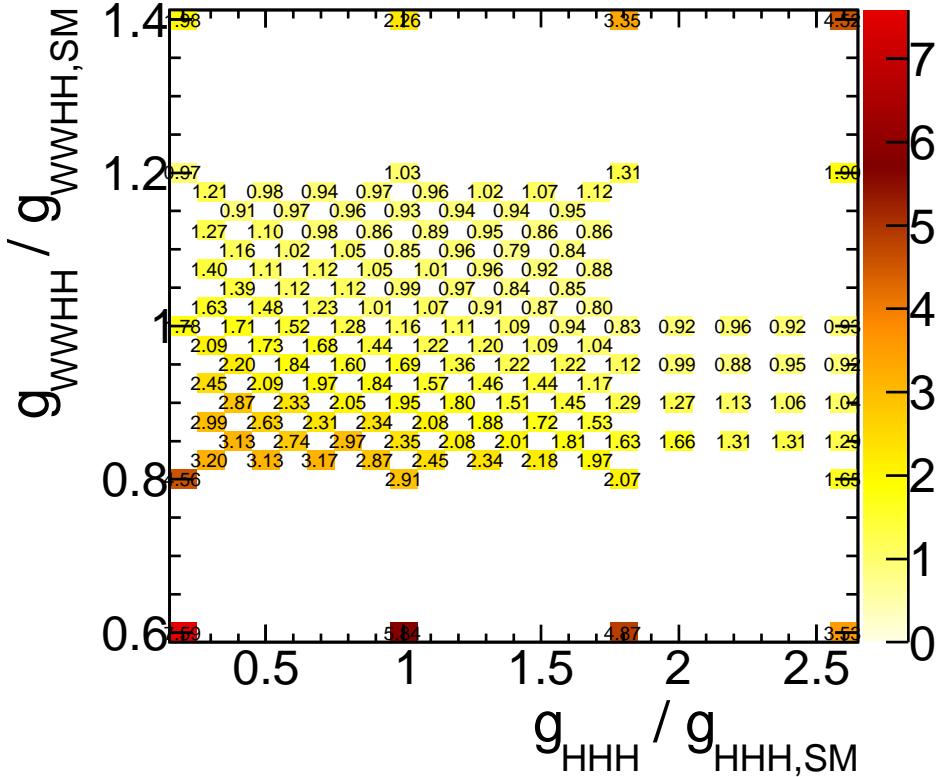


**Figure 7.17:** Normalised cross section for the  $e^+e^- \rightarrow HH\nu_e\bar{\nu}_e$  process as a function of the  $g_{HHH}/g_{HHH,SM}$  and  $g_{WWWH}/g_{WWWH,SM}$  at  $\sqrt{s} = 3$  TeV.

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

The normalised cross section of the  $e^+e^- \rightarrow HH\nu_e\bar{\nu}_e$  as a function of  $g_{HHH}$  and  $g_{WWWH}$  is shown in figure 7.17. The SM cross section is normalised to 1. Around the SM coupling value, the cross section increases with the decrease of  $g_{HHH}$  and with the increase of  $g_{WWWH}$ . Hence the cross sections along the anti-diagonal are nearly constant, which would be difficult to precisely determine the statistical uncertainty on the coupling measurements.

To determine the uncertainty on the coupling measurements, the variables proposed in the generator-level study in section 2.9 are used: the invariant mass of the two Higgs system,  $m_{HH}$ , and the sum of their transverse momenta,  $H_T$ .

