Search for effective field theory parameters for H $\rightarrow$ ZZ* $\rightarrow 4\ell$ using Normalizing Flow models

A Thesis

submitted to

Indian Institute of Science Education and Research, Pune

in partial fulfillment of the requirements for the

BS-MS Dual Degree Programme

by

Dhruvanshu Parmar

Registration No : 20171191

Indian Institute of Science Education and Research, Pune

Dr Homi Bhabha Road,

Pashan,Pune 411008,

INDIA

May 2022

Supervisor: Dr. Seema Sharma

Certificate

This is to certify that this dissertation entitled **Search for effective field theory parameters for H $\rightarrow$ ZZ* $\rightarrow$ 4ℓ using Normalizing Flow models** towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Dhruvanshu Parmar at Indian Institute of Science Education and Research under the supervision of Dr. Seema Sharma, Associate Professor, Department of Physics , during the academic year 2021-2022.

Dhruvanshu Parmar

Dr Seema Sharma

Committee:

Dr Seema Sharma

Dr Kin Ho Lucien Lo

Dedication

This thesis is dedicated to my parents, and the cohort of hostel flat no 306-D

Declaration

I hereby declare that the matter embodied in the report entitled **Search for effective field theory parameters for H $\rightarrow$ ZZ* $\rightarrow$ 4ℓ using Normalizing Flow models**, are the results of the work carried out by me at the Department of Physics, Indian Institute of Science Education and Research, Pune, under the supervision of Dr. Seema Sharma and the same has not been submitted elsewhere for any other degree.

Dr Seema Sharma Dhruvanshu Parmar

Acknowledgements

I would like to thank my supevisor Dr Seema Sharma for constantly providing her relentless support and encouragement for this project work on studying effects of new physics beyond Standard Model effects in Higgs boson interactions using advanced machine learning concepts. My sincere thanks to Dr. Kin Ho Lucien Lo for his invaluable suggestions and discussions during the course of this project, especially for the machine learning aspects which was an unexplored area for me. Also, I am grateful to our system administrator, Nisha Kurkure Ma'am and IT department for their online-support to use supercomputer PARAM-BRAHMA with ease, without which we would not have managed to finish this project in time. A special mention to our Saturday Group meetings with lots of discussions and helpful suggestions from Alpana, Bhumika, Shubham, Abhishek, Dr Aditee, Rahul, Nitish and Pradyun. Special mention of my junior buddy Dipayan Pal who helped me in the aspects of coding the scripts for the model when it was in its initial stages of development.

A BIG thanks to my Mom and Dad for their love and support required for keeping me motivated to work on the project. Special thanks to Somesh, Dheeraj and Rakshit for hanging out with me and helping each other on campus. Thanks to my D building buddies, namely Manish Gupta, Mohit, Manish Ahire who were not on campus yet who stay connected via online means.

The support and the resources provided by 'PARAM Brahma Facility' under the National Supercomputing Mission, Government of India at the Indian Institute of Science Education and Research (IISER) Pune are gratefully acknowledged.

Contents

Abstract

The ATLAS and CMS collaborations announced the discovery of a new scalar particle of 125 GeV mass in 2012 whose measured properties like production cross-sections, couplings with other Standard Model (SM) particles, charge-parity are consistent with the predictions of SM within current measurement uncertainties. In the SM, interactions of the gauge bosons with Higgs boson lead to their masses via the mechanism of spontaneous symmetry breaking. However, the SM does not explain the existence of dark matter and dominance of matter over anti-matter. Also, mass of the Higgs boson gets large corrections from quantum fluctuations in the SM theory. In absence of any direct evidence of beyond SM (BSM) physics, an alternative approach is to investigate BSM interactions of Higgs boson modelled by an Effective Field Theory (EFT). An EFT incorporates higher dimensional interaction terms which could potentially modify the already observed SM interactions. This thesis aims to explore machine learning techniques like normalizing flows based on a real-valued non-volume preserving approach to search for effects of EFT interaction dubbed the cWW interaction on modifying the SM HZZ interactions in four lepton final state i.e. $H \rightarrow ZZ^* \rightarrow \ell^+\ell^-\ell^+\ell^-$ where ℓ could be electron or muon. The analysis is implemented as a parameter estimation using likelihoods for which a complete statistical analysis is performed for the ML model.

Chapter 1

Introduction

Particle physics is the study of matter and its interactions with the environment at a level smaller than the atomic scale. In the latter half of the 20^{th} century, collective insights obtained from physicists based on experimental results and theoretical developments helped in formulation of a theoretical framework called the Standard Model (SM) to classify all the known elementary particles and interactions among them. The SM [1] showed great experimental agreements in proving existence of the quarks, the weak neutral currents, the Higgs boson, etc. There are still experimental results which are not explained by SM like dark matter, neutrino masses, etc. In order to explain this unanswered questions, attempts are being made to extend the existing SM theory to form the Beyond Standard Model (BSM) theories. One of the approaches to constrain the BSM effects is to develop a general framework extending the SM via effective field theories. This thesis presents the analysis of one such EFT model named the Higgs Effective lagrangian (HEL) in the golden decay channel $H \rightarrow ZZ^* \rightarrow 4\ell$. The EFT interaction is analysed using a machine learning technique based on normalizing flows, to constrain the Wilson coefficient cWW associated to the HZZ vertex interaction.

1.1 A brief recap of the Standard Model

The Standard Model (SM) is a theoretical model based on the quantum field theory (QFT) which encapsulates quantum mechanics and relativistic mechanics, and is renormalizable. In the SM, the elementary particles are categorized as fermions (spin-1/2 matter particles) and bosons (spin-1 force carriers). The fermions are further divided into six leptons and six quarks arranged in three generations with increasing masses. The interactions among these particles are explained by the exchange of gauge bosons namely gluons for strong interactions, photons for electromagnetic interactions, and $W^{\pm}$ and Z_0 bosons for weak interactions. The gravitational force is not incorporated in the SM yet. The SM also incorporates anti-particles corresponding to each of its particle. The Higgs boson is the spin-0 scalar particle which is known to interact with all the SM particles with non-zero masses. Each of the fermions and bosons under QFT can be expressed as fields in space and time. Hence, the various interactions between the particles can be explained as the interaction of these fields. Figure 1.1 shows the current known particle content of the

SM.

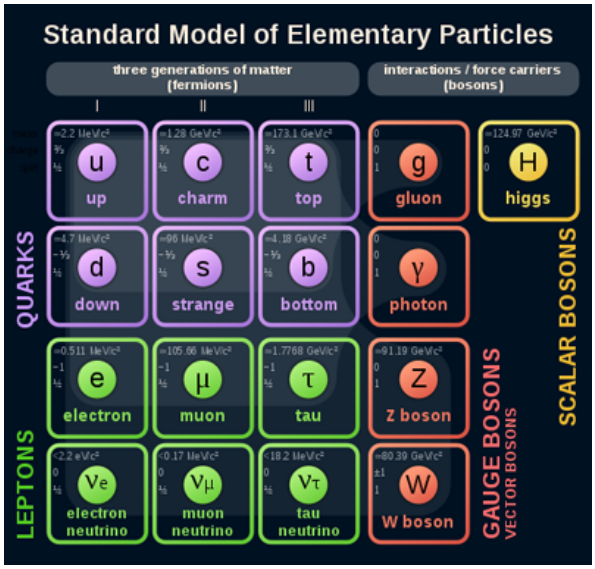


Figure 1.1: The particle content of Standard Model.

The SM is expressed as a complex gauge group with symmetry group $SU(3)_C \times SU(2)_L \times U(1)_Y$, with the components pertaining to the fundamental interactions except gravitational interaction.

- **Strong interactions** : The strong interactions are mediated by the gluons described by the $SU(3)_C$ group as quantum chromodynamics (QCD). The $SU(3)_C$ group has 8 generators which are the Gell-Mann matrices $\frac{\lambda_a}{2}$ [2]. The associated quantum number to QCD is colour, characterising quarks into three colour schemes of red, green and blue. Their antiparticles, anti-quarks similarly are represented by anti-colour. Quark fields q_f transforms under the gauge symmetry

$$q_f(x) \rightarrow e^{-i\alpha_o(x)\frac{\lambda_a}{2}} q_f(x) \tag{1.1}$$

This transformation encourages to introduce the covariant derivative D_μ to replace ∂_μ to keep Dirac lagrangian [3] gauge invariant.

$$D_\mu = \partial_\mu + ig_s \frac{\lambda_a}{2} G^a_\mu \tag{1.2}$$

G^a_μ corresponds to the gluon fields and g_s is the strong coupling constant. Gluon fields transform under $SU(3)_C$ symmetry as

$$G^a_\mu \rightarrow G^a_\mu + \alpha^b(x) f^{abc} G^c_\mu + \frac{1}{g_s}\partial_\mu \alpha^a(x) \tag{1.3}$$

f^{abc} are the structure constants in the adjoint representation of QCD group (obtained from commutation properties of the generators as $[\frac{\lambda_a}{2}, \frac{\lambda_b}{2}] = if^{abc}\frac{\lambda_c}{2}$). Finally the full QCD lagrangian is then given as :

$$\mathscr{L}_{QCD} = \bar{q}_f(i\gamma^\mu \partial_\mu - m)q_f - g_s \bar{q}_f \gamma^\mu \frac{\lambda_a}{2} q_f G^a_\mu - \frac{1}{4} G^{\mu\nu}_a G^a_{\mu\nu} \tag{1.4}$$

$$where \quad G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu - g_s f^{abc} G^b_\mu G^c_\nu \tag{1.5}$$

The first term of equation 1.4 is the free Dirac lagrangian for the quark fields (q_f), the second term describes the interactions between the quarks and gluon fields and the last third term is the gauge-invariant kinetic term consisting of gluon field strength tensors given by equation 1.5. It should be noted that the QCD lagrangian does not incorporate mass terms.

- **Electroweak interaction** : The electroweak interactions are gauge-invariant interactions under $SU(2)_L \times U(1)_Y$ symmetry.

  - $SU(2)_L$ is a non-abelian group with associated quantum number as weak isospin (I_3). This group has three generators as the Pauli matrices $\frac{\sigma_i}{2}$, invoking three gauge fields W_μ^i (i=1,2,3).
  - $U(1)_Y$ is an abelian group with hypercharge (Y) as the associated quantum number. With a single generator, the associated gauge field is the field B_μ.

The weak hypercharge (Y) and the electric charge (Q) (generator of the symmetry group for electromagnetism $U(1)_{em}$) are connected with isospin as

$$Q = I_3 + \frac{Y}{2} \tag{1.6}$$

Under the above groups, the fermion fields can be seen as consisting of a left-handed doublet and right-handed singlet. $SU(2)_L$ generators acts only on the left-handed field, while the $U(1)_Y$ generator acts on both left-handed and right-handed components. Thus, the gauge transformations for the electroweak theory are given as

$$L \to e^{-i\alpha_i(x)\frac{\sigma_i}{2} - i\beta(x)\frac{Y}{2}} L, \quad \psi'_R \to e^{-i\beta(x)\frac{Y}{2}} \psi'_R \tag{1.7}$$

where $L = \begin{pmatrix} \psi_L \\ \psi'_L \end{pmatrix}$ is a $SU(2)_L$ doublet and ψ_R *and* ψ'_R are the singlets, which introduces new covariant derivatives

$$Doublet: \quad D_\mu = \partial_\mu + ig_w \frac{\sigma_i}{2} W_\mu^i + ig\frac{Y}{2} B_\mu, \tag{1.8}$$

$$Singlet: \quad D_\mu = \partial_\mu + ig\frac{Y}{2} B_\mu, \tag{1.9}$$

The corresponding gauge fields associated with the coupling constants g_w (for $SU(2)_L$ interactions) and g (for $U(1)_Y$ interactions) transforms as

$$W_\mu^i \to W_\mu^i + \alpha^j(x)\varepsilon^{ijk}W_\mu^k + \frac{1}{g_w}\partial_\mu \alpha^i(x), \quad B_\mu \to B_\mu + \frac{1}{g}\partial_\mu \beta(x) \tag{1.10}$$

Hence the full electroweak lagrangian is given as

$$\mathscr{L}_{EW} = \bar{L} i\gamma^\mu D_\mu L + \bar{\psi'}_R i\gamma^\mu D_\mu \psi'_R - \frac{1}{4} W_i^{\mu\nu} W_{\mu\nu}^i - \frac{1}{4} B^{\mu\nu} B_{\mu\nu} \tag{1.11}$$

implying summation over all doublets and singlets in equation 1.11, the first two terms are the free Dirac lagrangian for the left-handed and right-handed fermions respectively, with last two terms representing the gauge-invariant kinetic terms of

field strength tensors given by

$$W^i_{\mu\nu} = \partial_\mu W^i_\nu - \partial_\nu W^i_\mu - g_w \varepsilon^{ijk} W^j_\mu W^k_\nu, \quad B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu \tag{1.12}$$

The electroweak lagrangian also do not include any mass terms due to gauge invariance.

1.2 The Higgs mechanism

The Brout-Englert-Higgs (BEH) mechanism is a way of generating masses of vector gauge bosons of the electroweak theory [4–6]. The mechanism introduces the gauge-invariant $SU(2)_L$ doublet complex scalar field to break the electroweak symmetry

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi^1 + i\phi^2 \\ \phi^3 + i\phi^4 \end{pmatrix} \tag{1.13}$$

So the Higgs field introduced to the SM lagrangian is

$$\mathscr{L}_{BEH} = (D^\mu \phi)^\dagger (D_\mu \phi) + V(\phi^\dagger \phi) \tag{1.14}$$

with D_μ given by equation 1.8 and the scalar potential term introduced is

$$V(\phi^\dagger \phi) = \mu^2 \phi^\dagger \phi + \lambda (\phi^\dagger \phi)^2 \tag{1.15}$$

The ground state for the potential of equation 1.15 is given by

$$\phi^\dagger \phi = \frac{-\mu^2}{2\lambda} \tag{1.16}$$

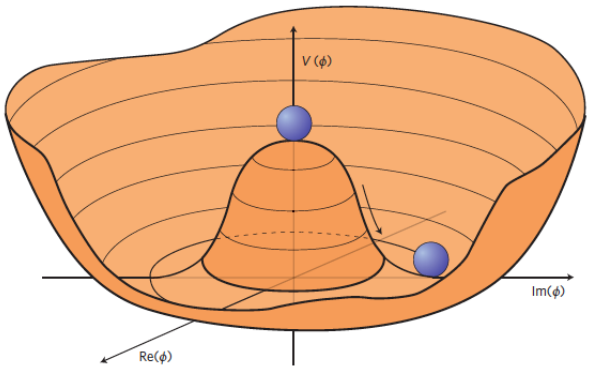


Figure 1.2: Shifting of the vacuum state from zero to a non-zero value on introducing Higgs field [7].

If μ is set to zero, then the minima would be zero which is the usual case with no symmetry breaking. Keeping $\lambda > 0$ for the square term and $\mu^2 < 0$, then the minima of the potential becomes a circle with radius given by vacuum expectation value (v) (see fig 1.2). Keeping the state on ϕ_0 axis, the vacuum state of equation 1.13 reduces to

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v+h \end{pmatrix}, \quad v^2 = \frac{-\mu^2}{\lambda} \tag{1.17}$$

where perturbations around the displaced vacuum v are represented by h. With this vacuum, one can see that this field breaks the $SU(2)_L \times U(1)_Y$ symmetry as

$$SU(2)_L: \quad \sigma_1\phi = \begin{pmatrix}0&1\\1&0\end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix}0\\v+h\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}v+h\\0\end{pmatrix} \neq 0 \tag{1.18}$$

$$\sigma_2\phi = \begin{pmatrix}0&-i\\i&0\end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix}0\\v+h\end{pmatrix} = \frac{-i}{\sqrt{2}}\begin{pmatrix}v+h\\0\end{pmatrix} \neq 0 \tag{1.19}$$

$$\sigma_3\phi = \begin{pmatrix}1&0\\0&-1\end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix}0\\v+h\end{pmatrix} = \frac{-1}{\sqrt{2}}\begin{pmatrix}0\\v+h\end{pmatrix} \neq 0 \tag{1.20}$$

$$U(1)_Y: \quad Y\phi = \begin{pmatrix}0&0\\0&1\end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix}0\\v+h\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\v+h\end{pmatrix} \neq 0 \tag{1.21}$$

So the operators of the electroweak symmetry gives non-zero doublets when acted on the Higgs doublet (equation 1.17), hence these symmetries are broken; the four gauge bosons of electroweak symmetry (W_1, W_2, W_3, B) would get masses via the Higgs mechanism. However, there exists one symmetry $U(1)_{em}$ which is not broken, which is the electromagnetic symmetry with operator

$$Q = \frac{1}{2}(\sigma_3 + Y) \tag{1.22}$$

$$\implies Q\phi = \begin{pmatrix}1&0\\0&0\end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix}0\\v+h\end{pmatrix} = 0 \tag{1.23}$$

where Q is the electric charge of the particle. This is expected because the vacuum is electrically neutral (Q=0). This implies that instead of four broken symmetries (as shown in equations 1.18-1.21), there exists one unbroken symmetry so instead of 4, there would be 3 massive particles and one of them would remain massless. Substituting $\phi(x)$ from equation 1.17 into equation 1.14 and the covariant derivative operator and expanding the terms, we get

$$\mathscr{L}_{BEH} = \mathscr{L}_H + \mathscr{L}_{VV} + \mathscr{L}_{HVV} + \mathscr{L}_{HH} \tag{1.24}$$

where

$$\mathscr{L}_H = \frac{1}{2}\partial_\mu h \partial^\mu h + \mu^2 h^2 \tag{1.25}$$

$$\mathscr{L}_{VV} = \frac{g_w^2 v^2}{4} W_\mu^- W^{+\mu} + \frac{g_w^2 v^2}{8cos^2\theta_w} Z_\mu Z^\mu \tag{1.26}$$

$$\mathscr{L}_{HVV} = \frac{g_w^2}{4}(vhW_\mu^- W^{+\mu} + h^2 W_\mu^- W^{+\mu}) + \frac{g_w^2}{8cos^2\theta_w}(vhZ_\mu Z^\mu + h^2 Z_\mu Z^\mu) \tag{1.27}$$

$$\mathscr{L}_{HH} = \frac{\mu^2}{v}h^3 + \frac{\mu^2}{4v^2}h^4 \tag{1.28}$$

- $\mathscr{L}_H$: This part of lagrangian defines the massive scalar boson with spin 0 namely the Higgs boson with mass given by

$$m_H = \sqrt{2}|\mu| \tag{1.29}$$

  Note that this mass is not predicted by the SM, because μ is actually a free parameter. So the mass of Higgs boson needs to be constrained experimentally.

- $\mathscr{L}_{VV}$: This part of lagrangian defines the mass terms of the gauge bosons. The breaking of the electroweak symmetry leads to this mass terms with the $W^{\pm}$, Z, A_μ bosons obtained by mixing of the four gauge fields of the electroweak symmetry as:

$$W_\mu^{\pm} = \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2) \tag{1.30}$$

$$Z_\mu = W_\mu^3 cos\theta_w - B_\mu sin\theta_w \tag{1.31}$$

$$A_\mu = W_\mu^3 sin\theta_w + B_\mu cos\theta_w \tag{1.32}$$

$$\tag{1.33}$$

  with the weak mixing angle for the $SU(2)_L$ field and $U(1)_Y$ field given as

$$cos\theta_w = \frac{g_w}{\sqrt{g_w^2+g^2}}, \quad sin\theta_w = \frac{g}{\sqrt{g_w^2+g^2}} \tag{1.34}$$

  Thus from this part of lagrangian, the masses of the vector bosons ($W^{\pm}$ and Z) can be given as

$$m_{W^+} = m_{W^-} = \frac{g_w v}{2}, \quad m_Z = \frac{g_w v}{2cos\theta_w} \tag{1.35}$$

  Photon field A_μ still remains massless.

- $\mathscr{L}_{HVV}$: This part of lagrangian represents the trilinear coupling terms of Higgs boson with the weak vector bosons. The coupling constants also define the decay widths during the decay of Higgs to these vector bosons.

- $\mathscr{L}_{HH}$: This part of the lagrangian contains trilinear and quadrilinear terms describing Higgs self couplings.

The Higgs mechanism can be extended to explain masses of fermions as well. This is done by introducing the Higgs interactions with the fermions terms as

$$\mathscr{L}_{Yukawa} = -\lambda_f(\bar{\psi}_L \phi \psi_R + \bar{\psi}_R \bar{\phi} \psi_L) \tag{1.36}$$

where $\psi_L and \psi_R$ are the left-handed and right-handed components of the fermion field. Before electroweak symmetry breaking, the mass terms were not allowed for fermions because $\bar{\psi}_L\psi_R$ or $\bar{\psi}_R\psi_L$ terms are not gauge invariant under electroweak symmetry. After breaking of the symmetry, the Yukawa terms of lagrangian can be expanded as

$$\mathscr{L}_{Yukawa} = \sum_f \frac{-\lambda_f(v+h)}{\sqrt{2}}(\bar{f}_L f_R + \bar{f}_R f_L) \tag{1.37}$$

$$= \sum_f \frac{-\lambda_f v}{\sqrt{2}}\bar{f}f + \sum_f \frac{-\lambda_f}{\sqrt{2}}h\bar{f}f \tag{1.38}$$

where the summation is carried out over all fermions. The first and second terms of equation 1.38 shows the mass terms and coupling interaction of fermion and higgs boson respectivey, the couplings proportional to the mass of the fermions. Note that the Yukawa coupling λ_f is a free parameter, so it does not actually predict the mass of the fermions. Also equation 1.38 is a simplified version to explain the higgs couplings, because it does

not include the effect of fermion mixing (similar to the weak mixing of gauge bosons) which is very evident in case of quarks [8, 9].

Because there is no observation of a right-handed counterpart of neutrino (or left-handed counterpart of anti-neutrino), neutrino masses (as been observed in experiments [10]) cannot be explained or visualized by Higgs mechanism.

1.3 Interactions of Higgs boson

In previous section, the Higgs mechanism was discussed where a complex scalar Higgs field was introduced generating mass terms for the gauge bosons and fermions by spontaneously breaking the electroweak symmetry. This field can be represented by a massive particle called the Higgs boson, which interacts with SM particles with interaction strength dependent on particle's mass. So, the interactions are maximum with top quarks and massive electroweak bosons. Interactions with very light particles like neutrinos are not observed yet.

1.3.1 Higgs boson production

Since the Higgs boson interacts with most of SM particles, it can be produced in pp collisions via a variety of processes [11]. The mass of the Higgs boson is a free parameter, which has been constrained to a value of 125 GeV by precise measurements at LHC. The couplings of Higgs boson with fermions (mass m_f) and gauge bosons (mass m_V) varies as m_f and m_V^2 respectively. Hence, the Higgs couplings are preferable to heavier particles namely the W and Z bosons, top quarks and these are major contributors in Higgs production. Feynman diagrams of the four major production mechanisms for the Higgs boson are illustrated in fig 1.3.

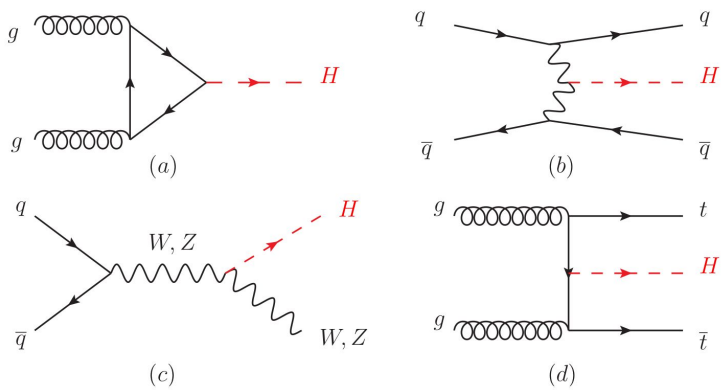


Figure 1.3: Major production modes of Higgs boson : (a) gluon-gluon fusion (ggF) (b) Vector boson fusion (VBF) (c) Associated boson production or HiggsStrahlung (VH) (d) Associated top-antitop pair production ($t\bar{t}H$) [12].

The cross-sections of processes of fig 1.3 for $\sqrt{s}$ =13 TeV are illustrated in fig 1.4.

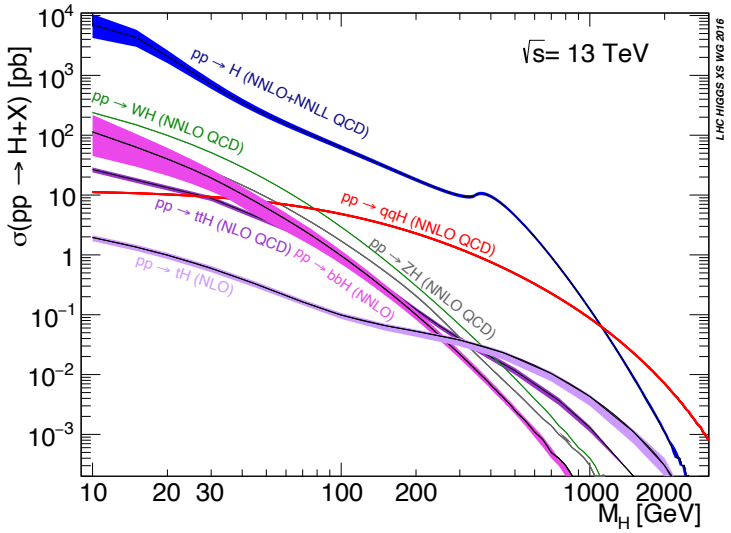


Figure 1.4: Variation of production cross-section with m_H for the major production channels of Higgs boson at $\sqrt{s} = 13\,TeV$. The coloured bands about the curves are the associated uncertainties [13].

Process	QCD/EW corrections	$\sigma_{\mathbf{prod}}(\mathbf{pb})$
ggF	N3LO QCD+NLO EW	48.61
VBF	NNLO QCD+NLO EW	3.766
WH	NNLO QCD+NLO EW	1.358
ZH	NNLO QCD+NLO EW	0.880
$t\bar{t}H$	NLO QCD+NLO EW	0.4987

Table 1.1: Production cross-sections for the major production channels of Higgs boson at $\sqrt{s} = 13\,TeV$ for $m_H = 125\,GeV$ [13].

.

Table 1.1 shows the cross-section values for $m_H = 125\,GeV$.

- **gluon-gluon fusion** : The gluon-gluon fusion process (ggF) is induced by the fusion of two gluons into a Higgs boson mediated by a virtual heavy quark loop of top quarks or bottom quarks (Fig 1.3a). Gluons are available in large amounts in pp collisions, and the high luminosity and $\sqrt{s}$ enhances it further, thereby making ggF as the most dominating production mode. Leading order (LO) cross-section can be improved by a magnitude of 2-3 by having higher QCD corrections. QCD corrections at Next-to-Next-to-Next-to-Leading-Order (N3LO) and electroweak corrections at Next-to-Leading-Order (NLO) are computed for the cross-section as shown in [14].