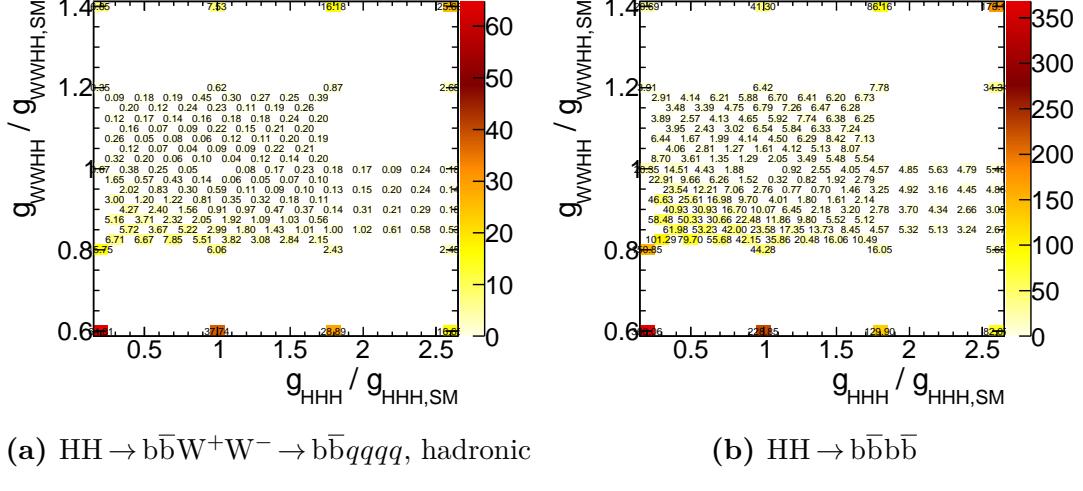


**Figure 7.18:** TODO TO UPDATE The significance for the  $e^+e^- \rightarrow HH\nu_e\bar{\nu}_e$  process as a function of the  $g_{HHH}/g_{HHH,SM}$  and  $g_{WWH}/g_{WWH,SM}$  at  $\sqrt{s} = 3$  TeV, using subchannel hadronic decay  $HH \rightarrow b\bar{b}W^+W^-$ , assuming a luminosity of  $3000 fb^{-1}$ .

The strategy for coupling extraction is described below. Simulated events with non-SM couplings are generated and reconstructed. These events went through the analysis chain discussed in this chapter with the same cuts and the same MVA classifier trained with sample of the SM coupling. The signal significance of the double Higgs events with sub-channel hadronic decay  $HH \rightarrow b\bar{b}W^+W^-$  as a function of  $g_{HHH}$  and  $g_{WWH}$  is shown in figure 7.18.

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

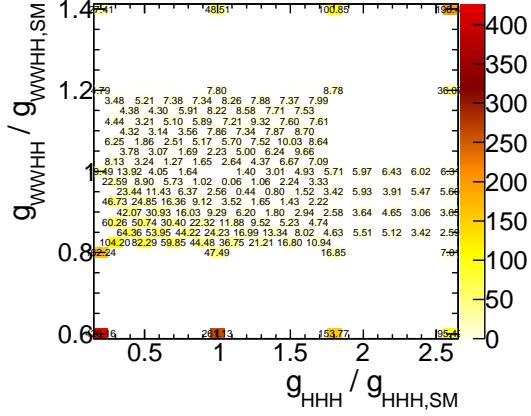
$$\chi^2 = \sum_i^{bins} \frac{(N_i - N_{i,observed})^2}{N_i}, \quad (7.17)$$



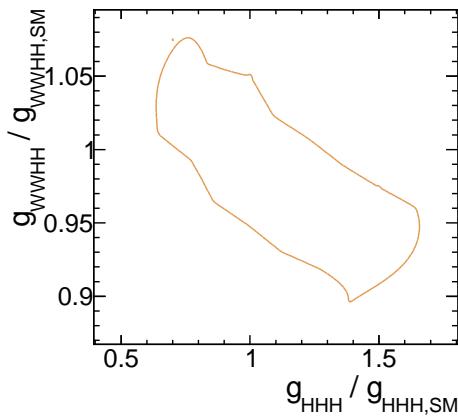
**Figure 7.19:** TODO TO UPDATE The  $\chi^2$  for the  $e^+e^- \rightarrow HH\nu_e\bar{\nu}_e$  process as a function of the  $g_{\text{HHH}}/g_{\text{HHH,SM}}$  and  $g_{\text{WWWH}}/g_{\text{WWWH,SM}}$  at  $\sqrt{s} = 3$  TeV, using sub-channel hadronic decay  $\text{HH} \rightarrow b\bar{b}W^+W^-$  and sub-channel, assuming a luminosity of  $3000 \text{ fb}^{-1}$ .

where  $N_i$  is the number of event expected in a kinematic bin  $i$  for a non-SM coupling sample.  $N_{i,\text{observed}}$  is the number of event observed in a kinematic bin  $i$ . Here the observed set can be the SM coupling sample. The expression is summing over all kinematic bins. By construction, the SM coupling point has a  $\chi^2$  of 0. Figure 7.19 shows the  $\chi^2$  as a function of  $g_{\text{HHH}}$  and  $g_{\text{WWWH}}$  for two sub-channels; hadronic decay  $\text{HH} \rightarrow b\bar{b}W^+W^-$  and  $\text{HH} \rightarrow b\bar{b}b\bar{b}$ . The  $\chi^2$  values for the  $\text{HH} \rightarrow b\bar{b}b\bar{b}$  sub-channel are larger due to larger signal significance obtained with this sub-channel.

Two sub-channels, hadronic decay  $\text{HH} \rightarrow b\bar{b}W^+W^-$  and  $\text{HH} \rightarrow b\bar{b}b\bar{b}$ , are combined to increase the statistical precision on the coupling measurements. Two  $\chi^2$  surfaces are summed. To avoid statistical fluctuations in the sample, a toy MC experiment is performed. The SM coupling samples are treated as a data template set. 100000 data sets are generated by fluctuating the event number in each kinematic bin in the data template set according to Poisson distribution. The  $\chi^2$  is performed and summed using these generated data sets as the observed data. The summed  $\chi$  is then averaged over the number of data sets (100000) and normalised such that the  $\chi^2$  at the SM coupling is 0. Since only the difference between the non-SM and SM  $\chi^2$  is used for the coupling measurements, the normalisation does not affect the measurements and helps to ease the visualisation. Figure 7.20 shows the normalised  $\chi^2$  after averaging over the toy MC experiments as a function of  $g_{\text{HHH}}/g_{\text{HHH,SM}}$  and  $g_{\text{WWWH}}/g_{\text{WWWH,SM}}$ . The  $\chi^2$  changes slowly along the anti-diagonal which is similar to the cross section plot.



**Figure 7.20:** TODO TO UPDATE Normalised  $\chi^2$ , after averaging the toy MC experiments, as a function of  $g_{\text{HHH}}/g_{\text{HHH,SM}}$  and  $g_{\text{WWHH}}/g_{\text{WWHH,SM}}$ , combining hadronic decay  $\text{HH} \rightarrow b\bar{b}W^+W^-$  and  $\text{HH} \rightarrow b\bar{b}bb$  sub-channels, assuming a luminosity of  $3000\text{fb}^{-1}$ .



**Figure 7.21:** TODO Contour plot of  $\chi^2$ , after averaging the toy MC experiments, as a function of  $g_{\text{HHH}}/g_{\text{HHH,SM}}$  and  $g_{\text{WWHH}}/g_{\text{WWHH,SM}}$ , combining hadronic decay  $\text{HH} \rightarrow b\bar{b}W^+W^-$  and  $\text{HH} \rightarrow b\bar{b}bb$  sub-channels, assuming a luminosity of  $3000\text{fb}^{-1}$ .

Since there are two couplings in this  $\chi^2$  surface, the degree of freedom for this fit is 2. A contour of 68% confidence ( $\chi^2 = 2.3$ ) can be drawn by interpolating between points on the surface. Figure 7.21 shows the contour. The counter can be sliced one dimensionally to extract the uncertainty of one coupling for a given value of the other coupling. For example:

$$\frac{\Delta g_{WWHH}}{g_{WWHH}} \simeq 4.9\% \text{ for } g_{HHH} = g_{HHHSM} \quad (7.18)$$

$$\frac{\Delta g_{HHH}}{g_{HHH}} \simeq 29\% \text{ for } g_{WWHH} = g_{WWHHS} \quad (7.19)$$

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

# Colophon

This thesis was made in L<sup>A</sup>T<sub>E</sub>X 2 <sub>$\varepsilon$</sub>  using the “heptesis” class [80].



# Bibliography

- [1] ATLAS Collaboration, G. Aad *et al.*, Phys.Lett. **B716**, 1 (2012), 1207.7214.
- [2] Particle Data Group, K. A. Olive *et al.*, Chin. Phys. **C38**, 090001 (2014).
- [3] M. Thomson, *Modern particle physics* (Cambridge University Press, New York, 2013).
- [4] D. Tong, Lectures on quantum field theory, 2006.
- [5] B. Gripaios, Lectures on gauge field theory, 2017.
- [6] SLD Electroweak Group, DELPHI, ALEPH, SLD, SLD Heavy Flavour Group, OPAL, LEP Electroweak Working Group, L3, S. Schael *et al.*, Phys. Rept. **427**, 257 (2006), hep-ex/0509008.
- [7] S. Weinberg, Phys. Rev. Lett. **19**, 1264 (1967).
- [8] D. Rainwater, Searching for the Higgs boson, in *Proceedings of Theoretical Advanced Study Institute in Elementary Particle Physics : Exploring New Frontiers Using Colliders and Neutrinos (TASI 2006): Boulder, Colorado, June 4-30, 2006*, pp. 435–536, 2007, hep-ph/0702124.
- [9] D. B. Kaplan and H. Georgi, Phys. Lett. **B136**, 183 (1984).
- [10] W. D. Goldberger, B. Grinstein, and W. Skiba, Phys. Rev. Lett. **100**, 111802 (2008), 0708.1463.
- [11] G. F. Giudice, C. Grojean, A. Pomarol, and R. Rattazzi, JHEP **06**, 045 (2007), hep-ph/0703164.
- [12] R. Contino, C. Grojean, M. Moretti, F. Piccinini, and R. Rattazzi, JHEP **05**, 089 (2010), 1002.1011.
- [13] R. Contino, C. Grojean, D. Pappadopulo, R. Rattazzi, and A. Thamm, JHEP **02**,