- **Vector boson fusion** : The vector boson fusion (VBF) is the second favorable production mode, albeit with cross-section less than ggF by a factor of approximately 10. At leading order, the process is expressed as $qq' \rightarrow qq'H$ where vector bosons are radiated off by the quarks which merge to form the Higgs boson (Fig 1.3b). The experimental signature of this process is the observation of two forward (high pseudorapidity) energetic jets of the quarks interacting with the Higgs boosted in the central region of the detector. QCD corrections for this process are also large.

- **HiggsStrahlung** : HiggsStrahlung, or associated vector boson production (VH) is the third prominent production mode, with a Higgs boson radiated off the vector boson (Fig 1.3c). Because of this, the Higgs decay products are boosted accompanied by products of associated on-shell vector boson.

- **Associated $t\bar{t}$ pair production** : Associated production with top quark pairs has much lower cross-section compared to the dominant ggF mode, but is the only process to probe the Higgs couplings to quarks. The $t\bar{t}$ pair generated alongside Higgs (Fig 1.3d) can decay in various possible modes, leading to many experimental signatures.

1.3.2 Higgs boson decay

Because the Higgs boson couples with almost every particle of SM having mass and can couple with massless particles as well via intermediate particle loops of heavier particles, it can decay into a variety of final states. The Higgs boson decay width and branching ratios of the different modes can be analysed as a function of Higgs mass m_H and associated coupling at the decay vertex. Figure 1.5 represent the branching ratios for Higgs decay modes as a function of the Higgs boson mass.

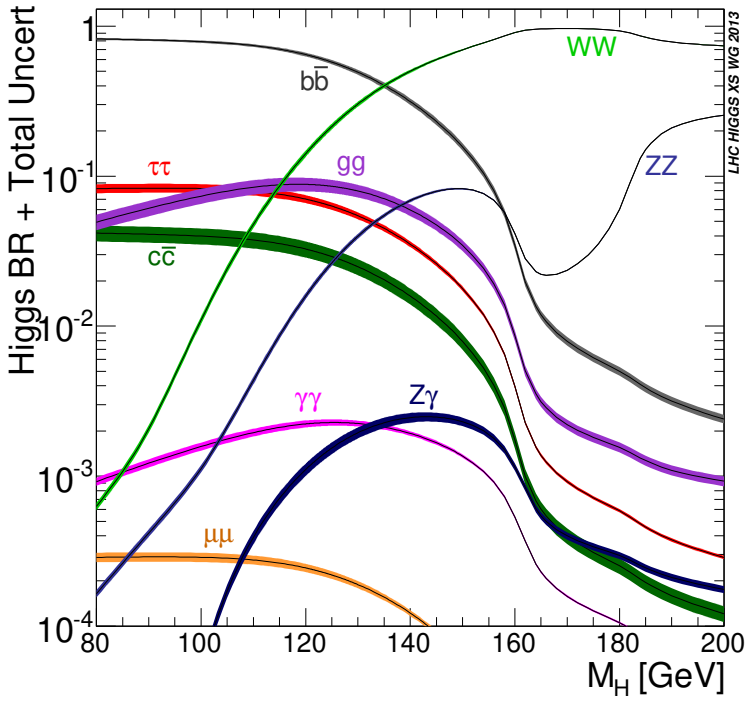


Figure 1.5: Higgs boson decay branching ratios for different values of Higgs boson mass [13].

.

For Higgs boson in the low mass regime, the decay to $q\bar{q}$ pairs are favoured, especially the decay $H \to b\bar{b}$. Although decay to bottom quarks is most probable, it is heavily dominated by the QCD background production of bottom quarks.

Experimentally, the decay of Higgs to a photon pair is studied in searches of low mass Higgs despite its smaller branching ratio; this is because of a cleaner signal signature which is observed as the Higgs resonance peak on top of the QCD background of photons coming from decay of neutral pions copiously produced in jets. This decay channel is mediated by a virtual loop of weak gauge bosons or top quarks. A similar counterpart of this decay mode, the decay to gluons which involves similar virtual loops although has a relatively larger BR but is faced with even more QCD background interference because the gluons produced are transformed to other particles via QCD interactions.

The decay to weak gauge vector bosons in the low mass regime involves an on-shell off-shell boson pair because $m_H < 2m_V$. The W and Z bosons respectively can decay into a $\ell\nu$ and $\ell^+\ell^-$ pair respectively, but due to invisible neutrinos the mass resolution in H→WW is poorer than in H→ZZ. The main focus for this thesis is going to be the decay channel $H \to ZZ^* \to 4\ell$, because unlike hadronic channels, leptonic channels do not face contamination from QCD background. Also, CMS and ATLAS detectors provide an excellent resolution in measurement of momentum of electrons and muons, which

gives a pretty good mass resolution for the reconstructed Higgs boson candidates.

1.4 Higgs effective field theory extension

Since the discovery of the Higgs boson at LHC [15, 16], its properties are being extensively studied evidently making it to be the scalar Higgs boson as predicted by the SM [17]. Complementary to that, dedicated studies are also being carried out to look for deviations in the interactions of Higgs with SM particles. So far, no such direct evidence for these deviations have been found from data [18, 19]. Under such circumstances, more generic theoretical framework, known as effective field theories (EFTs), are being developed to assist searching for new physics at the LHC. EFT introduces higher dimensional operators made from the usual SM operators which are valid upto an energy scale Λ. These operators introduced by EFT are suppressed in powers of Λ to keep the theory renormalizable. EFTs are studied extensively to study modifications of SM, however as it is bounded by an energy scale it fails to give valid results beyond that scale. The general expression of the EFT lagrangian is given by

$$\mathscr{L}_{eff} = \mathscr{L}_{SM} + \sum_i \frac{c^{(5)}}{\Lambda} O_i^{(5)} + \frac{c^{(6)}}{\Lambda^2} O_i^{(6)} + \dots \tag{1.39}$$

Each of the $O_i^{(D)}$ are D dimensional operators invariant under the full SM symmetry $SU(3) \times SU(2) \times U(1)$. These operators define the modified interactions of the EFT which are parametrised by parameters $c_i^{(D)}$ suppressed by energy scale Λ depending on the dimension D. These are called Wilson coefficients, which are constrained based on experimental data and act as tools to model BSM scenarios. The main idea behind introducing EFTs is to find parameters which models a class of general interactions, which can be used to modify BSM parameters depending on the model considered. So this results in constraining the parameters once, instead of repeating the procedure for different BSM models modifying the same interaction. Contribution from each operator to the physical processes scales as $(v/\Lambda)^{D-4}$, and by construction $(v/\Lambda) < 1$, so generally EFT effects are small modifications to the SM interactions.

Set of these higher dimensional operators which can be constructed from SM operators are theoretically derived and known [20, 21]. Out of the higher dimensional operators, new physics effects are expected to be from dimension six operators (D=6), and any higher contributions coming from $D \geq 8$ are suppressed by factors of $(v/\Lambda)^4$ so there effects can be neglected from analysis.

In this thesis, specific focus is being made on the Higgs effective lagrangian (HEL) using Feynrules and Madgraph as described in [22, 23]. The HEL has around 37 linearly independent operators which modifies the couplings of Higgs boson to the leptons, quarks, gauge vector bosons and Higgs itself. Explaining all the operators of the EFT lagrangian is beyond the scope of this thesis, so focus would be kept only on the EFT operators which affect the couplings of Higgs boson and the weak vector bosons, and table 1.2 lists some of those relevant operators.

These operators are constrained from the experimental observations of the VBF and VH production modes and Higgs decay to vector bosons. Because the Higgs couples to photons via weak gauge bosons, a modification in photon couplings also has an effect on couplings with weak gauge bosons. Hence, operators O_W and O_B both affects the couplings with weak gauge bosons. Experimentally, the operators O_W and O_B are constrained

in their linear combinations $O_W + O_B$ and $O_W - O_B$ because of this net effect. The former combination has been constrained from precise eletroweak measurements [24]. Individually, they can be constrained by either keeping one of them to zero or a fixed small value.

Operator	Coefficient	HEL constraint
$O_{HW} = i(D^\mu\phi)^\dagger\sigma^a(D_\nu\phi)W^a_{\mu\nu}$	$\frac{c_{HW}}{\Lambda^2} = \frac{2g}{m_W^2}$cHW	-0.035 < cHW < 0.015
$\tilde{O}_{HW} = i(D^\mu\phi)^\dagger\sigma^a(D_\nu\phi)\tilde{W}^a_{\mu\nu}$	$\frac{c_{H\tilde{W}}}{\Lambda^2} = \frac{g}{m_W^2}$tcHW	|tcHW| < 0.060
$O_{HB} = i(D^\mu\phi)^\dagger(D_\nu\phi)B_{\mu\nu}$	$\frac{c_{HB}}{\Lambda^2} = \frac{g'}{m_W^2}$cHB	-0.045 < cHB < 0.075
$\tilde{O}_{HB} = i(D^\mu\phi)^\dagger(D_\nu\phi)\tilde{B}_{\mu\nu}$	$\frac{c_{H\tilde{B}}}{\Lambda^2} = \frac{g'}{m_W^2}$tcHB	|tcHB| < 0.24
$O_W = i(\phi^\dagger\sigma^a\overleftrightarrow{D}^\mu\phi)D^\nu W^a_{\mu\nu}$	$\frac{c_W}{\Lambda^2} = \frac{g}{m_W^2}$cWW	-0.035 < cWW - cB < 0.005
$O_B = i(\phi^\dagger\overleftrightarrow{D}^\mu\phi)\partial^\nu B_{\mu\nu}$	$\frac{c_B}{\Lambda^2} = \frac{g'}{2m_W^2}$cB	-0.0033 < cWW + cB < 0.0018

Table 1.2: Dimension-6 HEL operators which affect the couplings of the Higgs boson and the weak vector bosons and the 95% confidence level constraints on the wilson coefficients expressed in terms of the HEL parameters. ϕ is the Higgs doublet, D_μ is the covariant derivative. There are also other operators which specifically affect the weak boson bilinear and trilinear interactions but can be ignored from the analysis of this thesis [23].

1.5 Motivation for this thesis

EFTs have been studied in searching for new physics phenomena which is currently not explained by the SM at the LHC [14]. Experimental data has been used to constrain the parameters of the HEL operators by studying the Higgs production and decay modes wherein the kinematic distributions of the final states are studied to reconstruct the entire event [25–27]. The kinematic distributions of the interacting particles and their decay products can be modelled by density distributions parametrised by the EFT parameter. Hence, a parametric estimation can be carried out by modelling the kinematic distributions as probability density functions (pdf), therein constraints can be put from measurement of confidence intervals for the estimated parameter given a level of significance.

Probability density functions from data are modelled via likelihoods. Likelihoods for kinematic observables with respect to EFT parameters can be evaluated as

$$p(x|\theta) = \int p(x,z|\theta)dz = \int p(x|z)p(z|\theta)dz \tag{1.40}$$

where x are the kinematic observables after parton shower, detector interactions and reconstruction, θ is the EFT theoretical parameter and z are the parton-level momenta [28]. From the two functions, the function $p(z|\theta)$ consists of the matrix element of the corresponding process which can be evaluated for arbitary z as done in Monte-Carlo (MC) event generators like MadGraph. However, the other component $p(x|z)$ can be decomposed into contributions from parton shower (z_{shower}) and detector ($z_{detector}$) as

$$p(x|z) = \int p(x|z_{detector})p(z_{detector}|z_{shower})p(z_{shower}|z)dz_{shower}dz_{detector} \tag{1.41}$$

The phase space for calculation of $p(x|z)$ is huge with millions of random numbers.

This makes its calculation intractable, making the calculation of the entire likelihood intractable.

In order to tackle the integration problem, algorithms are developed to model the pdfs by analysing the inter correlations among the data points. A better estimate of the pdf can be obtained as more data would be analysed. This can be achieved by using flexible learning algorithms like machine learning (ML) and deep learning (DL). The flexibility of the deep learning models is what makes them reliable for fitting functions to a set of data, as compared to fitting based on a fixed polynomial or a known analytical function. One of the key features of this study is to develop the DL methodology which can use a limited amount of information like the detector effects only to constrain EFT parameters which modifies the interactions at the parton level without worrying about the large phase space of shower components. This can make the computations much faster and also makes the likelihood integration tractable.

In this thesis, emphasis is laid on developing one such deep learning model based on normalizing flows [29]. As a check for constraining the EFT parameters, an EFT interaction mediated by the coefficient cWW (see table 1.2) is studied which modifies the couplings of Higgs and Z boson affecting the decay width of the decay channel. For developing strategy for this analysis, all other parameters are set to zero.

One of the key aspects of the applicability of the model is to relate detector level observations to a parton level phenomena modelled by the EFT. Given a set of kinematic variables of the final states observed in the detector, the model fits it to a probability function parametrized by the scalar EFT parameter, thereby it gives an estimation to the quantity $p(x|z)$ without having the information of the shower effects. With the parton level term $p(z|\theta)$ obtained from MC generators and is dependent of EFT parameter, this simplification of $p(x|z)$ makes the calculation of likelihood of equation 1.40 tractable.

Chapter 2

Event simulation and analysis

The Large Hadron Collider (LHC) is a large two-ring superconducting accelerator and collider which accelerates protons to an energy of 6.5 TeV and collides them at $\sqrt{s}$ = 13 TeV. It was designed to test the existence of Higgs boson by probing the scalar sector over a mass window approximately in the vicinity of 125 GeV, and to search for new physics phenomena not explainable by the SM. Having the Higgs discovered in 2012, detailed analysis of the Higgs boson interactions with other particles is carried out in search of new physics phenomena. The LHC collides pp bunches at a rate to 40M collisions per second which needs to be processed by state-of-art detectors and interesting events be stored offline for further analysis. The CMS and ATLAS detectors are two of these detectors located at the collision points of LHC to study the events for SM precision measurements and for searching BSM physics.

2.1 The CMS detector

The CMS detector is a large multi-layered cylindrical detector with each layer dedicated to perform a specific function. It mainly consists of a tracker, electromagnetic and hadronic calorimeters, superconducting magnet and muon chambers (see fig 2.1). The detector is meant to identify the particles produced, measure their kinematic properties like energy, momentum, angular variables, etc and combine this information to reconstruct the full collision event. The beam-axis is selected as the z-axis with the x-axis pointing towards the center of the ring and y-axis pointing in the vertically upward direction. The collision happens at the interaction point, about which one measures the azimuthal angle ϕ from the x-axis in the x-y plane and the polar angle θ is measured from the beam axis. Instead of θ, the Lorentz invariant pseudorapidity $\eta = -log_e(tan(\frac{\theta}{2}))$ is preferred. The Lorentz invariant quantity, transverse momentum defined as $p_T = \sqrt{p_x^2 + p_y^2}$ as the net momentum in the plane transverse to the beam axis is measured to analyse the events.

The tracker system consists of silicon chips which keeps a record of trajectory of the particle which constitutes the track of the particle. Tracks are used to meaure momentum of particles, and to reconstruct collision vertices. Enveloping the tracker is the scintillating type electromagnetic calorimeter made up of lead tungstate to measure the energy of the charged electrons and photons. Beyond ECAL is another calorimeter made up of brass and plastic scintillators called hadronic calorimeter for measuring energy of hadrons, mesons, etc. Many of such particles can also deposit small amounts of energy in the ECAL. Muons within the detector length are stable and they do not bremsstrahlung,

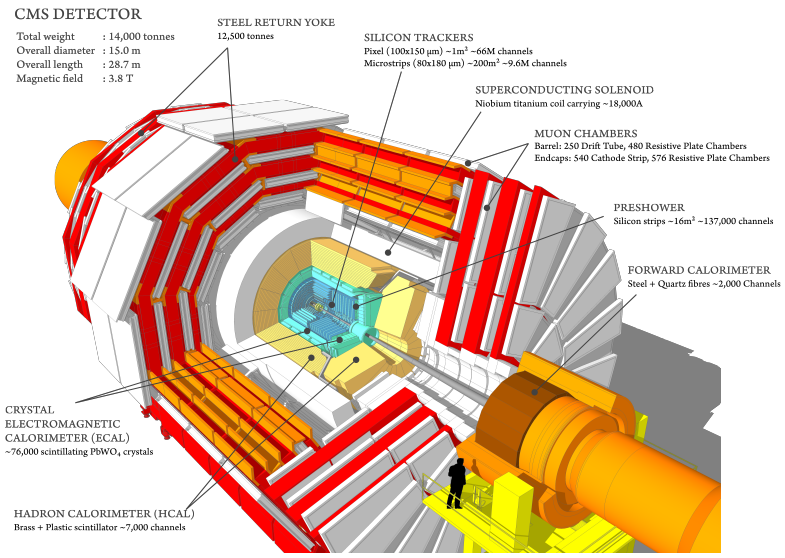


Figure 2.1: A schematic diagram of the CMS detector illustrating its major components namely the tracker system, calorimeters, superconducting magnet and muon chambers [30].

.

hence muons pass without showering through the calorimeters. In order to get very precise muon momentum resolution, its momentum is measured in tracker as well as in muon chambers outside the reach of superconducting coil.

A detailed description of the CMS detector along with the coordinate system used for the kinematic analysis of the collision events can be found here [31].

2.2 Delphes simulation

To study the pp collisions at LHC, highly sophisticated detectors like CMS, ATLAS, etc are designed to precisely measure the particles originating from these collisions. For developing the analysis strategies, we rely on Monte-Carlo event generators simulating the pp collisions, and the interaction of particles entering detector with the detector material fully simulated using GEANT4 package [32] for high accuracy. However, the full simulation is computationally expensive and can be handled only by large collaborations. Hence in order to make use of the limited computing resources for analysis of large MC event samples, fast-simulation techniques have been devised by LHC which parametrizes the detector geometry to simplify the computations. Delphes [33, 34] is one such C++ based fast detector simulation framework developed to achieve this task.

The layout of the detector simulation performed by Delphes is shown in fig 2.2. It consists of a central tracking system used to reconstruct the tracks of the particles and measure their transverse momentum via a centralised solenoidal magnetic field (not shown in fig 2.2). After the tracker are the central electromagnetic (ECAL) and hadronic (HCAL) calorimeters, along with a forward calorimeter (FCAL). The calorimeters measure the energy deposited by the particles, where electrons and photons leave their energy in ECALs, charged and neutral hadrons in HCAL, and some particles like kaons in both ECAL and HCAL. Particles like muons and neutrinos do not deposit their energies in the calorimeters. The central detector volume (consisting of the tracker and calorimeter) is enclosed by a muon chamber to measure the transverse momentum of muons.

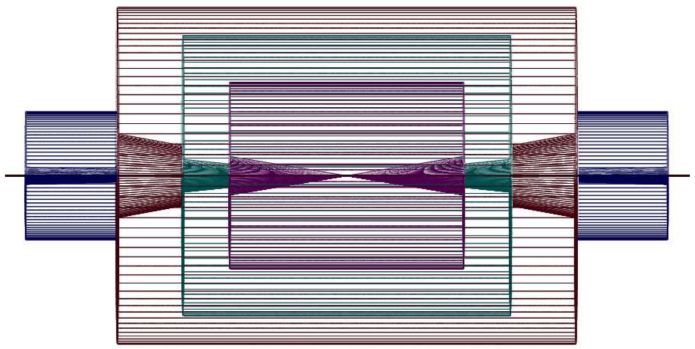


Figure 2.2: Profile layout of the simulated detector in Delphes showing its majour components namely the tracker (pink),central calorimteters (green) and endcap calorimeters (blue) and muon chambers (red) [33].

.

Delphes, though quite effective in simulating the detector response has certain limitations:

- The detector geometry is idealised. All the components are symmetrical and uniform about the beam axis. It cannot simulate physical defects like cracks, defective materials,etc.
- It neglects secondary detector interactions, multiple particle scatterings, bremsstrahlung, etc.

In order to make the Delphes simulation reliable to tackle the above limitations, the corresponding reconstruction efficiencies of Delphes needs to be validated with CMS observations.

2.3 Data samples

For analysing the golden decay channel H $\rightarrow$ ZZ* $\rightarrow 4\ell$ channel, we need to apply certain selection rules on the events of the signal and background samples to reduce as much of the background as possible in the signal region. To study the kinematic invariant mass distributions for the cWW, we study two sets of data samples for the same. The details of the samples analysed are shown in table 2.1 and table 2.2. Note that in the thesis the samples for which full detector simulation using GEANT4 package in full CMS event reconstruction framework is carried out are termed as full simulation samples, while the samples for which fast-simulation is carried out using Delphes are termed as Delphes samples.

Sample Type	**Sample Name**	$\sigma \times$ **BR(infb)**	**Total number of events**
Signal	GluGluToHToZZTo4L	12.1885	999800
Background	ZZTo4L	1256	6669988

Table 2.1: The full simulation samples for signal and background used for the analysis. The samples provided are passed through the HZZ algorithm (section 2.5) to obtain permissible events for signal and background in the signal region 118 GeV $< m_{4l} <$ 130 GeV. Each of the samples would later be scaled as per the luminosity considered.

The full simulation samples are the data samples used for the analysis of the golden decay channel as mentioned in this reference [26]. For generating the signal samples,

POWHEG V2 [35] was used for the main production mode gluon-gluon fusion ($gg \rightarrow H$), while the decay process $H \rightarrow ZZ^* \rightarrow 4\ell$ is modelled using JHUGEN generator [36]. Showering of parton-level events is done using PYTHIA 8.209 [37], and in all cases matching is performed by allowing QCD emissions at all energies in the shower and vetoing them afterwards according to the POWHEG internal scale. All samples are generated with the NNPDF 3.0 NLO parton [38] distribution functions (PDFs). For background samples, the POWHEG v2 generator was used keeping the same settings as for the Higgs signal to generate the SM $ZZ \rightarrow 4\ell$ samples obtained from quark-quark annihilation.The full simulation samples mentioned above correspond to the SM case, i.e. cWW = 0.

To take into account non-zero cWW values for the EFT analysis, signal samples (with non-zero cWW) and background samples (with cWW = 0) were generated using the Higgs Effective Lagrangian as described here [22]. Using the UFO model [39], MC samples were generated using Madgraph, parton showering done using Pythia and detector simulation using Delphes.

Sample type	**cWW**	**Total number of events**
Signal	> 0	100000
Background	$= 0$	500000

Table 2.2: The Delphes samples for signal and background events generated using Madgraph, Pythia and Delphes. The signal events are generated using non-zero cWW values while the background corresponding to the SM case is generated at cWW = 0.

The following table lists the cWW values used for making the cWW samples

0.00231	0.00307	0.00536	0.00596	0.00721	0.00952	0.01093	0.01761
0.01837	0.02113	0.02176	0.02196	0.02238	0.02514	0.02753	0.02898
0.03256	0.03257	0.033	0.03409	0.03597	0.0439	0.04441	0.04473
0.04687	0.04781	0.04818	0.05096	0.05427	0.05492	0.05653	0.05704
0.06041	0.06048	0.06052	0.06068	0.06086	0.06102	0.06566	0.06664
0.06863	0.07019	0.07384	0.07889	0.07948	0.08154	0.08315	0.09567
0.09963	0.10076	0.1047	0.11241	0.11404	0.11519	0.11585	0.12099
0.12244	0.12332	0.12384	0.12446	0.12465	0.12514	0.13007	0.1332
0.14245	0.14282	0.14429	0.14665	0.15203	0.15571	0.15672	0.16213
0.17031	0.17319	0.17563	0.17784	0.18037	0.18084	0.18156	0.18205
0.18345	0.18744	0.18936	0.1896	0.19272	0.19387	0.19807	

Table 2.3: The non-zero cWW values considered for the EFT samples.

For the analysis purposes, both the samples (full simulation and Delphes) were used. We make use of the lowest non-zero cWW sample (here cWW = 0.0023) as a proxy for the SM case.

2.4 Event reconstruction

For analysing the HZZ decay channel, precise reconstruction of events is required from the final lepton states obtained. A high lepton selection efficiency is required over a wide range of momenta in order to maximize the sensitivity of Higgs mass in the vicinity of 125 GeV. The CMS particle flow (PF) algorithm [40] combines the information of the subdetectors (namely the tracker, calorimeters, muon chambers, etc) to characterise the event in form of reconstructed particle candidates. The lepton reconstruction as per the PF algorithm is carried out as follows:

- Electron candidates reconstructed are required to have a transverse momentum $p_T^e > 7\,GeV$ which are required to be within the geometrical acceptance region $|\eta| < 2.5$. For electron reconstruction, information is obtained both from the ECAL and the tracker. Resolution for electron momenta measurements are favoured more from ECAL measurements for high pT electrons and from tracker measurements for low pT electrons [25].

- Muon reconstruction is carried out based on the information obtained from the particle tracker and the muon system. Muons reconstructed are required to have a transverse momentum of $p_T^\mu > 5\,GeV$ and in the geometric acceptance region $|\eta| < 2.4$.

- In addition to muons and electrons, tau leptons are also possible leptons obtained from the Z decay (Z boson has a decay chance of approximately 3% per leptonic flavour). However, a tau lepton is not directly detected because it decays before reaching the detector. Tau leptons can decay leptonically or hadronically, with a larger branching fraction for hadronic products [41]. The leptonic decay of tau lepton involves first decay to a tau neutrino and an off-shell W boson (represented as $(W)^*$ below) which decays leptonically.:

$$\tau^- \to (W^-)^* \nu_\tau \to \ell^- \bar{\nu}_\ell \nu_\tau$$
$$\tau^+ \to (W^+)^* \bar{\nu}_\tau \to \ell^+ \nu_\ell \bar{\nu}_\tau$$

  where ℓ corresponds to electron or muon. Because the leptonic tau decays are less probable (leptonic decays correspond to 35% of the tau decays), interesting events from tau leptons are less in number. Also due to neutrinos being invisible, it affects the resolution of reconstructed Higgs.

- **Relative isolation selection** : The leptons of the ZZ candidates are required to be well isolated in order to differentiate between the leptons coming from the decay of Z from the ones coming from jets or decays of in-flight hadrons like pions. The relative isolation parameter [26] is defined as :

$$R_{Iso}^\ell = \frac{(\sum p_T^{charged} + max[0, \sum p_T^{neutral} + \sum p_T^\gamma - p_T^{PU}(l)]}{p_T^\ell} \tag{2.1}$$

  Leptons forming the Z candidates are required to satisfy $R_{Iso}^l < 0.35$.