- 006 (2014), 1309.7038.
- [14] V. Barger, T. Han, P. Langacker, B. McElrath, and P. Zerwas, Phys. Rev. **D67**, 115001 (2003), hep-ph/0301097.
- [15] H. Abramowicz *et al.*, (2016), 1608.07538.
- [16] B. K. Bullock, K. Hagiwara, and A. D. Martin, Phys. Lett. **B273**, 501 (1991).
- [17] Y.-S. Tsai, Phys. Rev. **D4**, 2821 (1971), [Erratum: Phys. Rev.D13,771(1976)].
- [18] CMS Collaboration, S. Chatrchyan *et al.*, Phys.Lett. **B716**, 30 (2012), 1207.7235.
- [19] J. Brau *et al.*, (2007).
- [20] L. Linssen, A. Miyamoto, M. Stanitzki, and H. Weerts, (2012), 1202.5940.
- [21] H. Baer *et al.*, (2013), 1306.6352.
- [22] M. Aicheler *et al.*, (2012).
- [23] M. Thomson, Nucl.Instrum.Meth. **A611**, 25 (2009), 0907.3577.
- [24] J. S. Marshall, A. Mížnnich, and M. A. Thomson, Nucl. Instrum. Meth. **A700**, 153 (2013), 1209.4039.
- [25] I. G. Knowles and G. D. Lafferty, J. Phys. **G23**, 731 (1997), hep-ph/9705217.
- [26] M. Green, *Electron-Positron Physics at the Z*Studies in high energy physics, cosmology, and gravitation (Taylor & Francis, 1998).
- [27] CALICE, C. Adloff, (2011), 1105.0511.
- [28] B. Parker *et al.*, (2009).
- [29] CALICE, JINST **7**, P04015 (2012), 1201.1653.
- [30] 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).
- [31] W. Kilian, T. Ohl, and J. Reuter, European Physical Journal C **71** (2011).
- [32] M. Moretti, T. Ohl, and J. Reuter, p. 1981 (2001), hep-ph/0102195.
- [33] T. Sjostrand, (1995), hep-ph/9508391.

- [34] OPAL, G. Alexander *et al.*, Z. Phys. **C69**, 543 (1996).
- [35] S. Jadach, Z. Was, R. Decker, and J. H. Kuhn, Comput. Phys. Commun. **76**, 361 (1993).
- [36] GEANT4, S. Agostinelli *et al.*, Nucl.Instrum.Meth. **A506**, 250 (2003).
- [37] P. Mora de Freitas and H. Videau, p. 623 (2002).
- [38] F. Gaede, Nucl. Instrum. Meth. **A559**, 177 (2006).
- [39] J. S. Marshall and M. A. Thomson, Eur. Phys. J. **C75**, 439 (2015), 1506.05348.
- [40] J. S. Marshall, Presentation on pandorapfa with lc reconstruction, [https://github.com/PandoraPFA/Documentation/blob/master/Pandora\\_LC\\_Reconstruction.pdf](https://github.com/PandoraPFA/Documentation/blob/master/Pandora_LC_Reconstruction.pdf), 2017.
- [41] A. Sailer, Luminosities for ee, eg, and gg interactions, [https://indico.cern.ch/event/233706/contributions/499053/attachments/390186/542711/130514\\_LuminosityNormalisation.pdf](https://indico.cern.ch/event/233706/contributions/499053/attachments/390186/542711/130514_LuminosityNormalisation.pdf), 2013.
- [42] D. Schulte, (1999).
- [43] T. Barklow, D. Dannheim, M. O. Sahin, and D. Schulte, (2012).
- [44] G. F. Sterman and S. Weinberg, Phys. Rev. Lett. **39**, 1436 (1977).
- [45] S. Moretti, L. Lonnblad, and T. Sjostrand, JHEP **08**, 001 (1998), hep-ph/9804296.
- [46] G. P. Salam, Eur. Phys. J. **C67**, 637 (2010), 0906.1833.
- [47] A. Ali and G. Kramer, Eur. Phys. J. **H36**, 245 (2011), 1012.2288.
- [48] M. Cacciari, G. P. Salam, and G. Soyez, Eur. Phys. J. **C72**, 1896 (2012), 1111.6097.
- [49] M. Cacciari and G. P. Salam, Phys. Lett. **B641**, 57 (2006), hep-ph/0512210.
- [50] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber, Nucl. Phys. **B406**, 187 (1993).
- [51] S. D. Ellis and D. E. Soper, Phys. Rev. **D48**, 3160 (1993), hep-ph/9305266.
- [52] S. Catani, Y. L. Dokshitzer, M. Olsson, G. Turnock, and B. R. Webber, Phys. Lett. **B269**, 432 (1991).
- [53] M. Battaglia and F. P., CERN Report No. LCD-Note-2010-006, 2010 (unpublished).

- [54] M. Boronat, J. Fuster, I. Garcia, E. Ros, and M. Vos, Phys. Lett. **B750**, 95 (2015), 1404.4294.
- [55] T. Suehara and T. Tanabe, Nucl. Instrum. Meth. **A808**, 109 (2016), 1506.08371.
- [56] LCFI, D. Bailey *et al.*, Nucl. Instrum. Meth. **A610**, 573 (2009), 0908.3019.
- [57] Linear Collider ILD Concept Group -, T. Abe *et al.*, (2010), 1006.3396.
- [58] H. Aihara *et al.*, (2009), 0911.0006.
- [59] A. Hocker *et al.*, PoS **ACAT**, 040 (2007), physics/0703039.
- [60] E. Farhi, Phys. Rev. Lett. **39**, 1587 (1977).
- [61] G. Hanson *et al.*, Phys. Rev. Lett. **35**, 1609 (1975).
- [62] Y. Freund and R. E. Schapire, Journal of Computer and System Sciences **55**, 119 (1997).
- [63] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition* Springer Series in Statistics (Springer New York, 2009).
- [64] G. Kačarević, Prelection for  $h \rightarrow \gamma \gamma$  at 3 tev, <https://indico.cern.ch/event/577810/contributions/2485070/attachments/1424897/2185427/GoranKacarevic.pdf>, 2017.
- [65] B. Xu, Improvement of photon reconstruction in PandoraPFA, in *Proceedings, International Workshop on Future Linear Colliders (LCWS15): Whistler, B.C., Canada, November 02-06, 2015*, 2016, 1603.00013.
- [66] E. Segrè, *Nuclei and particles: an introduction to nuclear and subnuclear physics* (W. A. Benjamin, 1977).
- [67] E. Longo and I. Sestili, Nucl. Instrum. Meth. **128**, 283 (1975), [Erratum: Nucl. Instrum. Meth. 135, 587(1976)].
- [68] ALEPH collaboration, S. Schael *et al.*, Phys. Rept. **421**, 191 (2005).
- [69] S. Berge, W. Bernreuther, and S. Kirchner, Phys. Rev. **D92**, 096012 (2015).
- [70] DELPHI collaboration, P. Abreu *et al.*, Phys. Lett. **B267**, 422 (1991).
- [71] F. Gaede and J. Engels, EUDET Report (2007).

- [72] TMVA Core Developer Team, J. Therhaag, AIP Conf.Proc. **1504**, 1013 (2009).
- [73] D. Lyth, Journal de Physique Colloques **35**, C2 (1974).
- [74] S. Dittmaier *et al.*, (2012), 1201.3084.
- [75] A. Míznnich, CERN Report No. LCD-Note-2010-009, 2010 (unpublished).
- [76] CLICdp, A. Sailer and A. Sapronov, (2017), 1702.06945.
- [77] S. Lukić, Forward electron tagging in the  $h \rightarrow \mu \mu$  analysis at 1.4 tev, <http://indico.cern.ch/event/262809/contributions/1595499/attachments/464689/643931/electronTagging.pdf>, 2013.
- [78] G. Milutinović-Dumbelović *et al.*, (2014), 1412.5791.
- [79] C. Grefe, T. Lastovicka, and J. Strube, Light Higgs Studies for the CLIC CDR, in *Helmholtz Alliance Linear Collider Forum: Proceedings of the Workshops Hamburg, Munich, Hamburg 2010-2012, Germany*, pp. 258–264, Hamburg, 2013, DESY, DESY, 1205.3908.
- [80] A. Buckley, The heptesis L<sup>A</sup>T<sub>E</sub>X class.



# List of figures

2.1	SM Higgs boson decay width and branching ratios . . . . .	10
2.4	Two-dimensional distribution of $Z \rightarrow \tau^+ \tau^-$ and $H \rightarrow \tau^+ \tau^-$ . . . . .	16
3.1	A layout of the International Linear Collider complex, taken from [21]. . . . .	18
3.2	A layout of the Compact Linear Collider at final stage of a centre-of-mass of 3 TeV, taken from [22]. . . . .	19
3.3	A typical topology of a 250 GeV jet. . . . .	21
3.4	International Large Detector and the Silicon Detector for the International Linear Collider. . . . .	23
3.5	A top quadrant view of the ILD silicon envelope system . . . . .	25
3.6	A cross section through electromagnetic calorimeter layers. . . . .	26
3.7	CALICE AHCAL technological prototype module and jet energy resolution. .	27
3.8	Sensitive layers of the ILD muon system, taken from [21]. . . . .	28
3.9	The forward calorimeters of the ILD. . . . .	28
3.10	Longitudinal cross section of top quadrant of the ILD and the CLIC_ILD detector concepts. . . . .	30
4.1	Illustration of the cone based clustering, taken from [40] . . . . .	37
4.2	Illustration of the clustering algorithm used in the PandoraPFA, taken from [40] . . . . .	38
4.3	Topological association in the PandoraPFA. . . . .	39