2.5 Event selections : HZZ algorithm

The HZZ algorithm defines a set of selection rules being applied to the reconstructed leptons defining the Higgs event to extract the intermediate Z candidates from which the leptons are obtained. For a Higgs mass of 125 GeV, we see that $m_H < 2m_Z$, so it is natural for one of the Z bosons to be off-shell. So we define the Z candidate with mass closest to the on-shell SM Z boson mass (roughly 90 GeV) as Z_1 while the other off-shell Z boson as Z_2. These are essential to maximize the signal-to-background ratio in the signal region for the H $\to$ ZZ* $\to 4\ell$ decay channel. Having obtained the reconstructed leptons, the following next selection rules are followed to construct the Z candidates which would reconstruct the original Higgs boson.

- **Number of leptons** : Events reconstructed are required to have minimum four leptons in the final state. Specifically, events with four lepton configurations of 4 electrons, 4 muons and 2 electron 2 muon are valid configurations, called proper configurations. More than four leptons are possible in the scenarios of leptons obtained from brehmsstrahlung of photons (usually for electrons), some fake leptons added for probing the real leptons, etc.

- **Lepton flavours** : Since a Z boson decays into two leptons of same flavour and considering only electrons and muons in the final state, the four lepton combinations which are not proper configurations are rejected. Any events involving tau leptons would be detected as electron or muon depending on the final state lepton of tau leptonic decay, hence flavour combination would depend on the final state of the tau lepton decay and would fall in the categories as shown above.

- **Z candidate formation** : Based on the lepton flavour configurations, Z candidates are formed from same flavour opposite charge leptons (e^+e^- and $\mu^+\mu^-$). Z candidates which have a non-zero mass are selected as acceptable Z or Z* candidates. Here we define the reconstructed Z boson with the dilepton mass $m_{\ell^+\ell^-}$ closest to nominal on-shell Z boson mass (equal to 90 GeV) to be Z_1 and the other Z boson as Z_2.

- **Low p_T rejection** : Out of the four leptons forming the ZZ candidates, the leading and subleading lepton should have a minimum transverse momentum of 20 GeV and 10 GeV respectively.

- **Angular seperation between leptons** : Each lepton should have a minimum angular seperation of 0.02 with any other lepton in the reconstructed event. This is done to bring a cutoff to a highly boosted Higgs i.e Higgs with very high p_z.

- **Low mass dilepton pair rejection** : Out of the four leptons, any dilepton pair of opposite charge (irrespective of flavour) is required to have a minimum dilepton mass $m_{\ell\ell}$ of 4 GeV. This cut is to remove leptons coming from QCD-induced decays of J/ψ mesons.

- **Z mass window** : The Z_1 and Z_2 candidates are requried to have their mass in range $40\,GeV < m_{Z1} < 120\,GeV$ and $12\,GeV < m_{Z2} < 120\,GeV$ respectively.

- **Alternate ZZ pair rejection** : This check is specifically carried out for scenarios where the ZZ candidates have all four leptons of same flavour. Define alternative Z bosons Z_a and Z_b with Z_a being the Z with mass closest to on-shell Z boson mass. If
$$|m_{Za} - m_Z| < |m_{Z1} - m_Z| \quad and \quad m_{Zb} < 12\,GeV \tag{2.2}$$
Then the corresponding event is discarded from further analysis. This cut specifically remove events including an on-shell Z boson with a low-mass dileptonic resonance (As the Z_2 candidate in this case would not satisfy the Z_2 mass window)

- **Higgs mass window** : Using the four vectors of the Z candidates, one can reconstruct the four vector of the Higgs boson. Defining the mass of the Higgs reconstruced as $m_{4\ell}$, this cut requires the Higgs candidate to have a minimum mass of 70 GeV.

Ideally it is expected for only one ZZ candidate to survive in the end to have the full event reconstructed for further analysis. There could be cases where more than one candidate survives in the end. In that case:

- Out of the surviving ZZ candidates, the one with Z_1 mass closest to nominal Z mass is chosen as the best ZZ candidate.

- If the surviving candidates do have same Z_1 mass, then the one with highest transverse momentum for Z_2 is chosen as the best candidate.

The criteria for Z_2 transverse momentum is not effective in cases of HiggsStrahlung production modes, but since we are considering only gluon-gluon fusion mode in the analysis, we can use this criteria for the best ZZ candidate selection. Note that detailed kinematic analysis is done by calculating kinematic discriminants D^{kin}_{back} [25, 26], but calculating these is beyond the scope of this thesis. Hence, a simple analysis using the above mentioned selection rules is carried out to obtain a clean sample of events with signal events dominating the background events. These events obtained for the selection would be used later for developing the ML model and statistical analysis.

The main focus on the analysis would be the study of invariant mass distributions of the Higgs, Z_1 and Z_2 bosons, because invariant masses are Lorentz invariant. For that, the invariant masses of Higgs and Z bosons are evaluated from four vectors of the final state leptons, using simple kinematic relations [42]. Speaking of Lorentz invariance, the transverse momentum is also a good Lorentz invariant quantity for analysis, but the effects of cWW on transverse momenta of Higgs and Z bosons are small.

2.6 Kinematics of SM Higgs and SM ZZ events

For the analysis of the SM case, we consider first the full simulation samples and then the Delphes ones (using cWW=0.0023 sample) as described above in section 2.3.

2.6.1 Event kinematics using full simulation samples

For the analysis of the full simulation samples, we first apply the HZZ algorithm to the events obtained after detector simulation to get the required event candidates surviving the selection cuts. Then after, the events are normalised according to the luminosity under study (here 35.9 fb^{-1}). That implies weights are passed to the samples to normalise the distributions at the given luminosity with weights w_i given by

$$w_i = \frac{\sigma_{eff} \times L}{N_{total}} \tag{2.3}$$

where 'i' stands for either signal or background, L is the luminosity in fb^{-1} and σ_{eff} is the effective cross-section obtained by multiplying the production cross-section with the branching ratio of the decay channel (as described in third column of table 2.1). Fig 2.3 shows the kinematic distributions for the four leptons reconstructing the full event.

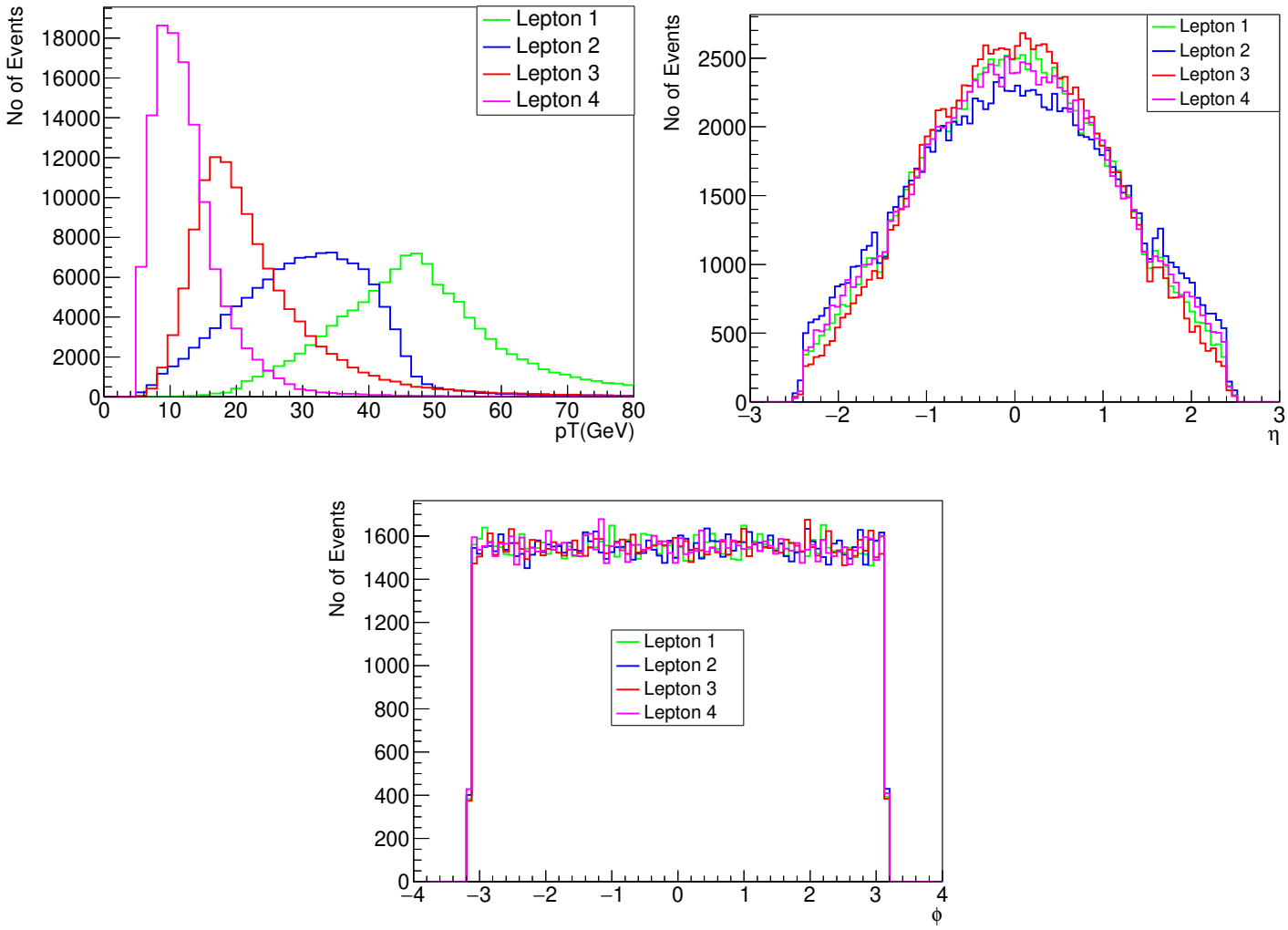


Figure 2.3: Distributions of the transverse momenta (upper left), pseudorapidity (upper right) and azimuthal angles (bottom) for the four lepton final states for the signal process $gg \to H \to ZZ^* \to 4\ell$ for leptons L1 (green), L2 (blue), L3 (red) and L4 (magenta). L1 and L2 constitute the Z_1 boson while L3 and L4 constitute the Z_2 boson.

CMS Sample	Weight calculated
Signal	0.000437
Background	0.00676

Table 2.4: Normalisation weights for the full simulation samples for computing event yields at luminosity 35.9 fb^{-1}.

This subsection will next present the outcome of the event selection carried out as described in the HZZ algorithm with event yields and distributions of the main kinematic variables.

Final State	4e	4μ	2e2μ	4ℓ
q$\bar{\text{q}}$ $\to$ ZZ	207.95	342.18	472.84	1022.98
Signal (m_H=125 GeV)	10.41	18.33	24.49	53.23

Table 2.5: Expected signal and background yields in the full signal region ($m_{4\ell} > 70$ GeV) for luminosity 35.9 fb^{-1}.

Final State	4e	4μ	2e2μ	4ℓ
$q\bar{q} \rightarrow ZZ$	3.64	8.19	10.59	22.43
Signal (m_H=125 GeV)	8.89	16.86	21.91	47.66

Table 2.6: Expected signal and background yields in signal region ($118\,GeV < m_{4\ell} < 130\,GeV$) for luminosity 35.9 fb^{-1}.

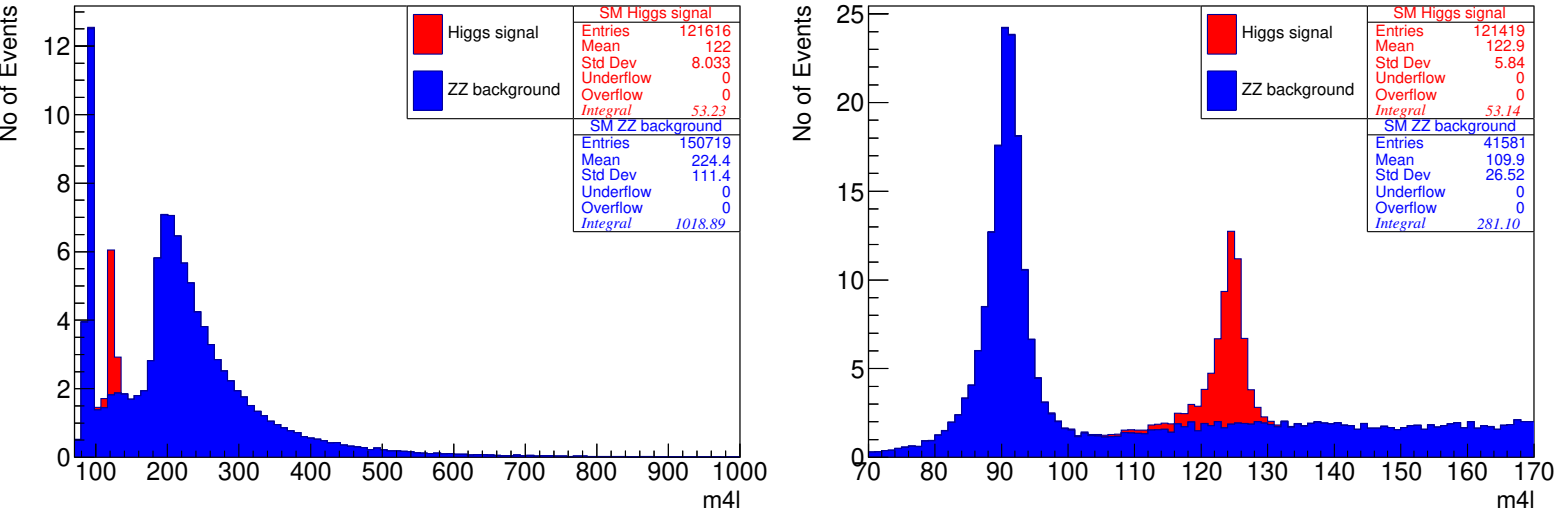


Figure 2.4: Distribution of the four-lepton reconstructed mass $m_{4\ell}$ in the full mass region (left) and the low mass region (right). The 125 GeV Higgs boson signal and the ZZ background are normalized to the SM expectation at 35.9 fb^{-1} luminosity. The distributions are stacked histograms.

From the distributions of fig 2.4, a clear peak is visible around 125 GeV coming from the signal events and not coming from the background which is in agreement with the narrow resonance expected to be observed around the Higgs mass window. The peak at around 180 GeV is the region where both the Z bosons generated are on-shell.

Selection criteria	**Surviving events**	**Ratio with total number of events**
No cut	999800	1
Number of leptons	141164	0.1412
Lepton flavours	138090	0.1381
Relative isolation selection	136794	0.1368
Low p_T rejection	135943	0.1359
Angular seperation between leptons	135922	0.1359
Low mass dilepton pair rejection	133436	0.1335
Alternate ZZ pair rejection	133383	0.1334
Z1 mass window	131875	0.1319
Z2 mass window	121670	0.1217
Higgs mass window	121616	0.1216

Table 2.7: The Cutflow table for the full simulation signal samples.

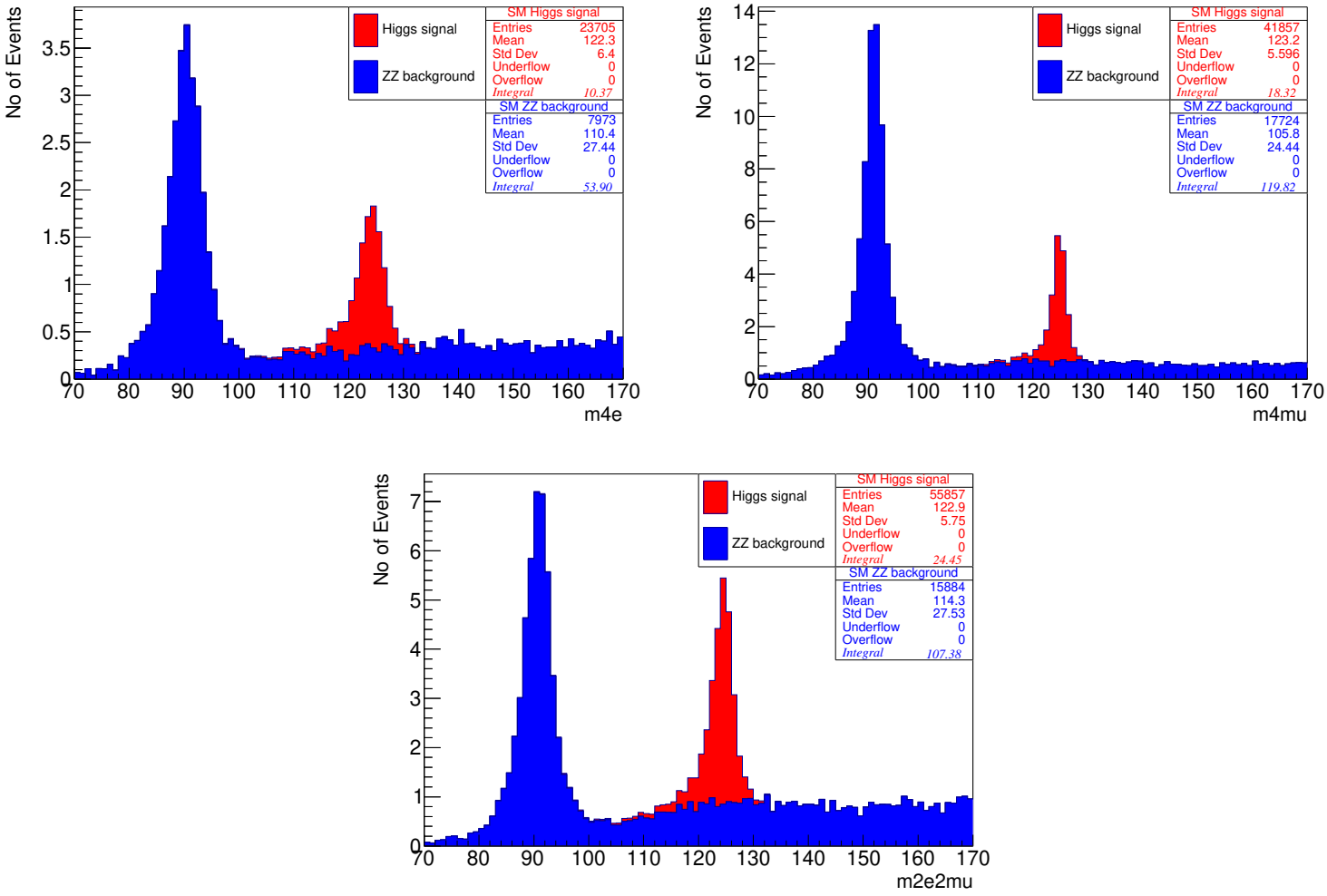


Figure 2.5: Distribution of the reconstructed four-lepton invariant mass distributions for each of the lepton flavours in the low mass range (4e case in upper left, 4μ case in upper right, 2e2μ case in bottom). The 125 GeV Higgs boson signal and the ZZ background are normalized to the SM expectation at 35.9 fb^{-1} luminosity. The distributions are stacked histograms.

Selection criteria	Surviving events	Ratio with total number of events
No cut	6669988	1
Number of leptons	244950	0.0367
Lepton flavours	237208	0.0355
Relative isolation selection	235541	0.0353
Low p_T rejection	229509	0.0344
Angular seperation between leptons	229467	0.0344
Low mass dilepton pair rejection	225109	0.0337
Alternate ZZ pair rejection	224974	0.0377
Z1 mass window	213155	0.0319
Z2 mass window	151195	0.0227
Higgs mass window	150560	0.0225

Table 2.8: The Cutflow table for the full simulation background samples.

Fig 2.5 show the four-lepton invariant mass distributions for individual lepton flavours.

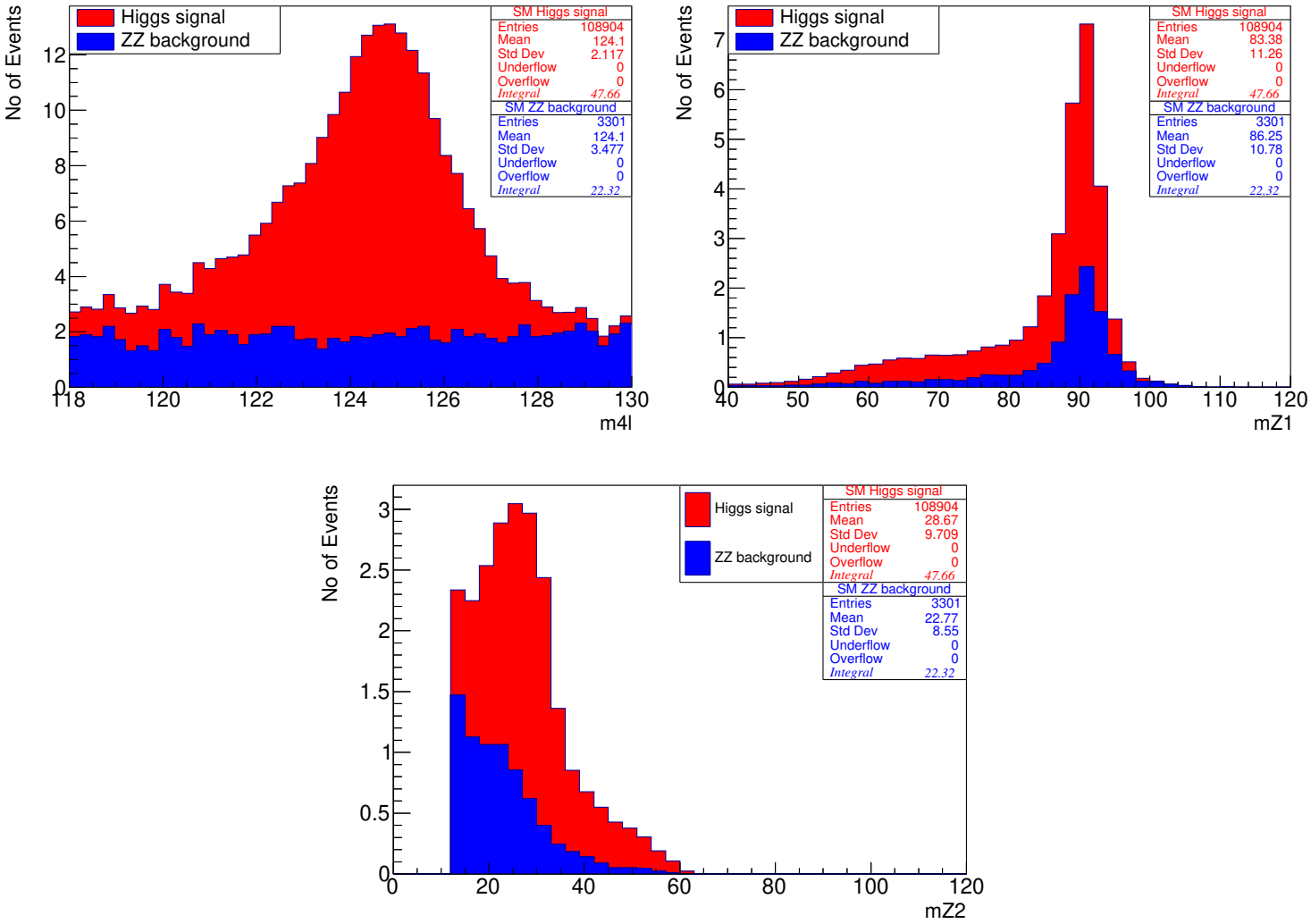


Figure 2.6: Distributions of the reconstructed four-lepton invariant mass in the signal region $118\,GeV < m_{4\ell} < 130\,GeV$ (upper left) and the distributions of the reconstructed Z1 (upper right) and Z2 (bottom) in the signal mass region. The 125 GeV Higgs boson signal and the ZZ background are normalized to the SM expectation at 35.9 fb^{-1} luminosity. The distributions are stacked histograms.

The observed number of events are consistent with the observed CMS results [27]. Because of electrons having tendency of going to brehmsstrahlung more than muons, the corresponding reconstruction efficiency for electrons is lesser than that of muons. Because of this, there is relatively poorer event yield in the 4e subchannel instead of the 2e2μ and 4μ subchannels. The efficiency for the 2e2μ subchannel is compatible with 4μ subchannel due to muon contributions. Fig 2.6 shows the invariant mass distributions for Higgs and Z bosons in the signal region ($118\,GeV < m_{4\ell} < 130\,GeV$).

The HZZ algorithm maximizes the signal yield in the region of interest by rejecting as many background events as possible. The signal-to-background yield ratio for the signal region here is equal to approximately 2. One of the reasons why this channel is extensively studied is because of this high signal purity in the signal region.

From the cutflow tables 2.7 and 2.8, one can see that a large number of events are already removed from the very beginning. This is due to majority of events with zero leptons detected and with some contribution from events with less than four leptons. Many

of them are not in the detector acceptance region and hence are not even selected for the event reconstruction. Excluding the large drop in the initial selection criteria, a large change in the number of events to be finally surviving comes at around the Z_2 mass cut for both signal and background events. The net efficiency for the signal and background events is 12% and 2% respectively.

2.6.2 Event kinematics using Delphes simulation samples

For the analysis of SM Delphes samples, the sample for lowest cWW (= 0.0023) was taken for as a proxy for SM case. The Delphes samples are the samples used for training the ML model (The full simulation samples were used to validate the algorithm). Since we do not have explicitly the SM Delphes sample (cWW=0) but are considering sample of cWW = 0.0023, the corresponding Delphes sample needs to be validated with the CMS observations for further analysis. For a complete validation to be done, one has to take care of the lepton reconstruction efficiencies in Delphes, but we have not done a complete validation of the event reconstruction of Delphes with respect to event reconstruction for full simulation samples but rather trying to find scale factors to match the final event yields. On applying the selection cuts on the Delphes samples for SM case, there is observed a discrepancy in the event yields and the kinematic distributions obtained (see fig 2.7).

Final State	4e	4μ	$2e2\mu$	4ℓ
$q\bar{q} \to ZZ$	1262.89	2699.47	3512.63	7475.0
Signal (m_H=125 GeV)	6.74	14.62	19.13	40.50

Table 2.9: Expected signal and background yields in the full signal region ($m_{4\ell} > 70$ GeV) for luminosity 35.9 fb^{-1} for Delphes SM sample before scale factor modification.

Final State	4e	4μ	$2e2\mu$	4ℓ
$q\bar{q} \to ZZ$	15.15	27.6	38.06	80.80
Signal (m_H=125 GeV)	6.45	13.9	18.3	38.67

Table 2.10: Expected signal and background yields in signal region ($118\,GeV < m_{4\ell} < 130\,GeV$) for luminosity 35.9 fb^{-1} for Delphes SM sample before scale factor modification.