---

4.4	Illustration of the re-clustering algorithm in PandoraPFA . . . . .	40
4.5	Effect of the suppression of the background with the tight PFO selection. . . . .	43
4.6	Example of model efficiency as a complexity of model parameter. Here the model is boosted decision tree. The model parameter is depth of tree. From tree depth 6 onwards, overfitting occurs. . . . .	48
4.7	Example of a decision tree. . . . .	51
5.1	Simulated longitudinal electromagnetic shower profile as a function of depth for electrons and photons. . . . .	59
5.2	A flow diagram of the PHOTON RECONSTRUCTION algorithm. . . . .	60
5.3	Example of projecting a large photon clusters containing two photons. . . . .	62
5.4	Flow chart for 2D PEAK FINDING algorithm neutral cluster variant. . . . .	63
5.5	Flow chart for 2D PEAK FINDING algorithm. . . . .	66
5.6	Variables for photon ID test, obtained from simulated 500 GeV di-jet $Z' \rightarrow u\bar{u}/d\bar{d}/s\bar{s}$ events. Photon ID in plots are determined using MC information. Figure 5.6a, Figure ??, , Figure 5.6c, and Figure 5.6d show distributions for start positions of longitudinal shower profiles, fractional difference from the expected longitudinal shower profile, energy weighted r.m.s. distance of a bin to its peak, and distance between the photon candidate and the closest track projection. . . . .	69
5.7	An event display of a typical 10 GeV photon (figure 5.7a), reconstructed into a main photon (figure 5.7b) and a photon fragment (figure 5.7c). . . . .	71
5.8	Illustration of distance metric, $d$ . . . . .	73
5.9	An event display of a typical 500 GeV photon, reconstructed into a main photon in the ECAL (yellow) and a neutral hadron fragment in the HCAL (blue). . . . .	75
5.10	Average number of photons using two photons of 500 and 50 GeV per event sample. . . . .	79
5.11	Jet energy resolution as a function of the total jet energy without and with photon related algorithms . . . . .	79

5.12	Average number of reconstructed photons and reconstructed particles, as a function of their true energy using single photon sample. . . . .	81
5.13	Average number of reconstructed photons and reconstructed particles, as a function of the MC distance separation. . . . .	82
5.14	Average fraction fragments energies of the total energy, as a function of the MC distance separation . . . . .	82
5.15	Jet energy resolution as a function of the di-jet energy . . . . .	83
5.16	Average number of photons, as a function of the MC distance separation for different algorithms combinations. . . . .	84
5.17	Single photon reconstruction efficiency as a function of energy. . . . .	85
5.18	Average numbers of photon and particle using two photons of 500 and 50 GeV per event sample and with different energy pairs. . . . .	86
6.1	An example event display of simulated $e^+e^- \rightarrow \tau^+\tau^-$ event. The top tau decay is $\pi^- \pi^0$ final state and the bottom is $\pi^+ \pi^- \pi^- \pi^0$ final state. Purple clusters are $\pi^\pm$ and yellow clusters are photons. Blue region is the transverse cross section of the ECAL barrel, looking along the beam line direction. . . . .	92
6.2	Normalised distribution for selected variables. . . . .	93
6.3	The correct classification efficiency for tau hadronic decay final states as a function of the ECAL square cell sizes . . . . .	101
6.4	The tau hadronic decay efficiency as a function of the ECAL cell sizes at different $\sqrt{s}$ with the nominal ILD detector model. . . . .	102
6.5	Two-dimensional plot of tau pair polarisation correlations from Z decay, using $\tau^- \rightarrow \pi^- \nu_\tau$ decay mode. . . . .	107
7.1	The main Feynman diagrams for the leading-order $e^+e^- \rightarrow HH\nu_e\bar{\nu}_e$ processes at the CLIC. . . . .	110
7.2	BeamCAL and LumiCAL electron tagging efficiency. . . . .	120
7.3	Example MC mass fit for jet optimisation in double Higgs analysis . . . . .	123

7.4	Fitted mass, and resolution of $H_{bb}$ , $H_{WW^*}$ and $W$ at $\sqrt{s} = 1.4 \text{ TeV}$ . . . . .	125
7.5	Fitted mass peak positions and relative mass resolution of $H_{bb}$ , $H_{WW^*}$ and $W$ at $\sqrt{s} = 3 \text{ TeV}$ . . . . .	126
7.6	Performance of b-jet tagging with training samples at a) $\sqrt{s} = 1.4 \text{ TeV}$ , b) $\sqrt{s} = 3 \text{ TeV}$ . . . . .	128
7.7	The normalised distribution of the highest b-jet tag value for the signal events at $\sqrt{s} = 1.4 \text{ TeV}$ . . . . .	129
7.8	Sum of b-jet tag values as a function of $-\log(y_{34})$ at $\sqrt{s} = 1.4 \text{ TeV}$ . . . . .	130
7.9	The normalised distribution of $m_{H_{bb}}$ after jet pairing, for a) signal channel, hadronic $W^+W^-$ decay of $HH \rightarrow b\bar{b}W^+W^-$ , b) sum of all background channels. All plots are shown for $\sqrt{s} = 1.4 \text{ TeV}$ . . . . .	131
7.10	Distributions of the invariant mass of the two Higgs system for $\sqrt{s} = 1.4 \text{ TeV}$ , assuming an intergraded luminosity of $1500 \text{ fb}^{-1}$ . . . . .	133
7.11	Distributions of the second highest b-jet tag value for $\sqrt{s} = 1.4 \text{ TeV}$ , assuming an intergraded luminosity of $1500 \text{ fb}^{-1}$ . . . . .	134
7.12	Distributions of the transverse momentum of the two Higgs system for $\sqrt{s} = 1.4 \text{ TeV}$ , assuming an intergraded luminosity of $1500 \text{ fb}^{-1}$ . . . . .	135
7.13	Distributions of the four variables with highest discriminating power: a) the invariant mass of $H_{bb}$ , b) the invariant mas of $H_{WW^*}$ , c) the acolinearity of the two jets associated with $H_{bb}$ , and d) the opening angles of the two jets associated with $H_{bb}$ in the decay rest frame of the $H_{bb}$ . All plots assumes an intergraded luminosity of $1500 \text{ fb}^{-1}$ at $\sqrt{s} = 1.4 \text{ TeV}$ after all pre-selection cuts applied before the MVA. . . . .	141
7.14	Distributions of variables used in discriminative pre-selection cuts at $\sqrt{s} = 3 \text{ TeV}$ . . . . .	148
7.15	Sum of b-jet tag as a function of $y_{34}$ at $\sqrt{s} = 3 \text{ TeV}$ . . . . .	148
7.16	Cross section for the $e^+e^- \rightarrow HH\nu_e\bar{\nu}_e$ process as a function of the ratio $\lambda/\lambda_{SM}$ . . . . .	154
7.17	Normalised cross section for the $e^+e^- \rightarrow HH\nu_e\bar{\nu}_e$ process as a function of the $g_{HH}/g_{HHSM}$ and $g_{WWH}/g_{WWHSM}$ at $\sqrt{s} = 3 \text{ TeV}$ . . . . .	156

7.18 TODO TO UPDATE The significance for the $e^+e^- \rightarrow HH\nu_e\bar{\nu}_e$ process as a function of the $g_{HHH}/g_{HHHSM}$ and $g_{WWHH}/g_{WWHHS}$ at $\sqrt{s} = 3$ TeV, using subchannel hadronic decay $HH \rightarrow b\bar{b}W^+W^-$ , assuming a luminosity of $3000fb^{-1}$ . . . . .	157
7.19 $\chi^2$ as a function of $g_{HHH}/g_{HHHSM}$ and $g_{WWHH}/g_{WWHHS}$ at $\sqrt{s} = 3$ TeV . . . . .	158
7.20 TODO TO UPDATE Normalised $\chi^2$ , after averaging the toy MC experiments, as a function of $g_{HHH}/g_{HHHSM}$ and $g_{WWHH}/g_{WWHHS}$ , combining hadronic decay $HH \rightarrow b\bar{b}W^+W^-$ and $HH \rightarrow b\bar{b}b\bar{b}$ sub-channels, assuming a luminosity of $3000fb^{-1}$ . . . . .	159
7.21 TODO Contour plot of $\chi^2$ , after averaging the toy MC experiments, as a function of $g_{HHH}/g_{HHHSM}$ and $g_{WWHH}/g_{WWHHS}$ , combining hadronic decay $HH \rightarrow b\bar{b}W^+W^-$ and $HH \rightarrow b\bar{b}b\bar{b}$ sub-channels, assuming a luminosity of $3000fb^{-1}$ . . . . .	159