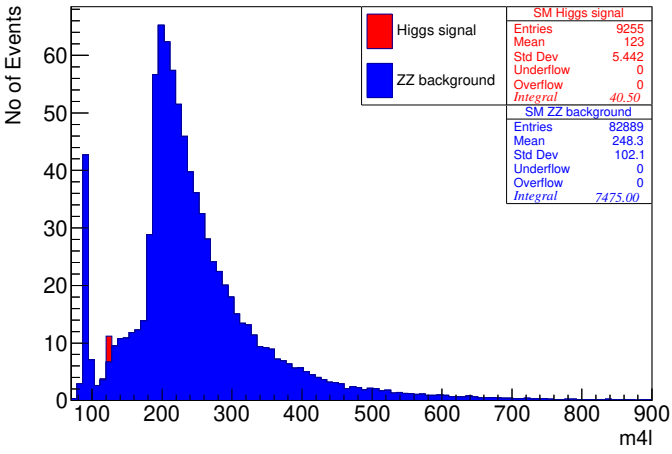
Delphes Sample	Weights calculated
Signal	0.004376
Background	0.09018

Table 2.11: Normalisation weights for the Delphes samples for computing event yields at luminosity 35.9 fb^{-1}.

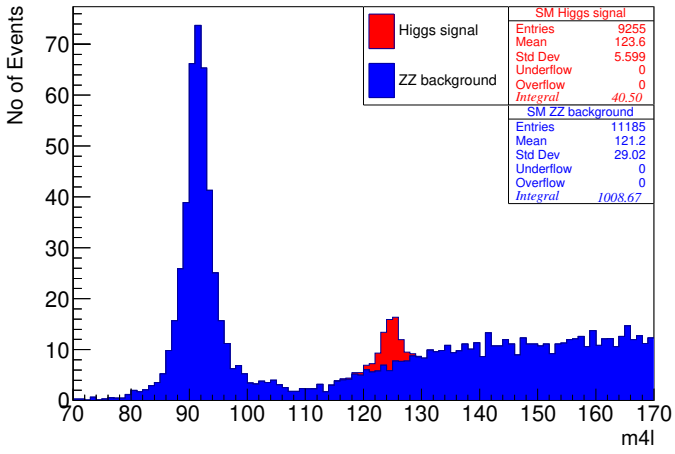
The Delphes samples are be normalised to meet the CMS expectations in event yields. To do that, the weights of delphes would be modified via scale factors (SF).

$$SF = \frac{(EventYield)_{CMS}}{(EventYield)_{Delphes}} \tag{2.4}$$

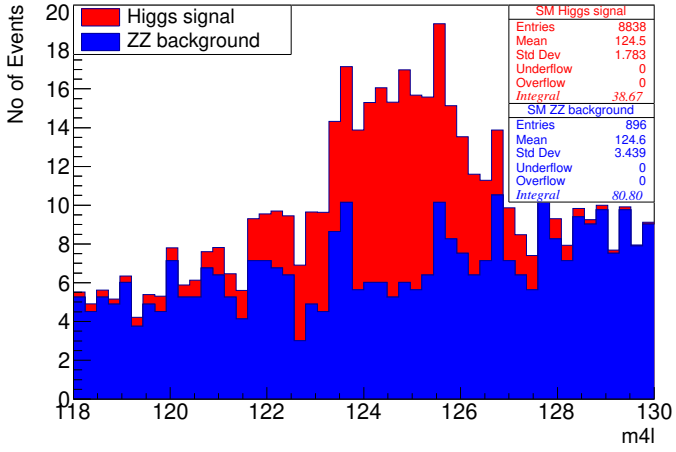
Since we need to ensure the two samples to agree in the signal region, the m_{4l} distributions are compared in the signal region as shown in figures 2.6 and 2.7c.



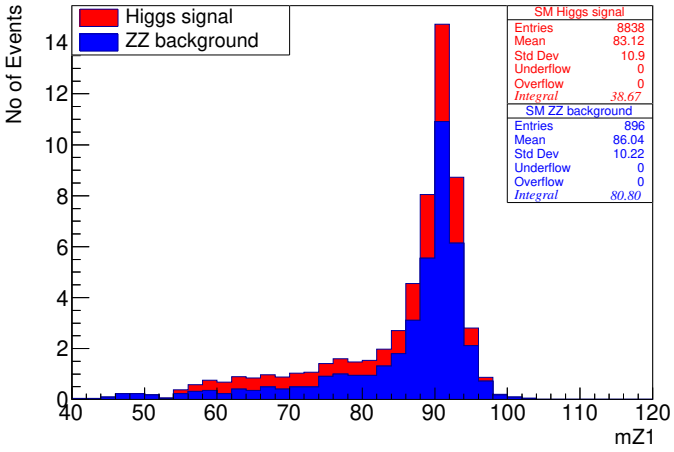
(a)



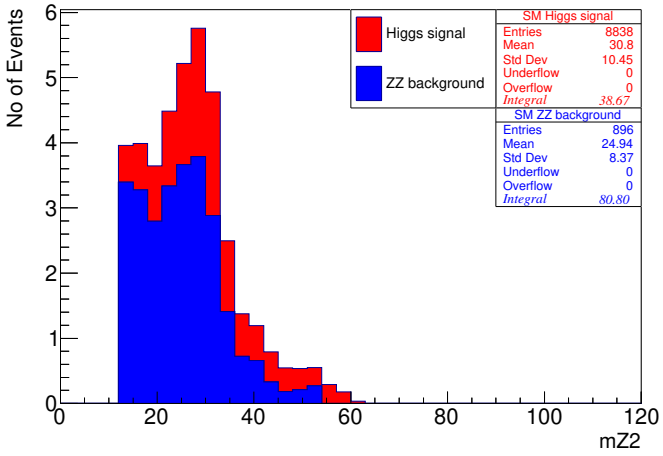
(b)



(c)



(d)



(e)

Figure 2.7: Distribution of the major kinematic variables for the Delphes samples for cWW=0. (a), (b) and (c) are the distributions of the reconstructed four-lepton invariant mass $m_{4\ell}$ for full mass range, low mass range and signal mass range respectively. (d) and (e) are the Z_1 and Z_2 distributions for events in the signal mass range. The 125 GeV Higgs boson signal and the ZZ background are normalized to the SM expectation at 35.9 fb^{-1} luminosity. The distributions are stacked histograms.

Sample	CMS Yield	Delphes Yield	SF
Signal	47.66	38.67	1.2325
Background	22.43	80.80	0.2775

Table 2.12: Scale factors (SF) calculated for the delphes samples to normalise the distributions with the full simulation distributions for luminosity 35.9 fb^{-1}.

.

Sample	$w^{old}_{delphes}$	SF	$w^{final}_{delphes}$
Signal	0.00437	1.2325	0.00538
Background	0.09018	0.2775	0.02502

Table 2.13: Updated weights for delphes samples as modified by the scale factors for luminosity 35.9 fb^{-1}. Old weights are obtained from table 2.11.

Finally, the modified weights are evaluated from the scale factors

$$w^{final}_{delphes} = w^{old}_{delphes} \times SF \tag{2.5}$$

This newly calculated weights would be used to calculate event yields and normalise the kinematic distributions for the Delphes samples. These SFs only affects the overall event yields and do not affect the shapes of the distributions since there is no reweighting carried out bin-by-bin for the distributions.

Final State	4e	4μ	$2e2\mu$	4ℓ
$q\bar{q} \to ZZ$	350.38	748.95	974.55	2073.88
Signal (m_H=125 GeV)	8.28	17.98	23.53	49.79

Table 2.14: Expected signal and background yields in the full signal region($m_{4\ell} > 70$ GeV) for luminosity 35.9 fb^{-1} for Delphes SM sample post SF modifications.

Final State	4e	4μ	$2e2\mu$	4ℓ
$q\bar{q} \to ZZ$	4.20	7.66	10.56	22.42
Signal (m_H=125 GeV)	7.94	17.11	22.49	47.55

Table 2.15: Expected signal and background yields in signal region ($118\,GeV < m_{4\ell} < 130\,GeV$) for luminosity 35.9 fb^{-1} for Delphes SM sample post SF modifications.

The same scale factors would be used for different cWW Delphes samples for the event yields.

2.7 Comparison of the EFT Higgs and SM ZZ event kinematics

The effect of the EFT interaction mediated by cWW parameter of the Higgs Effective Lagrangian can be understood by studying the kinematic distributions of the reconstructed invariant masses for different cWW values. In order to get the proper kinematic picture for the EFT interaction, the kinematic distributions are needed to be normalized with the corresponding EFT effective cross-sections.

Because cWW modifies the couplings of the vector bosons with Higgs [22], the parameter has the tendency to modify the decay channels HZZ and HWW at the same time. To understand the effect of EFT modifications, measurements are carried out for a particular physical quantity (say the decay width of the HZZ decay) and its deviations are measured from SM predictions. Associated with the cWW parameter, define

$$\kappa_z = \frac{\Gamma(H \to ZZ^*)_{EFT}}{\Gamma(H \to ZZ^*)_{SM}} \cong \frac{\Gamma^{EFT}_{H\to ZZ}}{\Gamma^{SM}_{H\to ZZ}} \quad and \quad \kappa_w = \frac{\Gamma(H \to WW^*)_{EFT}}{\Gamma(H \to WW^*)_{SM}} \cong \frac{\Gamma^{EFT}_{H\to WW}}{\Gamma^{SM}_{H\to WW}} \tag{2.6}$$

Based on the collision data obtained from LHC for the above two decay channels under gluon-gluon fusion production mode, the ATLAS and CMS collaborations reported measurements of the ratio of the EFT branching ratios relative to SM calculations.

$$\lambda_{WZ} = \frac{\kappa_W}{\kappa_Z} \tag{2.7}$$

As mentioned in [22], Monte Carlo events were generated using Madgraph which were used to parametrise the EFT branching ratios κ_W and κ_Z. This parametrization is then used to parametrise the ratio λ_{WZ} and compared with one measured experimentally.

$$\kappa_W = 1 + 2.23cWW + 1.27cWW^2 \tag{2.8}$$

$$\kappa_Z = 1 + 1.97cWW + 1.0cWW^2 \tag{2.9}$$

Note that in addition to the cWW parameter, there is another interaction mediated by cB parameter which can affect the couplings with vector bosons. Because of which, the expressions 2.8 and 2.9 would include cB (EFT parameter known to modify photon couplings, which can modify specifically decay channel $H \to Z\gamma$ [22]) contributions also. In our analysis, any other EFT contributions are set to zero so we can ignore their contributions for the purpose of proof of concept developed in this thesis.

Using the above information, the branching ratio for the HZZ channel can be calculated to obtain effective EFT cross-sections. Defining $\Gamma^{SM}_{H\to all}$ as the total decay width of the Higgs boson in SM, the following expression of branching ratio can be obtained by some simple algebraic manipulations on equations 2.8 and 2.9 and using the fact that the branching ratio is the ratio of partial decay width to total decay width.

$$BR(\Gamma^{EFT}_{H\to ZZ}) = \frac{\Gamma^{SM}_{H\to ZZ}(1 + 1.97cWW + cWW^2)}{\Gamma^{SM}_{H\to all} + cWW(2.23\Gamma^{SM}_{H\to WW} + 1.97\Gamma^{SM}_{H\to ZZ}) + cWW^2(1.27\Gamma^{SM}_{H\to WW} + \Gamma^{SM}_{H\to ZZ})} \tag{2.10}$$

Based on [22], any modification to the total decay width would come from the vector boson channel. To get the branching ratio of the full decay channel H $\to$ ZZ* $\to 4\ell$, one needs to include the contribution from decay of Z bosons. Because the couplings of Z with leptons are not affected by cWW [43], one can set the effective branching ratio for Z $\to \ell^+\ell^-$ to be equivalent to the SM branching ratio. So

$$BR(\Gamma^{EFT}_{H\to ZZ\to 4\ell}) = BR(\Gamma^{EFT}_{H\to ZZ}) \times BR(\Gamma^{SM}_{Z\to \ell^+\ell^-})$$

Also, the production mode under study is the gluon-gluon fusion which is unaffected by cWW. Hence the effective cross-section for the EFT case is given by

$$\sigma^{EFT}_{eff} = \sigma^{EFT}_{prod} \times BR(\Gamma^{EFT}_{H\to ZZ\to 4\ell}) \tag{2.11}$$

$$\cong 48.58(pb) \times BR(\Gamma^{EFT}_{H\to ZZ\to 4\ell}) \tag{2.12}$$

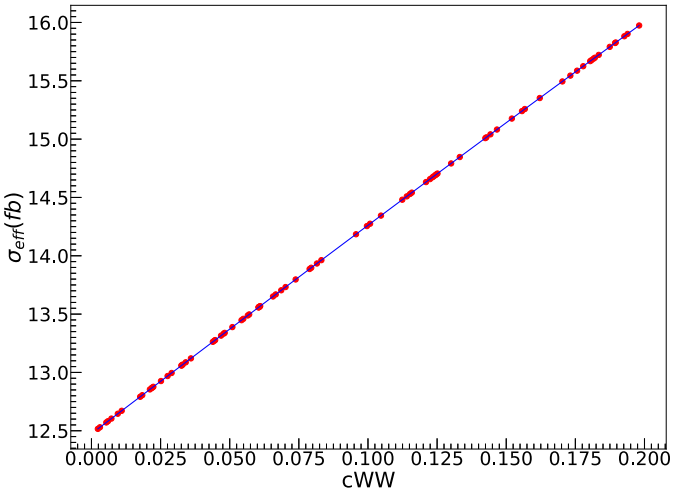


Figure 2.8: Variation of the effective cross-section with cWW parameter. The effective cross-sections are computed for cWW values as tabulated in table 2.3. An approximate linear increase is observed in the cross-section.

Fig 2.8 shows the variation of effective cross-section with cWW. To carry out the analysis of cWW samples using the EFT cross-sections, the weights evaluated for Delphes in table 2.13 are to be modified. The scale factor to ensure consistency is already included in the weights computed using SM cross-section. We have the updated EFT weights as

$$w^{EFT}_{delphes} = w^{SM}_{delphes} \times \frac{\sigma^{EFT}_{eff}}{\sigma^{SM}_{eff}} \tag{2.13}$$

where $w^{SM}_{delphes}$ are the weights as evaluated in table 2.13. Note that this would remain constant for all cWW samples, the modification would be done by the shift from SM cross-section to EFT cross-section. Also, equation 2.13 is applicable for signal samples here, because we are considering SM background with cWW=0.

The kinematic distributions for six example cWW samples over SM background are shown in figures 2.9, 2.10 and 2.11. Luminosity normalisation is carried out using EFT cross-sections for signal events using weights calculated from equation 2.13 and the ZZ background is normalised to SM cross-section using weights as shown in table 2.13. The distributions are shown in the signal region where relevant changes are observed for different cWW parameters.

Because of the modification by the EFT interactions, there is observed a relative increase in the event yields with increasing cWW. This is on account of the increasing effective cross-section of the decay channel. The EFT parameter cWW does not modify the Higgs self couplings to change the mass of Higgs boson, hence there is relatively no difference in the $m_{4\ell}$ distributions of the non-zero cWW samples and SM zero cWW case.

Deviations in the distributions are expected for the Z_1 and Z_2 distributions, because the HZZ vertex is modified by the cWW parameter. For small cWW values close to zero, the Z_1 distributions are similar with the 90 GeV resonance peak. The probability that the resonance peak of Z_1 would be around 90 GeV changes with a change in cWW. This changes the probability for Z_2 boson in an anti-correlated manner. With increasing cWW,

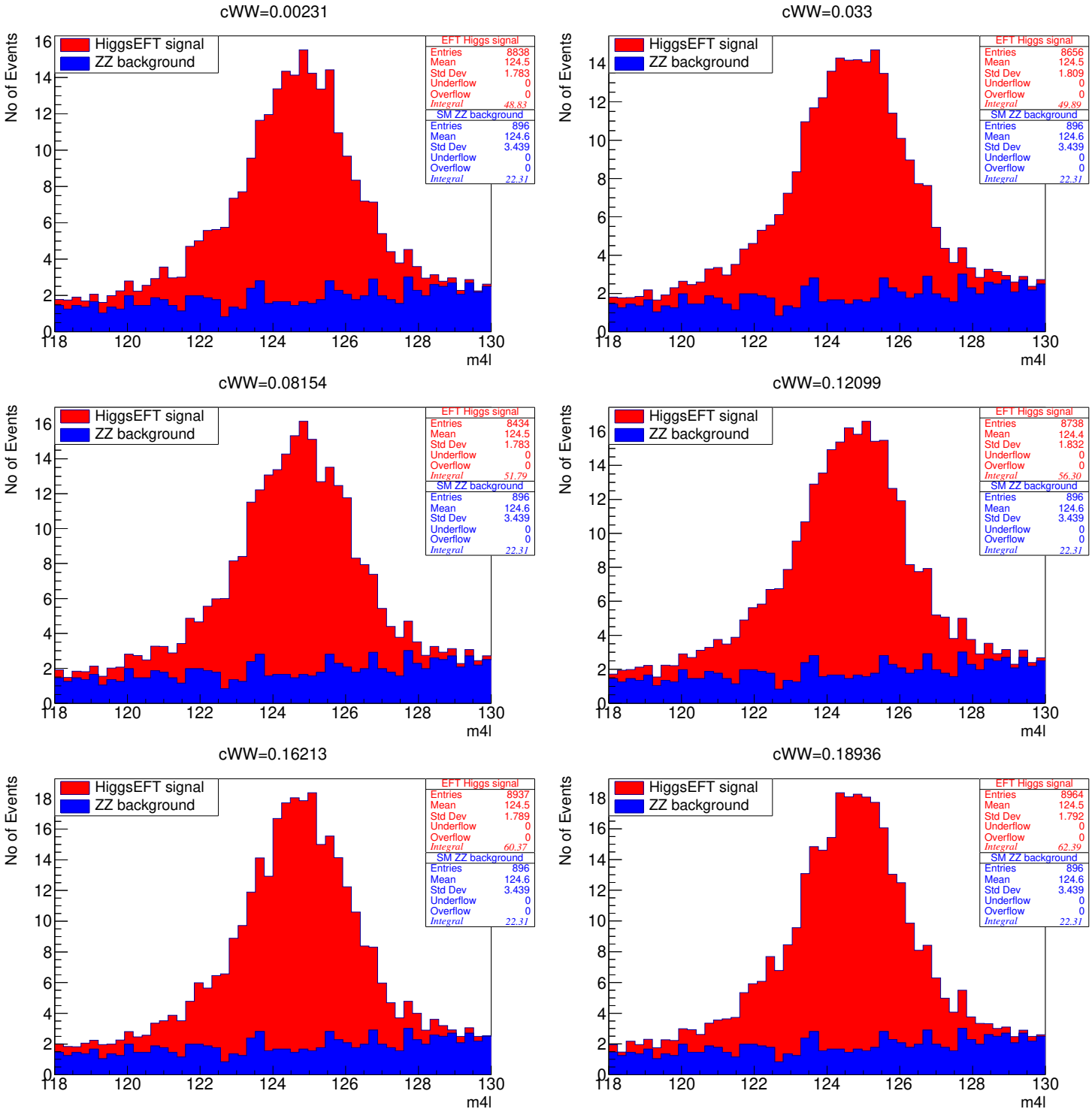


Figure 2.9: Distributions of the reconstructed four-lepton invariant mass $m_{4\ell}$ in the signal region ($118\,GeV < m_{4\ell} < 130\,GeV$). The 125 GeV EFT Higgs boson signal and the ZZ background are normalized to the EFT and SM expectation respectively at 35.9 fb^{-1} luminosity. The corresponding cWW value is mentioned at the top of the plots. The distributions are stacked histograms.

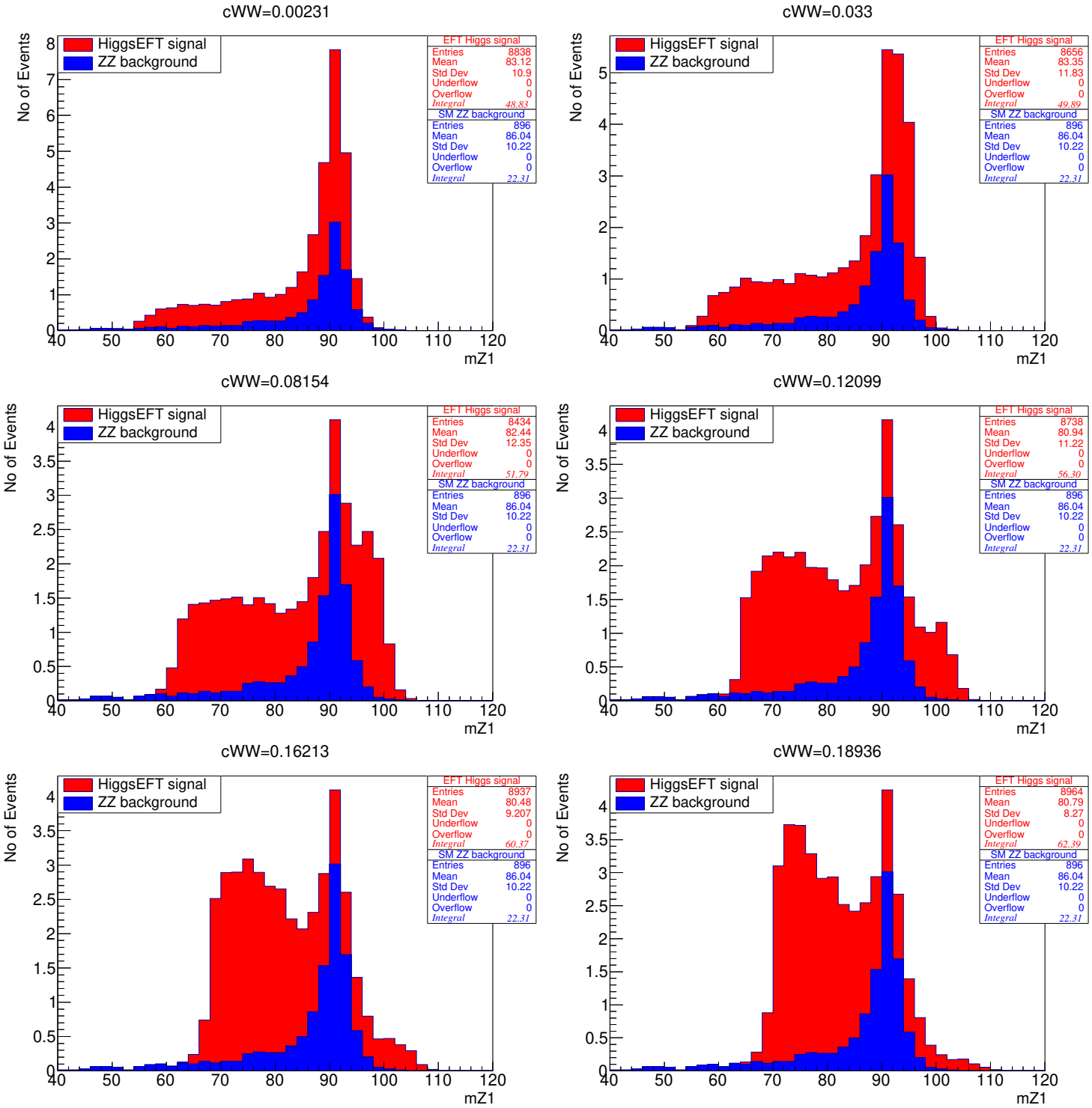


Figure 2.10: Distributions of the reconstructed Z_1 boson in the signal region ($118\,GeV < m_{4\ell} < 130\,GeV$). The 125 GeV EFT Higgs boson signal and the ZZ background are normalized to the EFT and SM expectation respectively at 35.9 fb^{-1} luminosity. The corresponding cWW value is mentioned at the top of the plots. The distributions are stacked histograms.

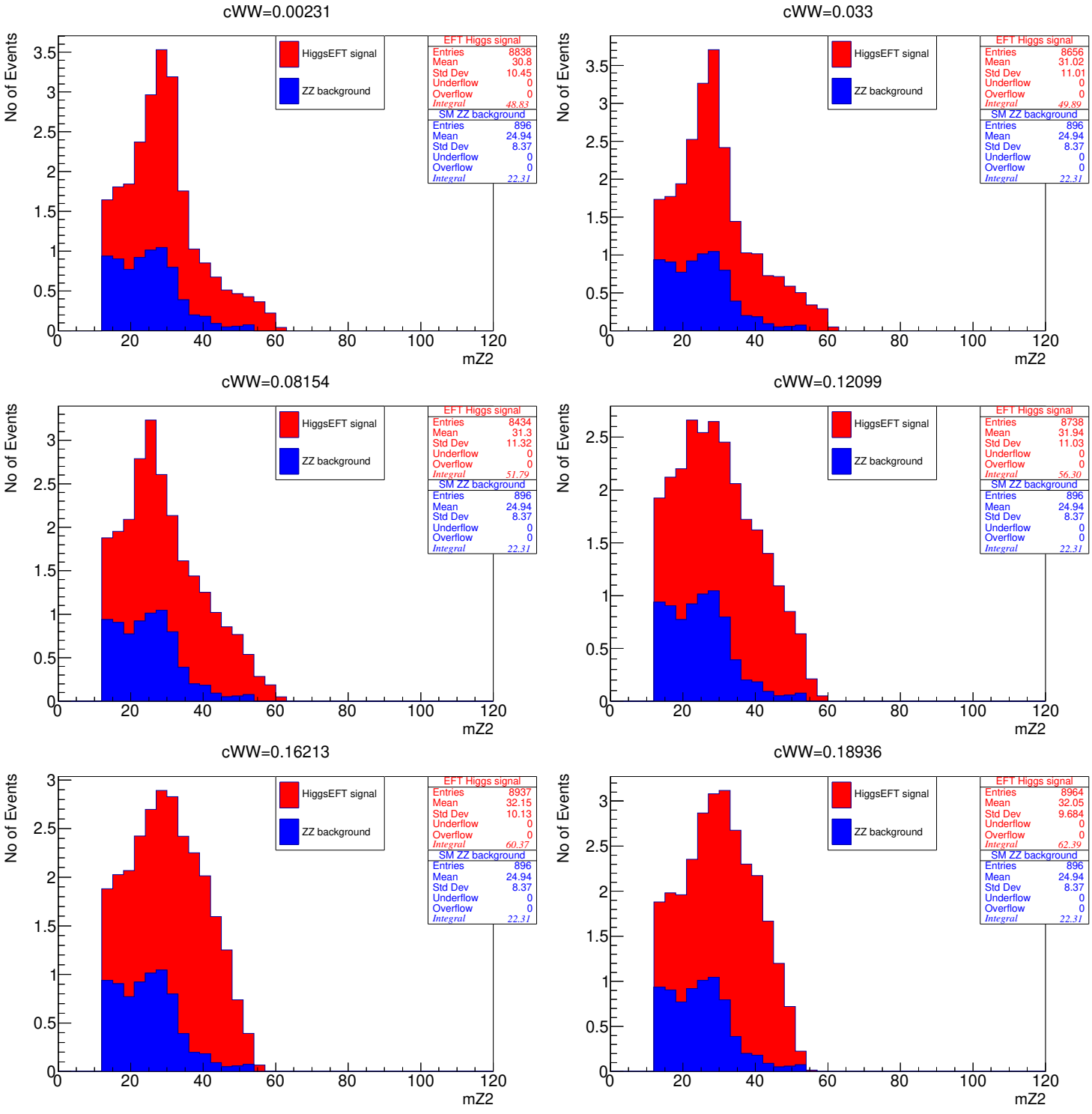


Figure 2.11: Distributions of the reconstructed Z_2 boson in the signal region (118 GeV $< m_{4\ell} <$ 130 GeV). The 125 GeV EFT Higgs boson signal and the ZZ background are normalized to the EFT and SM expectation respectively at 35.9 fb^{-1} luminosity. The corresponding cWW value is mentioned at the top of the plots. The distributions are stacked histograms.

cWW	4e	4μ	2e2μ	4 ℓ
0.0	7.94	17.11	22.49	47.55
0.0023	8.15	17.57	23.10	48.83
0.033	8.55	17.79	23.55	49.89
0.08154	8.39	18.86	24.53	51.79
0.12099	8.98	20.14	27.18	56.30
0.16213	9.65	21.79	28.93	60.37
0.18936	10.70	22.01	29.67	62.39

Table 2.16: Expected signal event yields for the delphes samples for different leptonic channels for luminosity 35.9 fb^{-1}. Normalisation is done using EFT cross-sections. The SM (cWW=0) is kept for comparison.

.

the Z_1 bosons with mass in the range of 50 GeV to 70 GeV gets more probable than SM 90 GeV, i.e. Z_1 starts getting more and more off-shell with respect to SM criteria as seen from distributions of fig 2.10. Because the Z bosons are coming from decay of Higgs of 125 GeV mass, the probability of Z_2 boson to be in 40 GeV to 60 GeV mass range increases in a similar fashion as seen from distributions of fig 2.11. Because of the change in these probabilities, the decay width of H$\rightarrow$ZZ changes which increases with increasing cWW due to larger cross-section. This effect on the HZZ decay channel due to cWW from one on-shell and one off-shell Z boson pair to off-shell Z boson pairs is also evident from [43] and [22].

Fig 2.12 shows the comparison between the Z_1 and Z_2 bosons for the SM and EFT scenarios. For small cWWs, the distributions are similar and deviations start with less probability around 90 GeV peak for Z_1 and around 35 GeV peak for Z_2 when cWW increases. It should be noted here that the distributions for large cWW deviate by large amounts from data obtained from LHC, so it is expected for the cWW parameter to be small and greater than zero.

2.8 Summary

The kinematic distributions and the event yields for a luminosity of 35.9 fb^{-1} are studied for the SM and EFT HZZ signal samples over a SM ZZ background. Deviations in the distributions of the dilepton pairs is observed for the EFT samples by which the physical effect of the cWW parameter can be visualized and studied.

To validate the HZZ algorithm, the event yields were calculated for full simulation samples and were compared with the official CMS results. Having obtained a consistency in the full simulation samples, the HZZ algorithm was tested on Delphes samples and a discrepancy in event yields was observed. A thorough validation of the Delphes is not done but instead as proxy for analysis scale factors are evaluated to match the event yields of the full simulation samples

The weights calculated in this chapter would form the basis for constraining the EFT parameter cWW based on the kinematic distributions of the invariant masses. For that, we look for a machine learning appoach which would be discussed in next chapters.

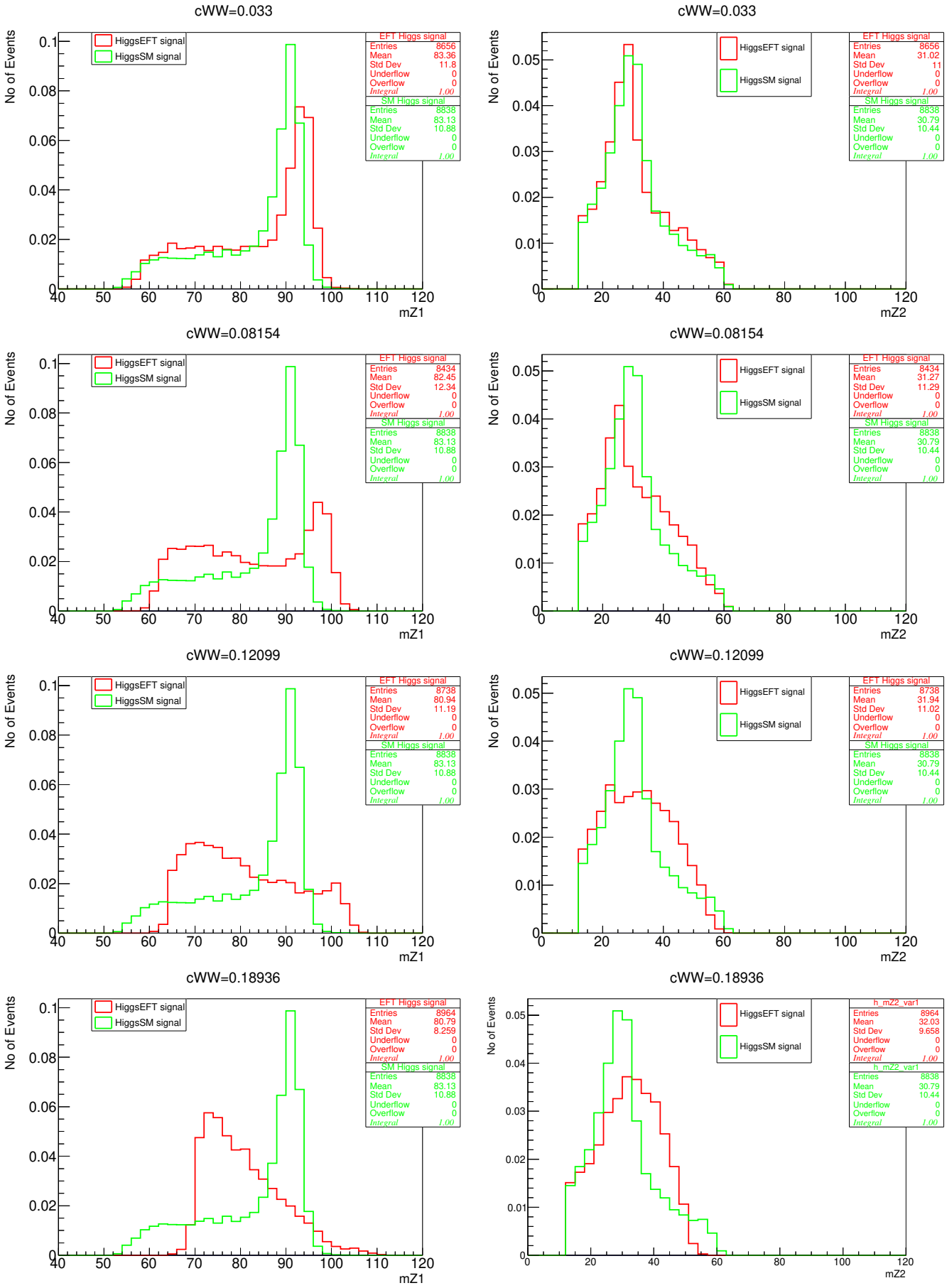


Figure 2.12: Comparison between the EFT and SM distributions for Z_1 (left) and Z_2 (right). The distributions are normalised such that their area is one. The red histogram corresponds the Z_1 and Z_2 distributions for positive cWW EFT samples, and the green one corresponds to SM case, which here is approximated by lowest cWW EFT sample (cWW = 0.0023).

Chapter 3

Parametric density estimation using machine learning

3.1 Brief overview of machine learning concepts

Machine learning (ML) has become one of the widely used artificial intelligence techniques in data analysis. ML models consists of learning the internal features and correlations among the given set of datapoints and making predictions based on these "learned" features. On a naive basis, ML sounds very similar to curve fitting, however the one basic difference between the two approaches is the flexibility in the function to be fitted in ML. This thesis emphasizes work on a sub-branch of ML called Deep learning (DL). DL includes learning algorithms mimicking the functioning of the neurons of the brain. Neural networks are interconnected circuits of data processors called neurons, where information flows from one neuron to another neuron to generate desired output. Fig 3.1 shows the basic structure of a neural network.

Each of the output of the neuron of a neural network is weighted by a weight w_i. This weighted wrapping of information is required to pass through the activation function, which regulates the way information should be passed to the next neuron. It is called activation function because its functioning is equivalent to firing synapses of an actual neuron (see fig (b) of 3.1). The basic mathematical idea behind the information exchange of two neurons is a linear transformation modelled by the weights. When the model is being trained, it is these weights which are modified so that the output predictions matches closest to the actual output. In ML terminology, weights are called parameters, which are dynamic in nature and change during the training procedure. On the other hand, there are some quantities, like the number of neurons of the hidden layer, which are decided by the user in the beginning which do not change during the training process. Such constants are called hyperparameters.

The measure of deviation from the true expected output in the prediction of model is the loss function or cost function of the model. Training is defined as the optimization of the model weights which can give minimum loss. The weights of the network are then updated via a backpropagation algorithm and gradient descent algorithm.

1. Using initial weights to the neurons, predictions are computed for the given inputs.

2. The predictions are compared with the actual output, and the extent of deviation is expressed using a loss function ($\mathscr{L}$).

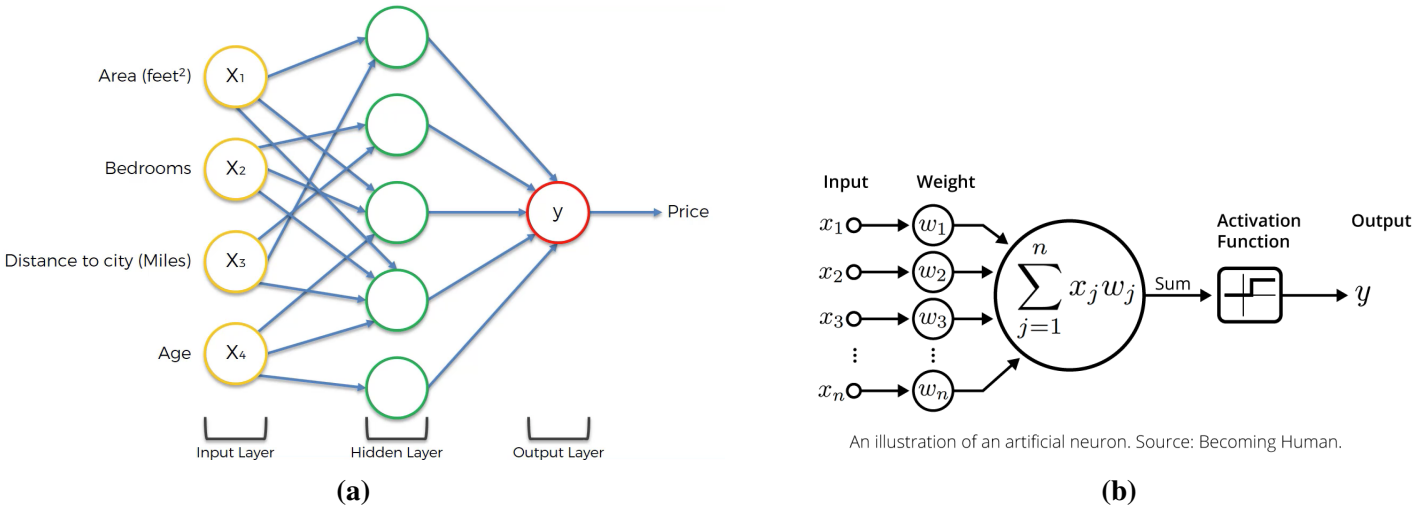


Figure 3.1: (a) Basic structure of a neural network. Information enters the network via the input layer and is processed by the hidden layers to send out output in the output layer. (b) The information through a neuron processed in the hidden layer by weighting inputs of previous layer passing it to next layer through the activation function [45].

3. **Backpropagation algorithm** : Going backwards from the output layer to the input layer, the derivatives of the loss function with respect to the weights of the neurons of a given hidden layer are calculated using chain rule. Intuitively speaking, the gradients of loss with respect to weights are computed using the chain rule on gradients of loss with respect to features and the gradients of features with respect to weights (see [44] for mathematical details).

4. **Gradient descent** : The derivatives computed using backpropagation are used to updates the weights as:

$$w_{i+1} = w_i - lr \times (\frac{\partial \mathscr{L}}{\partial w_i}) \tag{3.1}$$

where w_i is the weight during i^{th} epoch updated to w_{i+1} for next epoch, and lr is the learning rate. Learning rate is a hyperparameter which controls the extent to which weights are to be updated by the gradients. Gradient descent optimization algorithm works on the principle of descending downwards towards the minimum of the cost function.

Training is an iterative process where the model goes through the entire training data for a given amount of iterations to optimize the weights. An epoch is defined when the entire training data is scanned through once.

3.2 Probability density estimation using machine learning

One of the fundamental ways to characterise a set of data points is to model its inherent probabilistic distribution, because the probability density distributions (or density distributions) helps in understanding the properties of data like to pick up set of points with more importance, classifying outliers based on their less probabilities of occurence, etc. Statistically, there are two approaches to density estimation:

- Non-parametric estimation : Density estimation carried out by taking the entire dataset as parameters of a probability density function (pdf). Examples include Kernel density estimation, K-nearest neighbours,etc.

- Parametric estimation : Density estimation based on known distribution models whose relatively limited number of parameters are optimized from the data. Examples include Maximum likelihood estimation for Gaussian prior, Binomial prior, etc.

ML and DL techniques make use of generative models to get an idea about the distribution of the data. Generative models are typically a class of models which learns the features of the distribution of data to generate new datasets. Generative models can be used for both parametric and non-parametric density estimations. Three well known approaches for generative models exist in deep learning:

1. **Generative adverseial networks (GAN)** : GANs are generative models consisting of two components : A discriminator model and generator model. The generator model generates random data from the sample it has learned followed by the discriminator to differentiate it with real data. This tug-of-war goes on wherein both the models improve themselves, the generator improving to generate as many new data samples while discriminator improving to differentiate well between fake and real data.

2. **Variational autoencoders (VAE)** : As name suggests, VAE compresses the data to a latent distribution via an encoder and a decoder decodes data from the latent distribution to generate data samples. The extent of similarity between real and generated data is measured via Kullback-Liebler Divergence (more in section 4.5). KL-Divergence helps in accessing the similarity of the distributions, so VAE approximates target distribution.

3. **Normalizing flows (NF)** : NFs are generative models which stacks bijective functions on a latent distribution to model a target distribution. Bijective functions are functions which are invertible in nature making a one-to-one connection between input and output variables. The idea behind normalizing flows is to obtain a composite function through a series of invertible functions, which would map the target data distribution to a latent distribution, usually taken as a gaussian.

GANs and VAEs are generative models which actually do not capture the pdf explicitly like the NFs. GANs have issues with precise hyperparameter tuning and at a time one has to train two different models. VAEs on other hand only approximates the pdf while one can get the exact functional approach for the pdf via normalizing flows. Fig 3.2 pictorially represent the normalizing flows approach. This thesis is focussed on NFs, and specifically on a modification of Real-valued Non-Volume Preserving (RealNVP) models for parametric density estimation.

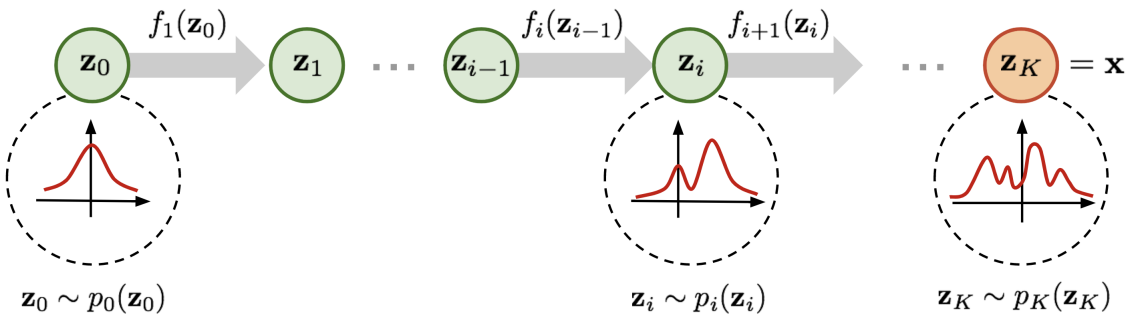


Figure 3.2: Pictorial representation of normalizing flows model. Bijective functions are applied on a gaussian distribution which can be trained via a neural network. The functional representation of the pdf of target data can be obtained as the composition of the transformation functions [46].

3.3 Conditional normalizing flow model

As described in the previous section, NFs are a class of generative models wherein probability distributions are generated by stacking bijective functions on a prior known density, generally preferred to be the standard normal distribution [46]. Specific to NFs, the model developed is based on the Real-valued Non-Volume Preserving (RealNVP) model class described below [47].

3.3.1 Change of variables

Consider a set of observations $x \in X$, where x could be a vector of dimensions m. In our case it would be the three vector of invariant masses $\overrightarrow{x} = (m_{4\ell}, m_{Z1}, m_{Z2})$, following some unknown distribution p_X. To get an idea about how p_X looks like, we follow a change of variables procedure where the input x is transformed to a latent variable z of known prior probability distribution p_Z (generally a gaussian distribution) via a bijective transformation function $f : X \to Z$ (dimension preserving function), whose inverse is $g : Z \to X$. Then the distribution of X is given by:

$$p_X(x) = p_Z(f(x)) \left| det\left(\frac{\partial f(x)}{\partial x^T}\right) \right| \tag{3.2}$$

$$log(p_X(x)) = log(p_Z(f(x))) + log \left| det\left(\frac{\partial f(x)}{\partial x^T}\right) \right| \tag{3.3}$$

where $\frac{\partial f(x)}{\partial x^T}$ is the Jacobian of f at x given by

$$\mathbf{J} = \frac{\partial f(x)}{\partial x^T} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix} \tag{3.4}$$

So what is being done here is that by using a known distribution, a mapping function is used to map the observed data to a point of this distribution. The information of the

transformation is contained in the Jacobian which is used to get the data distribution. The above equation forms the basis for the training of the ML model, especially equation 3.3 which defines log-likelihood of the point x. One of the challenges which comes in the normalizing flows is the choice of this bijective function f. The function should be chosen in such a manner that its Jacobian (equation 3.4) is easy to compute, and the function also needs to be easily invertible. RealNVP model makes use of scaling functions in form of exponentials and translation functions to achieve this task because it is very easy to invert an exponential function and a translation function.

3.3.2 Conditional RealNVP model

When considering Conditional RealNVP model, a slight modification comes into play with the inputs. Instead of only the vector as the input, the bijective functions also take into account a constant vector (scalar) $\overrightarrow{\theta}(\theta)$. The significance of the constants is to parametrise the data distributions by the parameter. For our purpose, we are trying to parametrise the invariant mass vector $(m_{4\ell}, m_{Z1}, m_{Z2})$ via the EFT parameter cWW, hence we consider a single parameter θ instead of a general vector.

Considering an input vector $\overrightarrow{x}$ having D dimensions which is transformed into a D dimensional output vector $\overrightarrow{y}$ via the bijective function $f : \overrightarrow{x} \longmapsto \overrightarrow{y}$. Parametrising the distributions based on a scalar θ , the Conditional RealNVP bijective functions divides the input vector into two parts:

- Constant component : First d components remain same (identity function)
- Scale-Shift Term : The remaining $D-d$ components are scaled and shifted by functions of the constant d components

$$\mathbf{y}_{1:d} = \mathbf{x}_{1:d} \tag{3.5}$$

$$\mathbf{y}_{d+1:D} = \mathbf{x}_{d+1:D} \odot exp(s(\mathbf{x}_{1:d}, \theta)) + t(\mathbf{x}_{1:d}, \theta) \tag{3.6}$$

where s and t are the scaling and translation functions defined on the space $R^d \longmapsto R^{D-d}$ and $\odot$ is the hadamard product (element wise product) [48]. The above transformations simplifies the Jacobian (equation 3.4):

$$\mathbf{J} = \frac{\partial y}{\partial x^T} = \begin{bmatrix} I_{d\times d} & 0 \\ \frac{\partial y_{d+1:D}}{\partial x_{1:d}^T} & diag(exp(s(x_{1:d}, \theta))) \end{bmatrix} \tag{3.7}$$

where $I_{d\times d}$ is the d dimensional identity matrix and $diag(exp(s(x_{1:d}, \theta)))$ is the diagonal matrix of elements of $exp(s(x_{1:d}, \theta))$. The determinant of this Jacobian is easier to evaluate, because it is just the product of the diagonal elements. Thus, the determinant of the Jacobian is given by

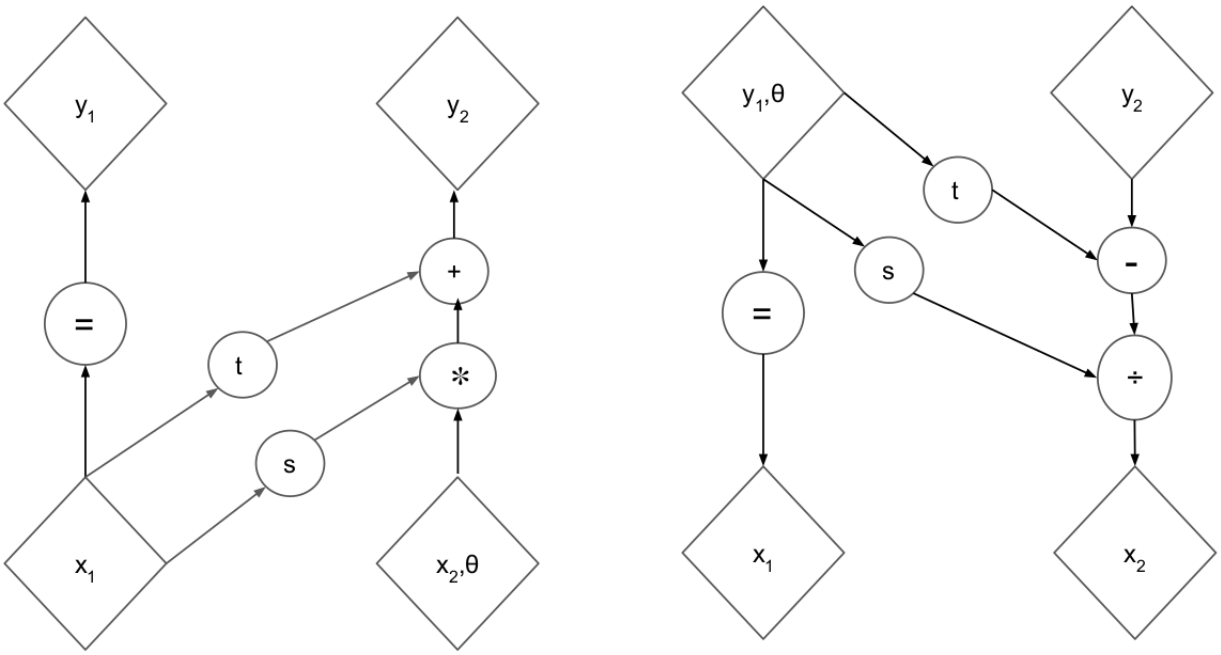
$$|\mathbf{J}| = exp(\sum_j s(x_{1:d}, \theta)_j) \implies log(|\mathbf{J}|) = \sum_j s(x_{1:d}, \theta)_j \tag{3.8}$$

Thus, the log determinant is the summation of the $D-d$ components of the scaling function. Note that the log-determinant is independent of the translation function and the determinant does not depend on the bijective nature of s and t functions. So one can make the scaling and translation functions as complicated as required. The complexity and flexibility of this functions can be modelled by deep neural networks.

As the bijective function in our case is a mix of identity functions, exponential functions and translation functions, these functions are invertible as shown:

$$\mathbf{y}_{1:d} = \mathbf{x}_{1:d} \Leftrightarrow \mathbf{x}_{1:d} = \mathbf{y}_{1:d} \tag{3.9}$$

$$\mathbf{y}_{d+1:D} = \mathbf{x}_{d+1:D} \odot exp(s(\mathbf{x}_{1:d}, \theta)) + t(\mathbf{x}_{1:d}, \theta) \Leftrightarrow$$
$$\mathbf{x}_{d+1:D} = (\mathbf{y}_{d+1:D} - t(\mathbf{y}_{1:d}, \theta)) \odot exp(-s(\mathbf{y}_{1:d}, \theta)) \tag{3.10}$$



(a) Forward Flow **(b)** Backward Flow

Figure 3.3: Computational graphs for forward and inverse flow. A coupling layer defines the transformation as scaling and translating a component of the vector x_1 and parameter θ and adding the contribution to component x_2 of the input vector. x_1 passes undisturbed. The roles of x and y are switched depending on the propagation flow considered. The computational cost between the two flows do not differ, although the log-likelihood calculated would differ.

The above invertible functions define the basis for the propagation of information via normalizing flows : which are the backward flows and forward flows as pictorially represented in fig 3.3. The transformation equations on the left of double implication in equations 3.9 and 3.10 corresponds to what is called the forward flow. So we start with a known prior distribution (represented by $\overrightarrow{x}$ here) and use the transformation equations to obtain data distributions. The equations on the right represents the backward propagation, wherein data points are mapped on a latent prior distribution. Backward flow is what governs the training of the model, because the model would try to fit the data to a prior distribution.

3.3.3 Coupling layers

The transformation equations described in equation 3.6 constitute a single function f, which defines a coupling layer or affine layer. Because the transformation functions leaves

some of the components unchanged, the choice of components to transform or not transform is controlled by coupling layers. The transformation functions in two consecutive coupling layers are kept in such a manner that if one component does not transform in the previous layer, it is transformed in the next consecutive layer. This can be achieved by masking some of the components of the input vector in a given coupling layer.