# List of tables

3.1	A comparison of key parameters of the ILD and CLIC_ILD detector concepts. . . . .	29
3.2	Comparison of the LumiCAL and the BeamCAL at the ILD and the CLIC_ILD. . . . .	31
4.1	Luminosity ratio for processes with initial-state photons from Beamstrahlung. . . . .	42
4.2	Masses of quarks and bosons used for generating Standard Model samples. . . . .	44
4.3	The attribute of samples for the decision tree example. . . . .	52
5.1	List of variables for the likelihood based photon ID test. . . . .	68
5.2	The cuts for photon fragment removal algorithm in the ECAL. . . . .	72
5.3	Cuts for merging high energy photon fragment in the HCAL. . . . .	75
5.4	Cuts for splitting photons. . . . .	77
5.5	Photon confusion as a function of energy for reconstruction with and without photon algorithms. . . . .	80
6.1	Decay modes, detectable final state particles and branching ratios of the seven major $\tau^-$ decays. . . . .	89
6.2	Pre-selection cuts for tau lepton decay final state classification. . . . .	90
6.3	MC level pre-selection cut efficiencies for tau lepton decay final state classification. . . . .	91
6.4	Variables used in the MVA . . . . .	92

6.5 Optimised parameters for the Boosted Decision Tree with Gradient boost multiclass classifier. See section 4.7.2 for detailed explanation of variables.	97
6.6 Classification efficiency for tau decay modes.	98
6.7 Optimised parameters of PHOTONFRAGMENTREMOVAL algorithm as a function of ECAL square cell size.	99
6.8 Optimised parameters of modified ISOLATEDTAUIDENTIFER.	104
7.1 Signal and background samples with the corresponding cross sections at $\sqrt{s} = 1.4 \text{ TeV}$ .	113
7.2 Optimised parameters of ISOLATEDLEPTONFINDER	114
7.3 Optimised parameters of ISOLATEDLEPTONIDENTIFER.	116
7.4 Optimised parameters of TAUFINDER.	117
7.5 Optimised parameters of ISOLATEDTAUIDENTIFER	118
7.6 The performances of lepton finders on the signal events and selected background events at $\sqrt{s} = 1.4 \text{ TeV}$ . Numbers represent the fractions of events where no leptons are identified by the individual lepton finder.	121
7.7 The performances of lepton finders on the signal events and selected background events at $\sqrt{s} = 3 \text{ TeV}$ . Numbers represent the fractions of events where no leptons are identified by the individual lepton finder.	121
7.8 The fitted mass parameters of optimal jet reconstruction at $\sqrt{s} = 1.4 \text{ TeV}$	124
7.9 Pre-selection efficiencies at $\sqrt{s} = 1.4 \text{ TeV}$ .	132
7.10 List of signal and background samples after pre-selection cuts at $\sqrt{s} = 1.4 \text{ TeV}$ .	136
7.11 Variables used in the MVA event selection for $\sqrt{s} = 1.4 \text{ TeV}$	140
7.12 Optimised parameters for the boosted decision tree classifier. See section 4.7.2 for detailed explanation of variables.	143
7.13 Selection efficiency and number of events for signal and background at $\sqrt{s} = 1.4 \text{ TeV}$ .	144

7.14	Cross sections of samples at $\sqrt{s} = 3 \text{ TeV}$ . . . . .	145
7.15	The extracted fitted parameters of optimal jet reconstructions at $\sqrt{s} = 3 \text{ TeV}$ . . . . .	146
7.16	Pre-selection cuts at $\sqrt{s} = 3 \text{ TeV}$ . . . . .	146
7.17	Signal and background events with selection efficiency and event numbers after the pre-selection cuts at $\sqrt{s} = 3 \text{ TeV}$ . . . . .	147
7.18	List of signal and background samples after loose cuts and mutually exclusive cuts at $\sqrt{s} = 3 \text{ TeV}$ . . . . .	149
7.19	List of signal and background selection efficiencies and event numbers after MVA application at $\sqrt{s} = 3 \text{ TeV}$ . . . . .	150
7.20	Pre-selection cuts at $\sqrt{s} = 3 \text{ TeV}$ for semi-leptonic analysis. . . . .	151
7.21	Variables used in MVA for semi-leptonic analysis at $\sqrt{s} = 3 \text{ TeV}$ . . . . .	152
7.22	List of signal and background selection efficiencies and event numbers after MVA for semi-leptonic analysis at $\sqrt{s} = 3 \text{ TeV}$ . . . . .	153
7.23	Number of signal and background events, and significance after MVA for all $\text{HH} \rightarrow b\bar{b}W^+W^-$ analyses. . . . .	154