Lets say one starts with input vector $\overrightarrow{x}$, and via passing through N coupling layers the vector is mapped to the prior distribution (gaussian) $\overrightarrow{z}$. For i^{th} coupling layer, the corresponding transformation function is f_i and the output vector is $\overrightarrow{x}_i$ obtained from output of consecutive previous coupling layer $\overrightarrow{x}_{i-1}$. Applying equation 3.3 for the distribution of $\overrightarrow{x_{i-1}}$, we have

$$x \xrightarrow{f_1} x_1 \xrightarrow{f_2} x_2 \xrightarrow{f_3} x_3 \ldots\ldots x_{N-2} \xrightarrow{f_{N-1}} x_{N-1} \xrightarrow{f_N} z$$
$$f_i : x_{i-1} \longmapsto x_i$$
$$log(p_{i-1}(x_{i-1})) = log(p_i(f_i(x_{i-1}))) + log\left|det(\frac{\partial f_i(x_{i-1})}{\partial x_{i-1}^T}\right|$$

Having obtained the distribution of x_{i-1} in terms of f_i, one can repeat the same calculation to obtain an expression for distribution of x_i in terms of x_{i+1}. This calculation is repeated until one reaches the prior distribution variable z after applying N transformations. One can see that at each individual step, the log of the determinant of the Jacobian for i^{th} coupling layer gets added. The net effective transformation function would be a composition of this individual functions of the coupling layers, with the effective function mapping the input $\overrightarrow{x}$ with prior $\overrightarrow{z}$. Thus equation 3.3 can be modified to include the effect of coupling layers to get:

$$log(p_X(x)) = log(p_Z(f(x))) + \sum_{i=1}^{N} log\left|det\frac{\partial f_i(x_{i-1})}{\partial x_{i-1}^T}\right| \tag{3.11}$$

where each of the Jacobians' determinant can be computed from equation 3.8, which on taking logarithms reduces to summation of the contributions from the scaling functions of the i^{th} coupling layer. Equation 3.11 is the basis for the training of the ML model.

3.4 Training of the model

3.4.1 Data samples

Performance of a ML algorithm heavily relies on the type of data it is trained on, because it is the features of the training dataset which is to be learned by the model, i.e. searching for a set of weights which gives minimum loss . The dataset used for training of the model consists of EFT Higgs signal events in the mass ranges for 88 cWW parameters (see table 2.3) each with 8k to 9k events. So for training of the model, the input vector is chosen as the three vector $\overrightarrow{x} = (m_{4\ell}, m_{Z1}, m_{Z2})$, each being parametrised by the scalar parameter θ = cWW.

In order to use these data for training, the data needs to be preprocessed first. Because we would be training a model which would map the mass distributions to a standard gaussian distribution, it would be beneficial if the data can be standardize accordingly. The Higgs peak is expected around 125 GeV, so the Higgs invariant masses would be translated about this value. Similarly the Z_1 invariant mass is translated around 90 GeV. So the minimum possible value in Higgs data and Z_1 data after this translation is around -50, which can be normalised to -5 by a division of 10. So the $m_{4\ell}$ and m_{Z1} are all

normalised to be in the range of -5 to +5. From the Z_2 distributions of chapter 2, the distributions are skewed to the left, equivalent to chisquare distributions. Hence for Z_2, the data points are translated about 17.5 GeV (and not 35 GeV) so that the minimum stays close to zero. Thus if x_i' represents the final normalized inputs obtained, we have

$$\begin{aligned} x'_{4\ell} &= \frac{m_{4\ell}-125.0}{10} \\ x'_{Z1} &= \frac{m_{Z1}-90.0}{10} \\ x'_{Z2} &= \frac{m_{Z2}-17.5}{10} \end{aligned}$$

The model is thererby trained by using the normalized input three vector $(x'_{4\ell}, x'_{Z1}, x'_{Z2})$.

3.4.2 Loss function

The goal of every machine learning model is to minimize a loss function or cost function. To carry on the analogy of the loss function, we consider negative log-likelihood as calculated using equation 3.11 as

$$\mathscr{L} = -\frac{\sum_{x\in D} log(p_X(x))}{D} \tag{3.12}$$

where we consider a sample of size D with the log-likelihood of x calculated using equation 3.11. The first component of the loss function is the log-probability of the prior distribution called the Gaussian loss, i.e. $p_Z(z)$. Considering the input vector is a 3D vector, one has p_Z defined as

$$p_Z(z_1,z_2,z_3) = \frac{exp(-\frac{z^T z}{2})}{\sqrt{(2\pi)^3}} = \frac{1}{\sqrt{(2\pi)^3}} exp(-\frac{z_1^2+z_2^2+z_3^2}{2}) \tag{3.13}$$

Note that the latent distribution is meant to be a standard Gaussian distribution, so the mean vector μ would be the zero vector because individual z_i would be independent 1D standard gaussians and the generalized variance matrix $\sum$ would be the 3×3 identity matrix. Of course the latent variable z itself would be a function of input x and scalar parameter θ. During the training process, the datapoints are being mapped on this Gaussian distribution. Initially it could start with random mappings which could imply some points getting mapped close to zero some very far away from zero. Thus the loss function would be higher. Because the mapping is far from mean, the associated log-probability on the prior distribution would be more negative which is equivalent to saying large negative log-likelihood. So when the loss function would be minimized via gradient descent, the model would map the points as close as possible to the central mean of the distribution. This is because the corresponding log-probability would be less negative due to the point being more probable.

This pushing of points climbing it up the Gaussian peak causes a saturation during training process. This is because the datapoints getting more and more concentrated about the mean would cover a fixed amount of area of the Gaussian curve, i.e. the concentration of points around the peak would contribute fairly equal amounts of log-probabilities to the loss function making the average to saturate after a point in training (see fig 3.4).

The other component of the loss function is the contribution from the Jacobian, called the Jacobian loss. The backward flow as described in equation 3.6 shows that the log determinant of the Jacobian would be negative sum of the scaling function contributions s (From equation 3.8, this change is equivalent to replacing exp(s) by exp(-s)). As the Gaussian loss gets fixated after enough training, the loss function can minimize further

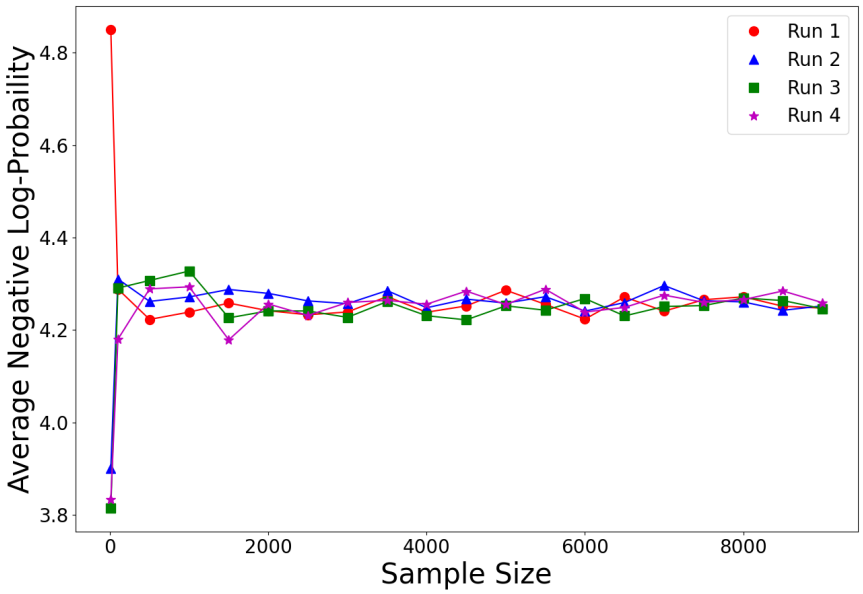


Figure 3.4: Variation of the average negative log gaussian probability with sample size. During the training of the model, when the datapoints are getting more and more mapped around the gaussian mean, the loss contribution from prior distribution converges at a point and beyond that epoch very minor changes are expected to be observed. The runs represent taking different random samples of given random sample size, so for a given sample size the average probability is evaluated 4 times using different samples. For 3D case, this term saturates somewhere between 4.2 and 4.3.

only if the contribution from scaling functions gets more and more negative. This would eventually push the loss function to go below the zero in the negative regime where probably defining it as negative log-likelihood would not make much sense. There is no limit on the Jacobian part because the flexibility of the scaling function is uncertain and can be modified as per model convenience.

3.4.3 Data splitting

In order to train the model and also to check for extreme fitting cases like overfitting or underfitting, it is a common practice in ML to divide the dataset into a training dataset and a validation dataset. Ideally one divides data into three sets of training, validation and testing but there is also a practice to use validation dataset for both validation and testing [49, 50]. For the dataset concerning our case, each cWW parametrised data corresponds to a sample of kinematic variables (3D ntuples) comprising of 8k-9k events. Hence if we naively follow the usual procedure of clubbing in all datasets as $[(m_{4\ell}, m_{Z1}, m_{Z2}), cWW]$, then there are chances of some of the kinematic variables going in training and validation with same cWW value. If such a splitted data would be used, it would not give valid results because the model has already been trained on a part of the distribution parametrised by that common cWW parameter. Thus the splitting strategy to be implemented here would be to divide the cWW values into training cWW and validation cWW values, and therein take into account the entire associated vector set in that category. Referring to fig 3.5, a 80:20 split is being performed for training and validation sets respectively.

Also, the splitting of data with the choice of cWW should be in such a manner that

the distribution of the parameters stay similar for training and validation sets. This is important to ensure the training or validation not to get biased to a local cWW distribution. Thus the data splitting is being done as shown below:

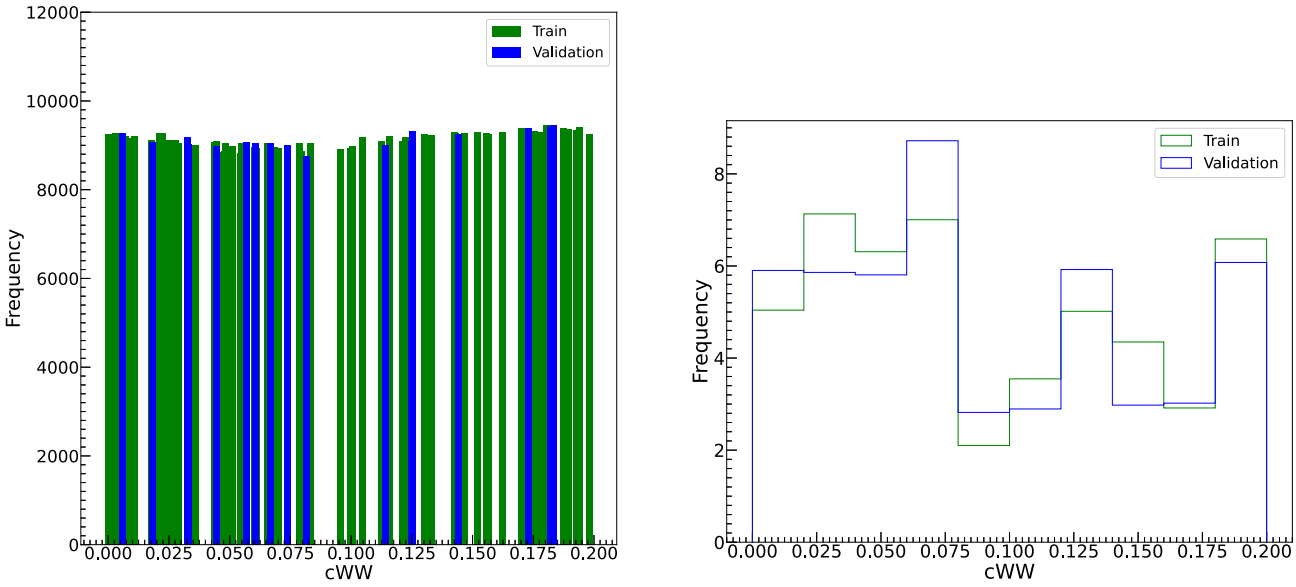


Figure 3.5: Data splitting into training and validation dataset. The left plot shows the discrete cWW values divided into training (green) and validation (blue). The right plot shows the distribution of the cWW values in training and validation dataset. Because of the differences in available values in a local region, the splitting is slightly non-uniform and because of which there are slight differences. But overall the distributions are fairly similar and equivalent. The frequency in the left plot signifies the number of events corresponding to the given cWW in the specific dataset, while the frequency in the right plot signifies the number of cWW samples available in a particular bin of the histogram a local range of cWW values.

3.4.4 Description of the trained model

Using the training dataset as described in previous section, the model was trained for 90 epochs. The details of the model are listed below.

- Number of coupling layers (affine layers) set to 10. To deal with the transformation issue as described in subsection 3.3.3, masking scheme was implemented in such a way that one of the three components of the input vector $\overrightarrow{x} = (x_{4\ell}, x_{Z1}, x_{Z2})$ is left unchanged and other two components changed with respect to scaling and translation functions defined by the unchanged component and cWW. Specific to the training of this model, the masking scheme was kept as [(1,0,0),(0,1,0),(0,0,1)...] $\times \overrightarrow{x}$, where '1' implies that component is not changed. This scheme was repeated in the sequential order shown until we reach the last coupling layer.

- The scaling and translation functions are the ones to be made flexible using neural networks. For a given coupling layer, a specific neural network architecture is followed. Between the input layer and the output layer, the hidden layers are made from dense layers [51] with number of neurons per layer defined by 16 $\times$ [1,2,4,4,2,1], so this means that given an affine layer, the neural architecture of s and t would have hidden layers with 16, 32, 64, 64, 32 and 16 neurons along with

input layer and output layer. The architecture is chosen to ensure as computationally simple of a model as possible.

- Each of the hidden layers are equipped with RELU activation functions [52]. RELU functions have the property to behave as linear functions for positive inputs and setting any negative input to zero. The reason for using RELU is to ensure the model not achieving saturation during training process.

- Model is trained using mini-batch gradient descent approach [53, 54]. Minibatch gradient descent is implemented because of its maximum advantage of rapid learning and faster computations. Compared to computing loss of the entire training dataset and then computing gradient of the loss for weight optimization, its computationally efficient to divide the dataset into small minibatches and update the model weights with each minibatch. In this respect, an epoch is defined when every minibatch has been accessed once and model weights updated with its gradient. For our case, the entire dataset of a particular cWW was taken as a single minibatch. Since the model is updated after every mini batch, the effective training loss is reported as the loss calculated on the training dataset after updating weights from last minibatch. The validation loss is also computed using the validation dataset at this instance to ensure comparison at the same instance of training.

- In order to ensure stability of the training, it is important to ensure that the gradient descent steps to be done in an optimized manner. The gradient descent step is updating weights with respect to gradients calculated, the extent of the update is controlled by the learning rate. A very high learning rate can push the model to instability and a very small learning rate can take the model too long to minimize the loss. To tackle this issue, a learning rate decay strategy was adopted wherein the training starts with a large learning rate to give an initial push to the model and then gradually the learning rate is reduced with each epoch to avoid the training getting unstable due to large updates to the weights (see fig 3.6).

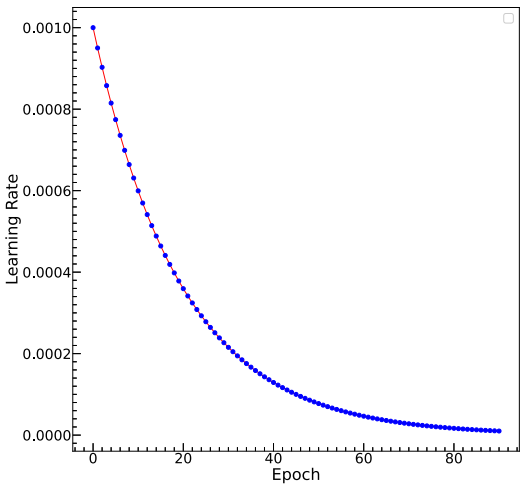


Figure 3.6: Variation of learning rate with epoch for training. Initially a learning rate of 0.001 is kept which is gradually decreased to 0.00001 by epoch 90.

- Optimizers are the components which updates the weights depending on the learning rate. For the training purpose, the generally used optimizer Adam optimizer is used.

The loss plot for the model trained with the details as described above is shown in fig 3.7. Initially the loss was higher for both the datasets which eventually decreases with each epoch. Validation loss is expected to be slightly higher than the training loss because it is the data not seen by the model. Because of the learning rate decay, the learning smoothens from epoch 40 leading to a stable model. Pre epoch 40, the loss is slightly noisy which accounts for because of two reasons:

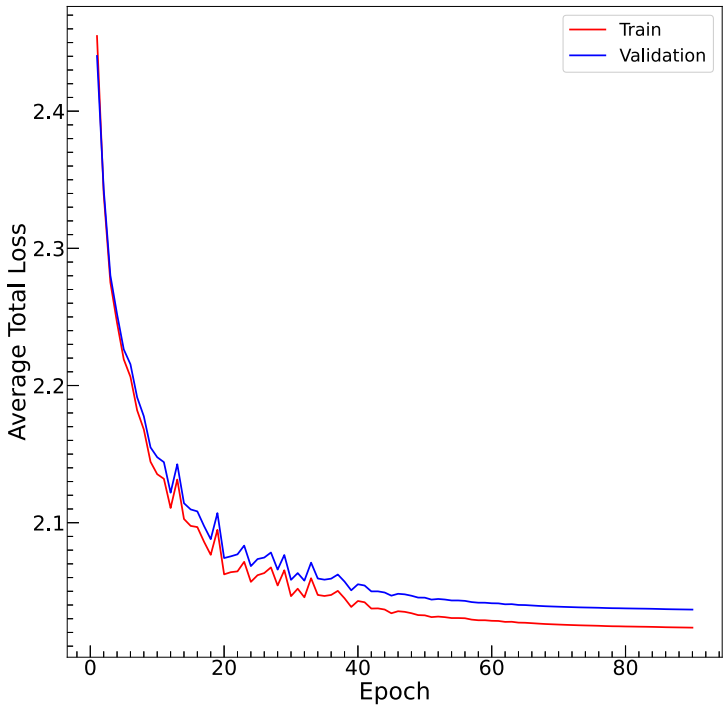


Figure 3.7: Average loss corresponding to trained model as a function of epochs. The loss is calculated using equation 3.12

1. Small noise from the minibatch considered during the training process. Each minibatch in itself is a small model, so if during a training instance one of the minibatches perform poorly its effect would be seen in the effective loss calculated at the end of epoch.

2. Oscillating gaussian and jacobian components of the loss function.

Because there is not any abnormal deviations in the training curves of the two datasets like validation loss (overfitting) or training loss (underfitting) gradually increasing after a certain epoch, the model is devoid of any abnormal fittings and hence can be used for further analysis. Fig 3.8 shows the variation of the gaussian and jacobian loss components with epochs.

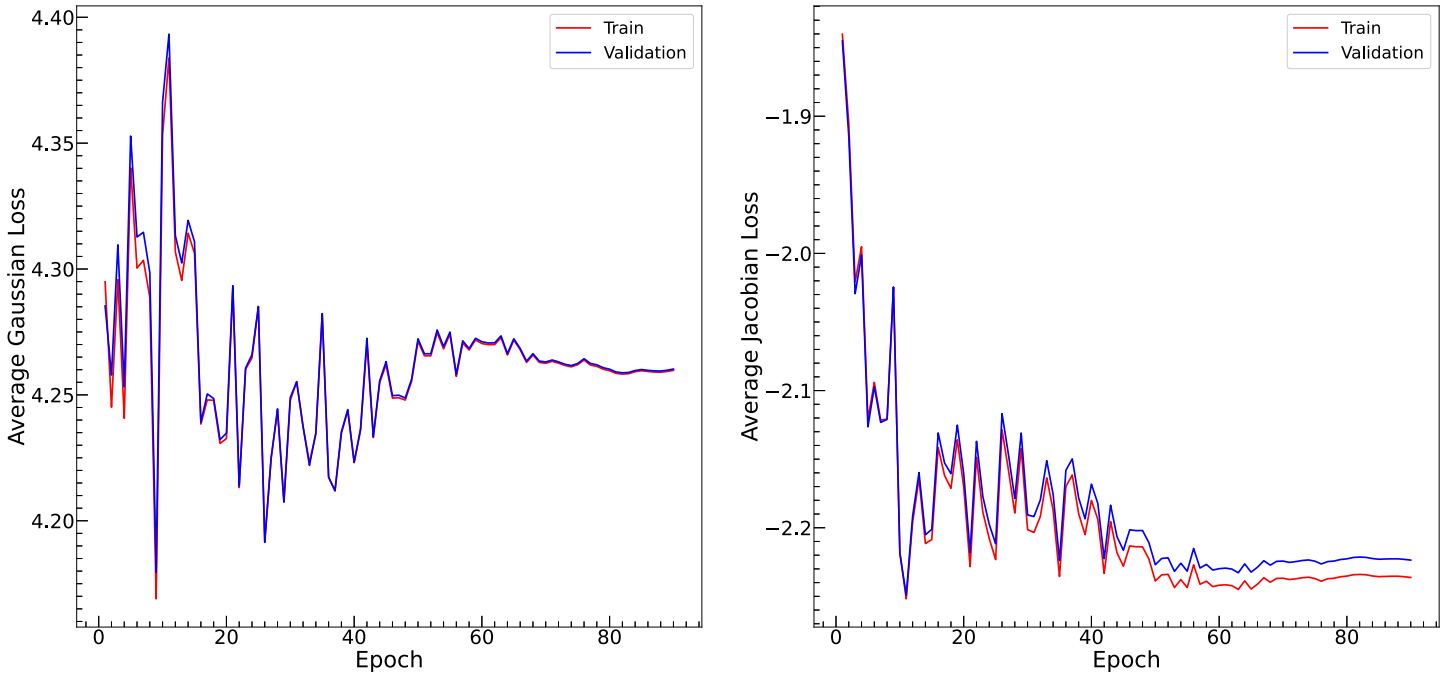


Figure 3.8: Variation of the gaussian (left) and jacobian (right) loss plots with epochs. To keep up with the loss plot of fig 3.12, average negative gaussian and jacobian component is plotted for the training and validation sets. As predicted in fig 3.4, the gaussian loss saturates at a fixed value for both training and validation sets. Contrary, the jacobian loss also got less noisy which is because of relatively small learning rate about that epoch.

3.5 Forward flow model inference

Having the trained model obtained as described in the previous section, one of the first inferences of the model which can be done is to do the forward flow inference. During the training phase, the hyperparameter direction was set to -1, which implies that the mapping is carried out from the original data distribution to a three dimensional (3D) standard normal distribution. For the forward flow inference, the direction of the flow would be reversed, so the hyperparameter direction is set to +1. This means that one takes a random sample of size N from a 3D standard normal distribution and the transformation functions applied would transform it to the expected data distribution. Then the two distributions, true data distribution and model predicted distribution can be compared to evaluate the consistency.

The model has been trained on the dataset $(m_{4\ell}, m_{Z1}, m_{Z2})$ with the values preprocessed to normalise the inputs. For carrying out this inference, firstly predictions are made using inputs of the gaussian distribution, using full sample of the EFT signal (see fig 7.5 in appendix). The predictions made by the model would be normalised back to the observed physical values for the invariant masses. If x_i corresponds to the prediction made by the model for invariant mass 'i', then

$$(m_{4\ell})_{pred} = 10(x_{4\ell})_{pred} + 125$$
$$(m_{Z1})_{pred} = 10(x_{Z1})_{pred} + 90$$

$$(m_{Z2})_{pred} = 10(x_{Z2})_{pred} + 17.5$$

In order to get a quantitative comparison between the true and predicted distributions, there are statistical divergences which can be calculated to compute numerically the similarity between two distributions. Naturally there is not a specific divergence measurement to consider, in our analysis we make use of Kullback-Liebler (KL) divergence [55]. If we define **p** as the true probability density or probability mass function (pmf) in our case since we would be dealing with discrete distributions via histograms and **q** is the model predicted distribution, then the KL divergence is given by

$$D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i) log_e(\frac{p(x_i)}{q(x_i)}) \tag{3.14}$$

where in our case 'i' stands for the i^{th} bin of the histogram considered of N bins. Intuitively KL Divergence act as a measure of information loss happening when we try to approximate the true distribution with the distribution predicted by the model. Note that KL Divergence is not symmetric, and switching p and q would change the interpretation of the results. Ideally, KL divergence is undefined if any of the probabilities in the given bin is zero (which is equivalent to saying there is an absence of a given entry in that bin). If the two distributions are exactly same, then the KL-divergence would be ideally equal to zero. There is no upper limit for dissimilar distributions, only relatively speaking a higher KL divergence indicates more dissimilarity.

Equation 3.14 is used to compute KL divergence for the one dimensional distributions of the invariant masses. As discussed, the binning has been kept in order to avoid the case of an entry of model distribution being empty with corresponding entry in the true distribution is non-zero. If there still exists an empty entry, that bin is ignored from the calculations. The following table lists the binning considered

Distribution	Min Value	Max Value	Range	No of Bins	Bin Width
$m_{4\ell}$	70 GeV	170 GeV	100 GeV	100	1 GeV
m_{Z1}	40 GeV	120 GeV	80 GeV	80	1 GeV
m_{Z2}	0 GeV	80 GeV	80 GeV	80	1 GeV

Table 3.1: Histogram information for the comparisons of invariant mass distributions between the true distribution and model prediction.

cWW	KL Divergence		
	$m_{4\ell}$	m_{Z1}	m_{Z2}
0.00231	0.0055	0.0078	0.0106
0.033	0.0056	0.006	0.00882
0.08154	0.0054	0.007	0.0106
0.12099	0.0056	0.0086	0.0072
0.16213	0.0057	0.0116	0.0074
0.18936	0.0032	0.0096	0.0062

Table 3.2: KL Divergence values for example cWW samples.KL Divergence for each mass dataset was evaluated using equation 3.14

From the KL Divergence plots of fig 3.12 and kinematic distributions of figs 3.9, 3.10 and 3.11, one can see that to a good extent the model has been able to map the invariant

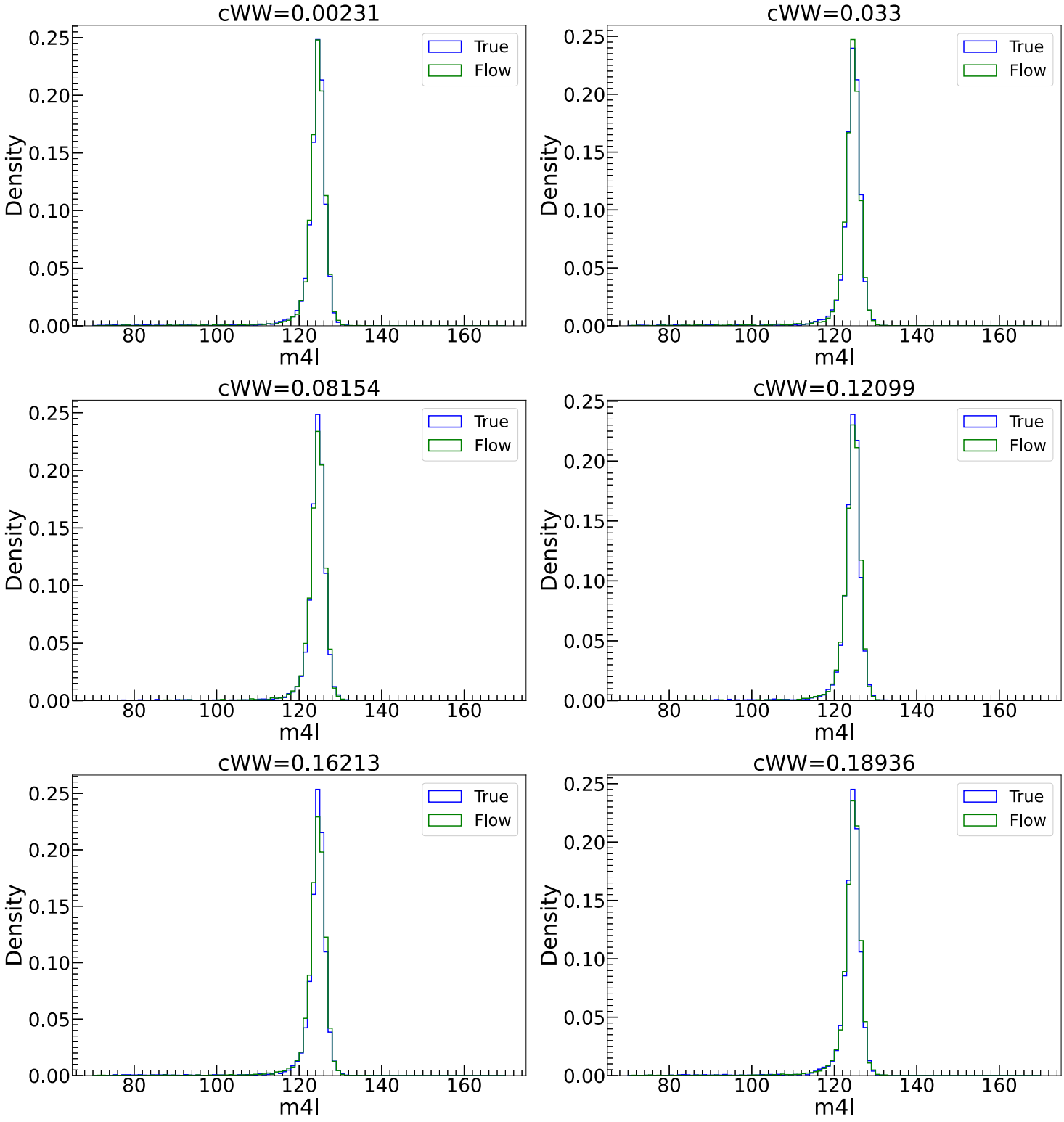


Figure 3.9: Distributions of the four-lepton invariant mass $m_{4\ell}$ for the true data (blue) vs the one predicted by the model (green). The corresponding cWW value is shown on top of the plots. The y-axis shows the probability density.

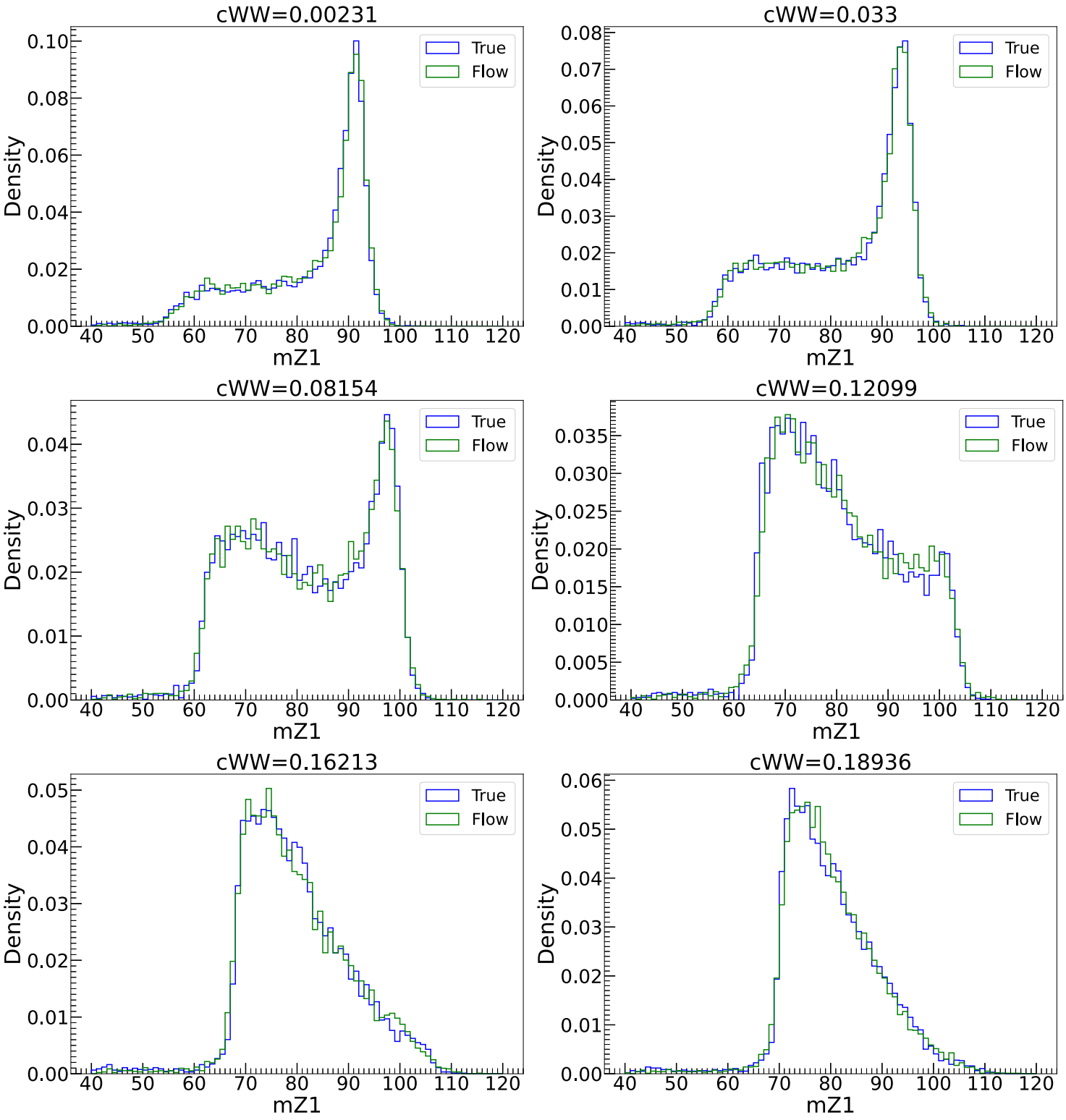


Figure 3.10: Distributions of the reconstructed Z_1 boson for the true data (blue) vs the one predicted by the model (green). The corresponding cWW value is shown on top of the plots. The y-axis shows the probability density.

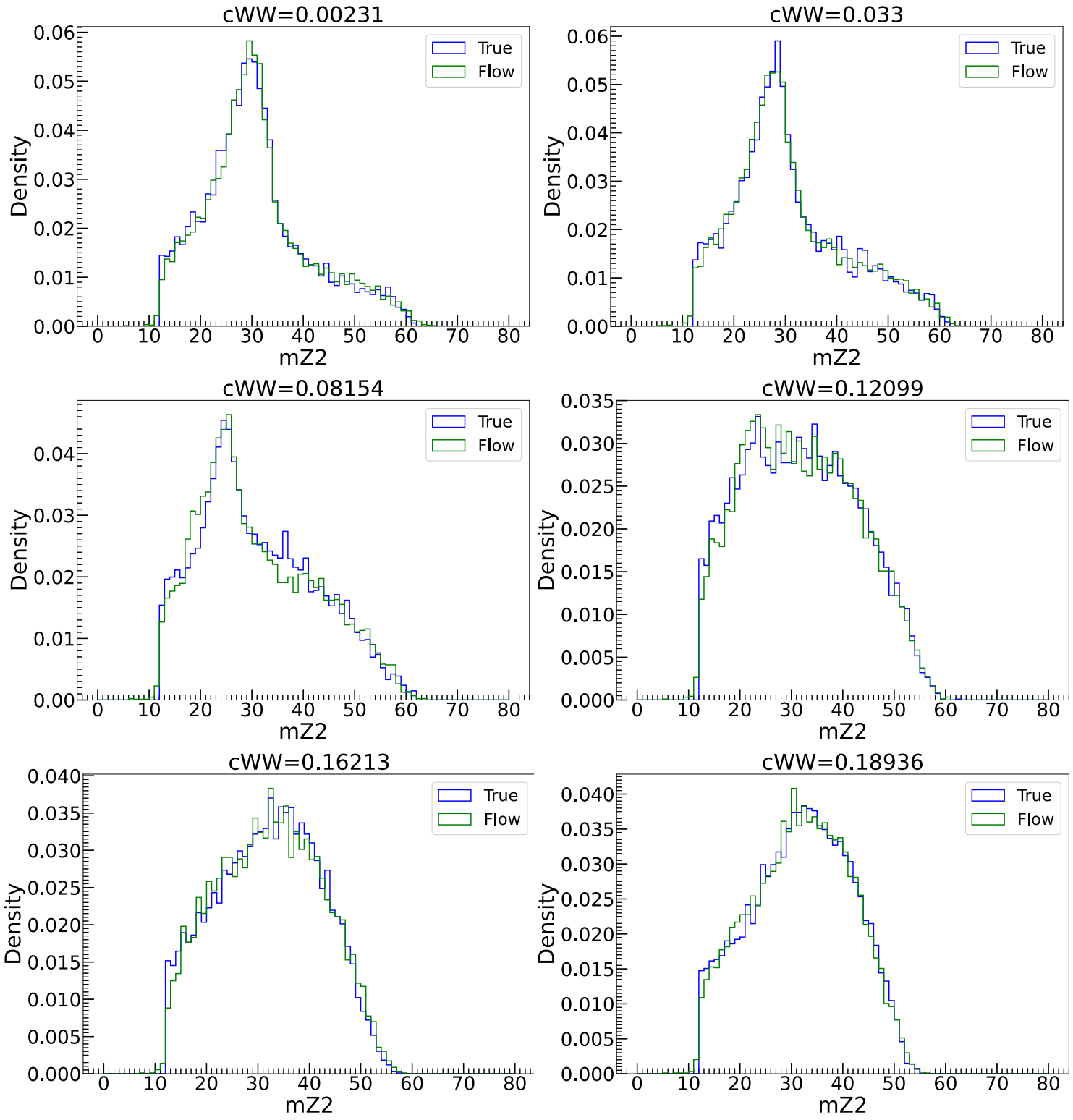


Figure 3.11: Distributions of the reconstructed Z_2 bosons for the true data (blue) vs the one predicted by the model (green). The corresponding cWW value is shown on top of the plots. The y-axis shows the probability density.

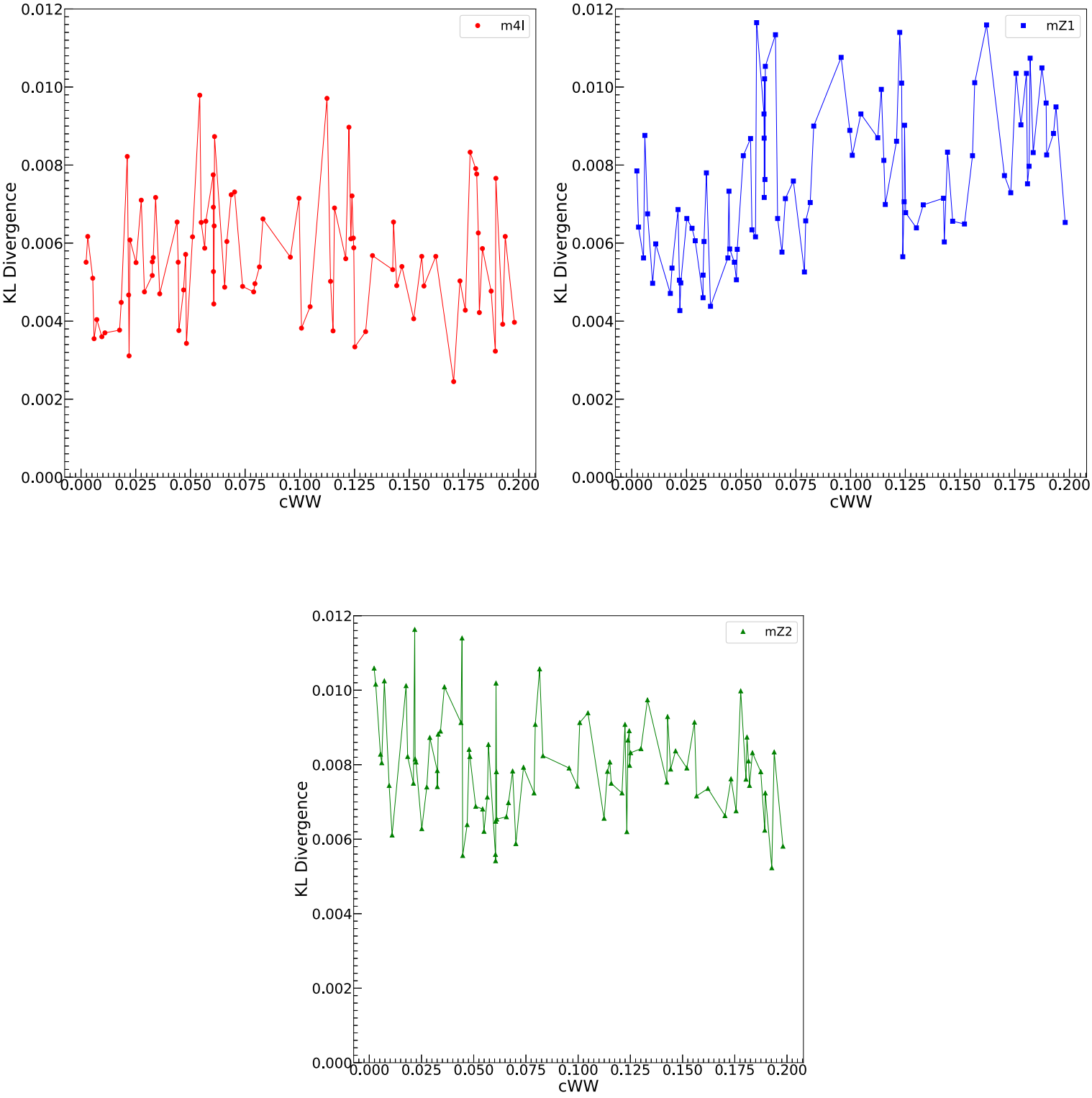


Figure 3.12: KL Divergence for all the cWW datasets for the 1D distributions of the invariant masses $m_{4\ell}$ (upper left), m_{Z1} (upper right) and m_{Z2} (bottom).

mass distributions considering the information loss by model approximation is of order of 10^{-3} to 10^{-2} units. Normally for KL Divergence, one uses natural logarithm or logarithm

with base 2. When used in base 2, the KL divergence values correspond to information loss in units of bits [56]. By comparing the kinematic distributions of $m_{4\ell}$ (fig 3.9), very small deviation about the 125 GeV peak is observed, and it is closest to true distribution. This is expected for the $m_{4\ell}$ case because the EFT Higgs $m_{4\ell}$ peak resembles a Gaussian distribution, so the model functions for $m_{4\ell}$ are equivalent to identity functions.

For m_{Z1} and m_{Z2} distributions (figs 3.10 and 3.11), the functions are not gaussian but from the distributions one can see good resemblance in the shape of the two distributions. The distinctive feature between the distributions of the $m_{4\ell}$ and the m_{Z1} and m_{Z2} bosons is the sharp cutoff around the tails for the Z_2 boson. Because the model is trained on a 3D input dataset (3 + 1 dimensional actually if counted the scalar parameter), the learning of exponentially decreasing tails instead of the sharp cutoff is coming from the Z_1 data. Z_1 has a relatively small density in the low mass range of 40 GeV - 60 GeV which is locally flat. For m_{Z2}, a sharp cutoff around 12 GeV is expected but on account of the decaying tails for the $m_{4\ell}$ and m_{Z1} in the vicinity of 12 GeV, these effect is learned by the model which is visible in m_{Z2} distributions. Expected was a sharp cutoff, but it has the decaying tail effect with relatively very small probability density for masses below 12 GeV.

The main feature of the cWW EFT study is the off-shell switching of the Z_1 boson, which is the area where the cWW comes into play during the density estimation. Overall shape predictions are fairly done well by the model. From fig 3.12, the KL divergence for the individual distributions are summarised. The values are relatively larger for m_Z distributions due to small deviations around the off-shell and nominal SM m_Z peaks.

Chapter 4

Statistical inference tools for parametric estimation

This chapter deals with the formulation of the statistical tools for analysing the properties of the trained ML model. For the sake of simplicity, in this chapter unless specified the normalised input invariant masses $\overrightarrow{x} = (x_{4\ell}, x_{Z1}, x_{Z2})$ as simply x and the EFT scalar parameter cWW as θ.

4.1 Likelihoods for parameter estimation

One of the important mathematical tool required to analyse the staistical model is the likelihood. Likelihoods are defined as the realisations of a probability distribution obtained from a set of data. In a sample of size N, each of the single data entry can follow a particular probability density function (pdf). Usually while sampling data, one considers each of the data point to be independent and identically distributed often abbrievated as i.i.d. This assumption simplifies the calculations a lot compared to points of different distribution source. Also, its quite logical for a set of data obtained from an experiment to have its points following a common distribution unless the point is a classified outlier meant to perform different. In such a scenario, the source distribution of the individual points can be represented by pdfs. One of the ways to estimate the nature of the pdfs is to parametrise it by a set of parameters, and the parameters can be estimated from the data collected by calculating likelihood which is given by:

$$\mathscr{L}(\theta|\mathbf{x}) = p_X(x_1, x_2, \ldots, x_n|\theta_1, \theta_2, \ldots, \theta_k) = \prod_{i=1}^{n} p_X(x_i|\theta_1, \theta_2, \ldots, \theta_k) \tag{4.1}$$

Equation 4.1 is a general expression for calculating likelihoods, which defines it mathematically as the joint probability of the sample points. Note that the likelihood is defined for the parameter θ with respect to the datasample $\mathbf{x}$. The likelihood is shown for a sample of size n defined by a model with k parameters. To simplify the products, take a negative logarithm on both sides of above equation and one recovers an expression of log-likelihood as described in equation 3.11. Likelihood and probabilities are two quantities often used interchangeably, although both are mathematically similar but have completely different interpretations when seen from statistical point of view. Probability is defined as the chances of occurence of sample of events $\mathbf{x}$ with respect to a known model with known parameters θ; likelihood on other hand follows an approach of hypothesizing the parameters θ of the model based on the event sample $\mathbf{x}$. In that respect, likelihood of

parameter set is equivalent to the joint probability of the sample modelled by the same parameter set.

One of the ways to estimate the parameters of the model is maximum likelihood estimation (MLE) (Other approach is known as method of least squares which can be equated to MLE above [57]). The idea behind the MLE is to find a set of parameters such that would maximise the likelihood (or log-likelihood) of the parameter space, which is equivalent to saying minimizing the negative log-likelihood. The set of parameters obtained after this minimization are defined as *estimators* of the parameter. We define the estimators of the parameter θ as $\hat{\theta}$, which in our case are defined by the MLE. Estimators $\hat{\theta}$ has the following properties [58]

- **Consistency** : Estimator $\hat{\theta}$ is said to be *consistent* if the estimator converges to the true value of the parameter with increasing sample size.

- **Bias** : Bias is defined as the difference of the expectation value of the estimator $\hat{\theta}$ and the true parameter θ_o, i.e. b = E[$\hat{\theta}$] - θ_o. An estimator is said to be unbiased if b = 0.

- **Efficiency** : If one has two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, then the estimator with lower variance would be a more reliable and efficient estimator than the other (more on this in section 4.3).

- **Sufficiency** : Given a sample of size N, the estimator $\hat{\theta}$ is said to be sufficient if the joint distribution of sample is independent of the parameter θ

- **Robustness** : $\hat{\theta}$ is said to be robust if its not affected due to presence of outliers or tiny variations around the distribution tails.

MLEs are obtained from basic rules of calculus; the parameter value for which the first order derivative of log-likelihood would be zero and has a negative double derivative of log-likelihood would be the MLE estimate. However, flexibility of the coupling functions of the model makes it tough to have a known analytical expression for evaluating the first derivatives; even if one compute it numerically solving for the roots of the first derivative would be trickier. So one of the ways to look for a rough estimate of the estimator via MLE is to carry out a likelihood scan. Following are the steps to carry out a likelihood scan

- Generate the cWW parameter space Ω by having a large set of cWW values randomly chosen following a uniform distribution in a given range. In our case, random cWW values are picked between 0 and 0.2 (see fig 4.1a).

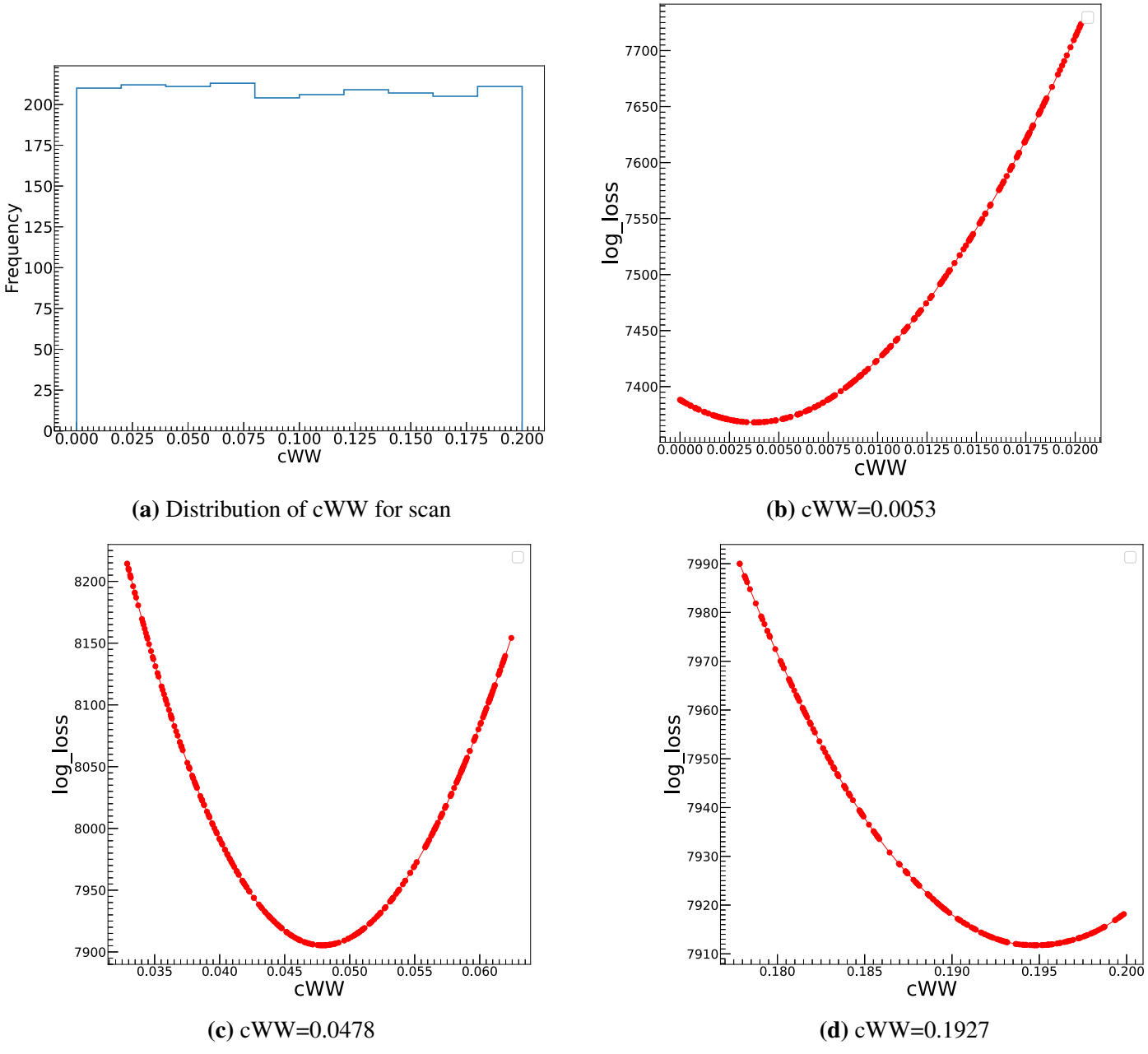
- From the available dataset labelled by the corresponding cWW value, pick a random sample of size N of the invariant mass vector x corresponding to that cWW value. Keep this sample fixed.

- From the parameter space Ω, choose a cWW value and evaluate the total negative log-likelihood (log_loss) for this cWW using equation 3.11 and the model trained as explained in section 3.4.4

$$\ell(\theta) = -\sum_{i=1}^{n} log(p_X(x_i|\theta))$$

ℓ is a simple modification of equation 3.11 wherein sum of the loss is considered instead of the average loss as done in training.

- Repeat this procedure by taking other cWW values from Ω and the corresponding model trained to obtain parabolic curves equivalent to ones shown in fig 4.1.
- The negative log-likelihood has a minima which would be the MLE for θ. The shape of the parabola can change depending on the cWW considered.



(a) Distribution of cWW for scan

(b) cWW=0.0053

(c) cWW=0.0478

(d) cWW=0.1927

Figure 4.1: Likelihood scans carried out for the cWW samples using the scan distribution shown in (a), the y-axis shows the number of randomly picked cWW values in the given bin. (b),(c) and (d) shows the negative log-likelihood for three sample cWW data samples, with cWW on x-axis and total negative log-likelihood on y-axis.

Because the mapping is done on a gaussian distribution, with the Jacobian controlling the way points would be mapped, the negative log-likelihood resembles to negative log-likelihood of a gaussian distribution in parameter space with mean equal to the MLE (or

true parameter). So if one takes the negative log of gaussian probability, the locus of points obtained would be that of a parabola with minimum around the MLE.

4.2 Confidence intervals for estimated parameters

The MLE estimator gives the rough estimate of the parameter which minimizes the negative log-likelihood based on the ML model. If the estimator would be unbiased, then the estimator would land on true value, else there would be some bias. Also, the prediction by the estimator not only depends on the size of the sample but also on the nature of the sample considered. The estimator can predict a different parameter value if we change the sample keeping its size fixed. Hence, there exist a bounded region on the parameter space Ω where the predictions can lie. This bounded region is whats intuitively considered as the confidence interval (CI).

Confidence intervals act as statistical errors as it defines a range for which the values of parameter are expected to be. Defining ω as a subset of the parameter space Ω such that $\theta_{MLE} \in \omega$ and $\omega\prime$ as the complement subset of Ω wherein $\omega \cup \omega\prime = \Omega$ there are two possible hypothesis which can be devised:

- **Null hypothesis** $(\mathbf{H_0})$: The parameter lies in the subset $\omega \implies \theta \in \omega$
- **Alternate hypothesis** $(\mathbf{H_1})$: The parameter lies in the complementary subset $\omega\prime \implies \theta \in \omega\prime$

Now in order to check the validity of the hypotheses, quantitiative representatives called test statistics are evaluated which are measures to show how much deviations are between the data and predictions of the null hypothesis. Thus the search for the confidence interval reduces to finding this appropriate subset ω which would be defined by two endpoints θ_1 and θ_2. So we look for these two endpoints.

Pertaining to the choice of hypothesis, the test-statistic divides the parameter set into two subsets called the acceptance region for H_0 and rejection region for H_0 [58]. However, human errors are possible and one might end up evaluating the incorrect test statistics or infering the validity of hypothesis incorrectly. Thereby the following two risks are to be considered in this scenario:

- If the null hypothesis turns out to be actually true but is still rejected, a type I error is committed with probability α
- If the null hypothesis actually is incorrect but is still accepted, then a type II error is committed with a probability β.

Given evaluating a test statistic followed by a decison making test, α is also defined as the level of significance of the test.

For finding confidence intervals for the MLE predictions done by our Conditional RealNVP model, the likelihood ratio test is performed. For that, likelihood ratio with respect to the MLE estimator is evaluated as the test statistic

$$\lambda(\theta|x) = \Delta\ell = \ell(\theta) - \ell(\theta_{MLE}) = -\sum_{i=1}^{n} log\frac{p_X(x_i|\theta)}{p_X(x_i|\theta_{MLE})} \tag{4.2}$$

The likelihood ratio test statistic evaluated above has some nice properties which makes it a useful quantity. Given the MLE can be approximated as a Gaussian given

sufficiently large sample of size n, then Wilks' theorem [59] states that the likelihood ratio statistic for a model of k parameters follows an asymptotic chi-square distribution of k degrees of freedom

$$2\lambda(\theta_1, \theta_2, \ldots, \theta_k | x_1, x_2, \ldots, x_n) \sim \chi^2_k \tag{4.3}$$

Our analysis is a simplified case of single parameter estimation so k=1. To define the corresponding region of acceptance, there exists a threshold for likelihood ratio which act as the border for defining the acceptance region. Considering the χ^2 distribution with one degree of freedom given by

$$p_{\chi^2_1}(x) = \frac{e^{\frac{-x}{2}}}{\sqrt{2\pi x}} \tag{4.4}$$

From fig 4.2b, if taken a non-zero $\chi^2_{threshold}$

$$\int_{\chi^2_{threshold}}^{\infty} p_{\chi^2_1}(x)dx \quad < \quad \int_0^{\chi^2_{threshold}} p_{\chi^2_1}(x)dx$$

So if we consider values less than the threshold, it will cover a large portion of the distribution, i.e. more coverage probability. Thus, the set of acceptable parameter values can be defined as the parameters for which the likelihood ratio is smaller than the threshold ratio ($\lambda_{threshold}$) where

$$\int_0^{\chi^2_{threshold}} p_{\chi^2_1}(x)dx = 1 - \alpha, \quad \lambda_{threshold} = \frac{\chi^2_{threshold}}{2} \tag{4.5}$$

So given a level of significance α, the parameter values for which the likelihood ratio falls below the threshold forms the acceptance region while for those larger than threshold forms the rejection region. Other way to look at it is the probability of accepting the null hypothesis when true with the expected true parameter truly in the acceptance region is $1 - \alpha$. For example, if the level of significance is 31.73%, then there is a 68.27% chance for the true parameter to lie in the acceptance region. The following table lists the thresholds for different significance levels.

Significance Level α(%)	$\chi^2_{threshold}$	$\lambda_{threshold}$
31.73	1.0	0.5
10	2.71	1.35
5	3.84	1.92

Table 4.1: Threshold values for defining the acceptance regions given a level of significance α for likelihood ratio test for single parameter model [57].

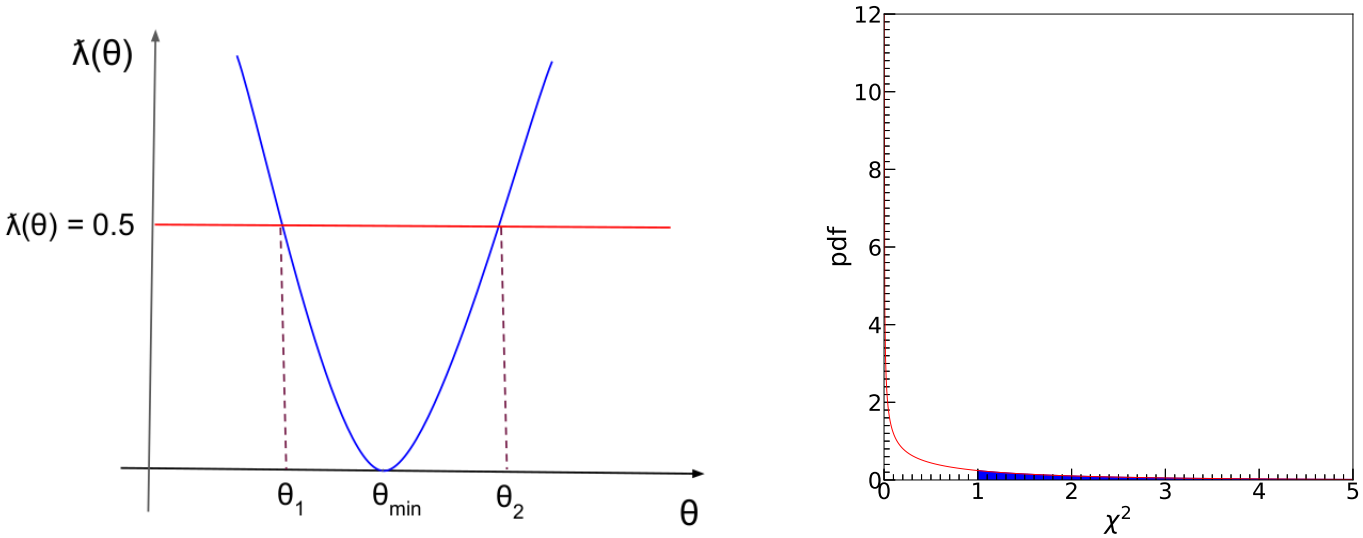
Having defined the threshold value, the acceptance region ω is defined by looking for parameter values θ_1 and θ_2 (assuming $\theta_1 < \theta_2$) for which the likelihood ratio equals the threshold value. The acceptance region is then defined by

$$\omega = [\theta_1, \theta_2] \implies CI = \theta_2 - \theta_1 \tag{4.6}$$

Hence the prediction of the model can be represented as

$$\theta_{pred} = (\theta_{MLE})^{\theta_2 - \theta_{MLE}}_{\theta_1 - \theta_{MLE}} \tag{4.7}$$

with θ_1 *and* θ_2 defined from equation 4.6. Lastly, considering the likelihood scan done for the MLE is rather a discrete approach which approximates continous behaviour, the



(a) **(b)**

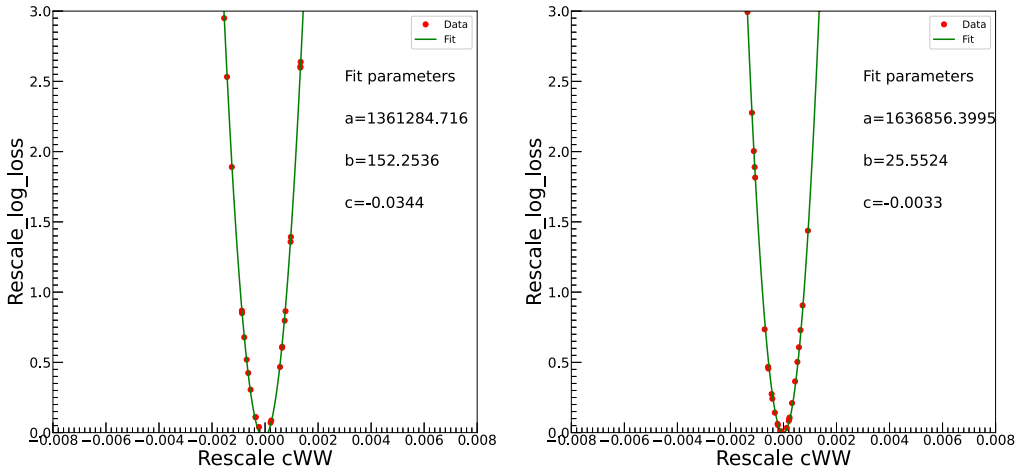
Figure 4.2: (a) Likelihood ratio test with threshold value (horizontal red line) for a LR curve defining the endpoints of the confidence interval. (b) Pdf of a chisquare distribution with one degree of freedom, with shaded portion showing $\alpha = 32\%$.

endpoints can be found by approximating the likelihood ratio curve in the threshold neighbourhood by fitting a quadratic curve to it (since its shape resembles a parabola as shown in fig 4.2a). To do that, least squares approach is followed using a general parabolic curve $ax^2 + bx + c$ whose fit parameters a, b, c can be obtained by solving the matrix equation 4.8. Fig 4.3 shows three examples of such fits.

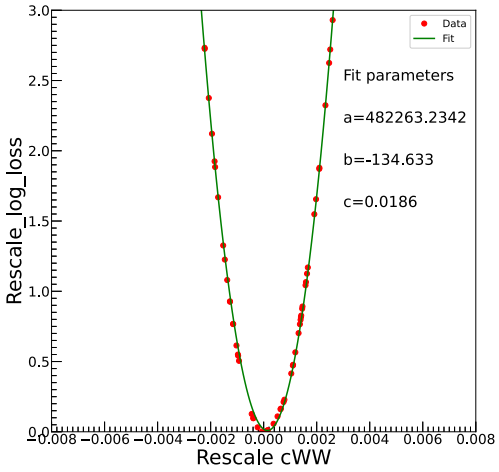
$$\begin{bmatrix} \sum_i x_i^4 & \sum_i x_i^3 & \sum_i x_i^2 \\ \sum_i x_i^3 & \sum_i x_i^2 & \sum_i x_i \\ \sum_i x_i^2 & \sum_i x_i & \sum_i 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \sum_i x_i^2 y_i \\ \sum_i x_i y_i \\ \sum_i y_i \end{bmatrix} \tag{4.8}$$

The interval endpoints are finally obtained as the solutions to the quadratic equation.

$$ax^2 + bx + c = \lambda_{threshold}$$



(a) cWW=0.0053 **(b)** cWW=0.0478



(c) cWW=0.1927

Figure 4.3: Least square fits to the likelihood ratio curves obtained for three example cWW datasets. Rescale log loss is the likelihood ratio (λ) evaluated with respect to the MLE, and rescale cWW is just translating the cWW about the MLE, i.e. $cWW - cWW_{MLE}$. Fitting is done for a region where likelihood ratio threshold lies.

4.3 Analytical expression for the confidence interval

The estimation of the confidence intervals via likelihood scans is a computationally long process and depends on the choice of the parameter space chosen. Even if it is not scanned through the entire parameter space, one needs to scan through sufficient points to ensure parameters with likelihood ratios upto the threshold value. For that purpose, an analytical expression of the confidence interval is appreciable; confidence intervals are required for the possible range for the true parameter.

One of the interesting properties of the estimators is consistency, which defines the approach of the estimation close to true parameter at asymptotic limits. Central limit theorem dictates that at asymptotic limits the sum of the negative log-likelihoods approaches a normal distribution of mean given by the MLE for the sample. The idea is to evaluate the standard deviation σ for this Gaussian distribution. By a simple translation about the MLE evaluated, the asymptotic result implies the variable $\theta - \theta_{MLE}$ follows a normal distribution $\mathcal{N}(0,\sigma^2)$ distribution.

If we define the total log-likelihood using equation 3.11 wherein

$$\ell_p = \sum_{i=1}^{n} log(p_X(x_i|\theta))$$

we have

- **Score function** : $U(X;\theta) = \frac{\partial}{\partial\theta} log(p_X(x|\theta))$. For the estimator to be consistent, the expectation value of the score function approaches zero at asymptotic limit.
- **Fisher information** : Under asymptotic limits,

$$\sqrt{n}(\theta - \theta_o) \xrightarrow{LargeSample} \mathcal{N}(0, I(\theta_o)^{-1}) \tag{4.9}$$

where the Fisher information $I(\theta_o)$ is given by

$$I(\theta_o) = E(-[\frac{\partial^2}{\partial\theta^2} log(p_X(x|\theta))]) \tag{4.10}$$

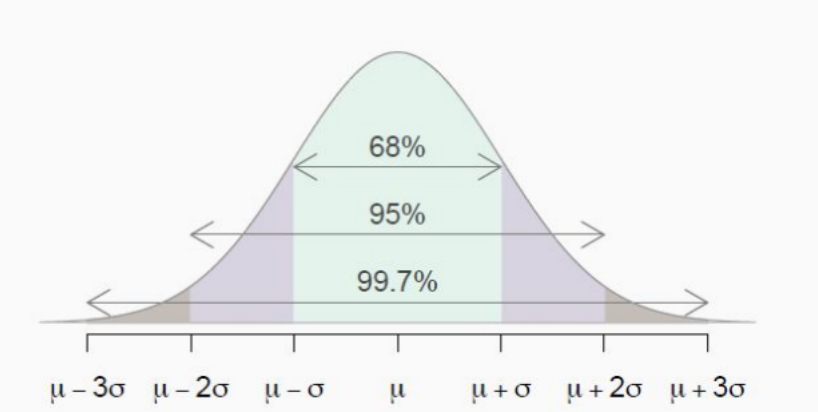


Figure 4.4: Acceptance regions defined for the standard normal distribution.

θ_o above corresponds to the true value of the parameter. Since the true value would be unknown, the true value can be replaced by the MLE and can show that the above relations still hold true. See chapter 5 of [60] for intuitive proofs of above results and relevant discussion.

Fisher information is a quantitative measure for the amount of information which can be obtained from the likelihood of the parameter. More the value of the Fisher information, the more precise the measurement is. From equation 4.9, a small algebraic manipulation will show that the standard deviation of the distribution varies with sample size as

$$\sigma_n = \frac{\sigma}{\sqrt{n}} \tag{4.11}$$

where σ is evaluated using the Fisher information as shown in equation 4.10. The above result forms the basis for estimating the confidence interval by analytical means. Considering the approximate gaussian has the standard deviation σ_n, acceptance regions are defined in a similar fashion as shown in section 4.2, the only difference in this case would be the symmetric nature of the gaussian compared to asymmetric nature of the χ^2 distribution. Fig 4.4 pictorially represents the acceptance regions with the acceptance probabilities $1-\alpha$.

Thus, given the true value of parameter (or the MLE estimate), the interval defined by endpoints θ_1 *and* θ_2 are given by

$$\theta_1 = \theta_{MLE} - z(\alpha)\cdot\sigma_n \quad and \quad \theta_2 = \theta_{MLE} + z(\alpha)\cdot\sigma_n \tag{4.12}$$

$$CI = 2\times z(\alpha)\cdot\sigma_n \tag{4.13}$$

where $z(\alpha)$ are numerical factors defining the allowed deviation for a significance α from the MLE. For eg, for significance level of 32 %, the allowed deviations are $1\sigma_n$. The interval using the gaussian likelihood approximation is symmetric in nature ($\theta_1 = \pm\theta_2$) while the one calculated using the Wilks' theorem is asymmetric. To obtain σ_n, the Fisher

information can be estiamted by approximating the expectation value.

$$\begin{aligned} I(\theta_o) &= E(-[\frac{\partial^2}{\partial\theta^2} log(p_X(x|\theta))]) \\ &= \int -[\frac{\partial^2}{\partial\theta^2} log(p_X(x|\theta)) p_X(x|\theta) dx \end{aligned}$$

From here on, two approximations would follow

- **Integral approximation** : The required integral to be evaluated over the phase space can be approximated by Monte Carlo sums [61]. The model is being used to guess the density p_X for the input invariant mass tuples. So if we take a sample of size n from this distribution (which is equivalent to saying taking a sample of size n from the data we have for the cWW parameter) the integral is approximated as

$$\int -[\frac{\partial^2}{\partial\theta^2} log(p_X(x|\theta)) p_X(x|\theta) dx \approx \sum_{i=1}^{n} \frac{-1}{n} \frac{\partial^2}{\partial\theta^2} log(p_X(x_i|\theta)) \tag{4.14}$$

- **Derivative approximation** : The double derivative of the log-likelihood can be approximated by numerically calculating the double derivative. Considering the Taylor expansions of function f in an ε neighbourhood of x, we have

$$f(x+\varepsilon) = f(x) + \varepsilon f'(x) + \frac{\varepsilon^2}{2!} f''(x) + \frac{\varepsilon^3}{3!} f'''(x) + \ldots \tag{4.15}$$

$$f(x-\varepsilon) = f(x) - \varepsilon f'(x) + \frac{\varepsilon^2}{2!} f''(x) - \frac{\varepsilon^3}{3!} f'''(x) + \ldots \tag{4.16}$$

which on solving gives the nice expression for the double derivative as

$$f''(x) \approx \frac{f(x+\varepsilon) - 2f(x) + f(x-\varepsilon)}{\varepsilon^2} + O(\varepsilon^2) \tag{4.17}$$

Thus the final expression for the Fisher information is given by

$$I(\theta_o) = -\frac{\sum_{i=1}^{n} log(p_X(x_i|\theta-\varepsilon) - 2log(p_X(x_i|\theta) + log(p_X(x_i|\theta+\varepsilon)}{n\varepsilon^2} \tag{4.18}$$

Hence the expression for the confidence interval in terms of Fisher information is given by

$$CI = \frac{2 \times z(\alpha)}{\sqrt{nI(\theta_o)}} \tag{4.19}$$

It should be noted that the analytical expression needs the information of the MLE estimate to evaluate the confidence interval. Because the precision in measurement is expected to get better when the model is trained for certain epochs, the confidence interval becomes a metric to track model behaviour with epochs. In that case, one can approximate the interval by picking the cWW parameter used for training (instead of the MLE estimate) considering the expected MLE is not deviating more than 10% of the true cWW. Note that Fisher information gives symmetric intervals, and that is why it is an approximation. One has to do the likelihood scan to get the required asymmetric interval.

Chapter 5

Statistical inference for the EFT parameter using H$\rightarrow$ ZZ* $\rightarrow 4\ell$ events

This chapter includes the results of the statistical analysis done on H$\rightarrow$ ZZ* $\rightarrow 4\ell$ events using the methods described in the previous chapters and the associated discussion for the same. The first two sections include the analysis done for the full sample using only signal events by considering all the events unweighted (more on this in section 5.3). The last section summarises a small physics analysis done using the EFT samples for signal and SM background for estimation of the cWW parameter at 35.9 fb^{-1} luminosity.

5.1 Model inference using signal events

The first set of results corresponds to the predictions for the cWW parameter depending on the invariant mass distributions. Model predictions are obtained via likelihood scans using the parameter space defined in Fig 4.1a for sample sizes of order 8k-9k events (see fig 7.5 in appendix). Two metrics, namely bias and model variance (confidence interval) are evaluated where

$$bias = cWW_{true} - cWW_{pred}$$

and the confidence interval is evaluated using equation 4.6. Confidence intervals are calculated for a significance of 32%.

Fig 5.1 shows the predicted cWW values using the trained model and the percentage deviation of the true parameter from the predicted parameter. The bias plots in fig 5.2 shows the deviations from true cWW value. From the bias plots, the deviations are observed to be smaller in magnitude but the average bias is less than zero. This implies that the model predictions are generally larger than the expected true values. From the bias and normalised bias (with respect to true cWW) plots, the relative deviations are relatively larger in magnitude for small cWW values but it gradually decreases and approaches zero for higher cWW values. The difference between the small and large cWW values is the inherent deviation from the SM expectations. From the data splitting, the SM case was involved in training. So when considering bias relative to a higher cWW (here 0.0023) the predictions by the model would still be same.

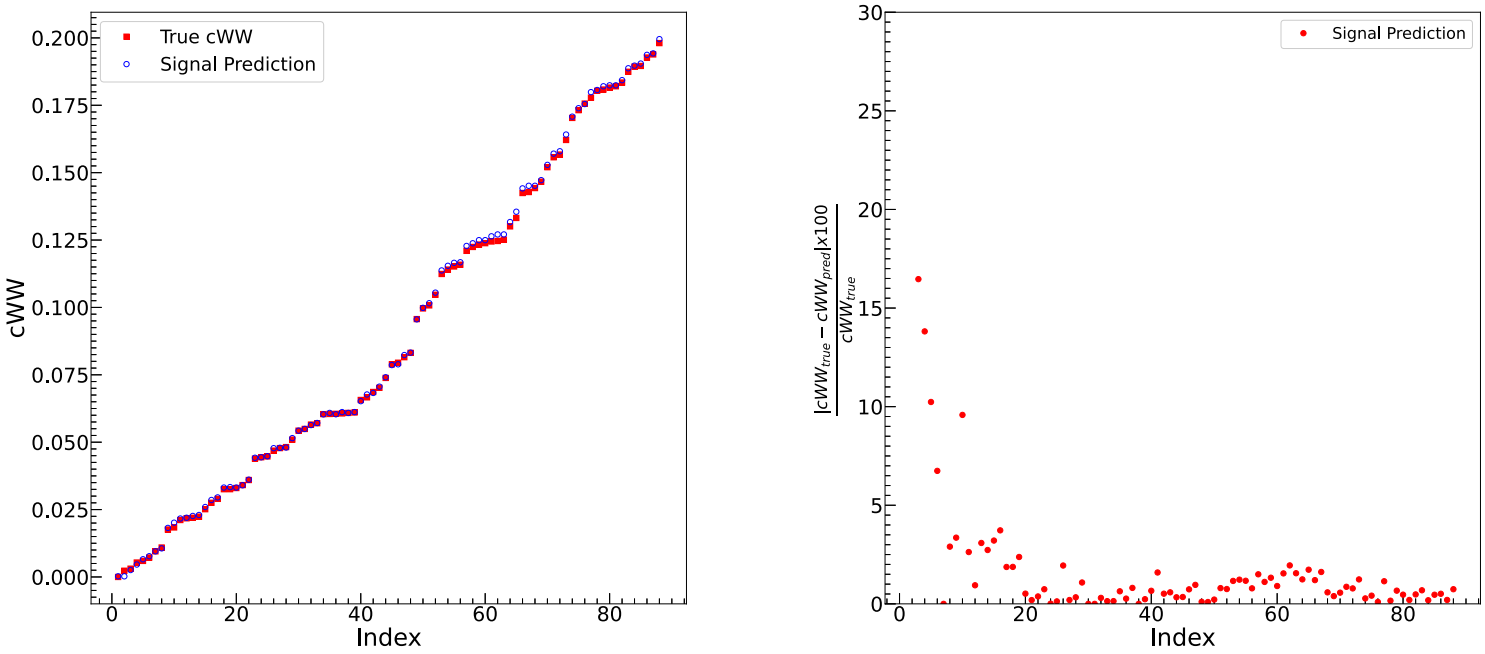


Figure 5.1: Predictions for the cWW parameter by the model with red squares as the true parameter and the blue circles as predictions of model (left) and the relative deviation from the true parameter value (right) shown as percentage change with respect to true cWW (i.e. $\frac{|cWW_{true} - cWW_{pred}| \times 100}{cWW_{true}}$) . For most of the cWW values, the model is making fairly good predictions with ratio close to one on average for majority of the cWW sets.

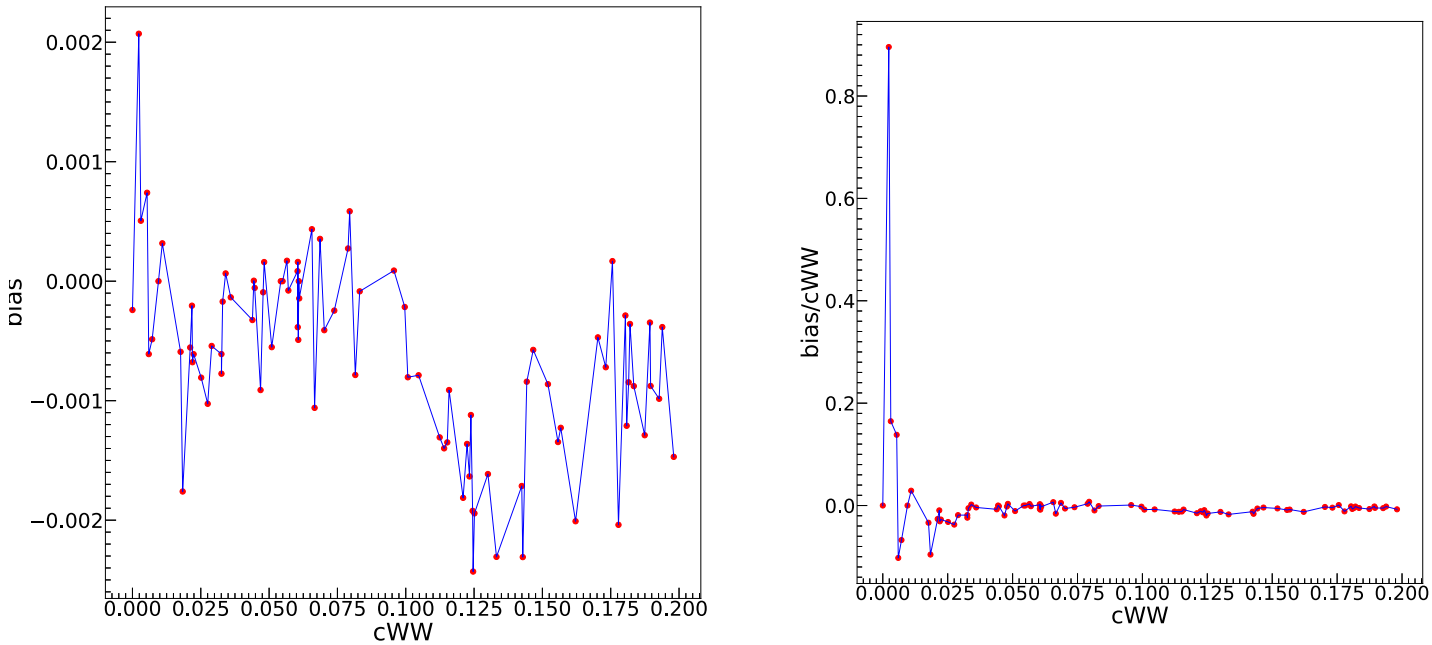


Figure 5.2: Bias plots for the measurement of deviations from the true parameter. A decrease in deviation is observed from smaller to larger cWW. Normalised bias is defined as the ratio of the bias with the cWW parameter, i.e. normalising bias in units of cWW.

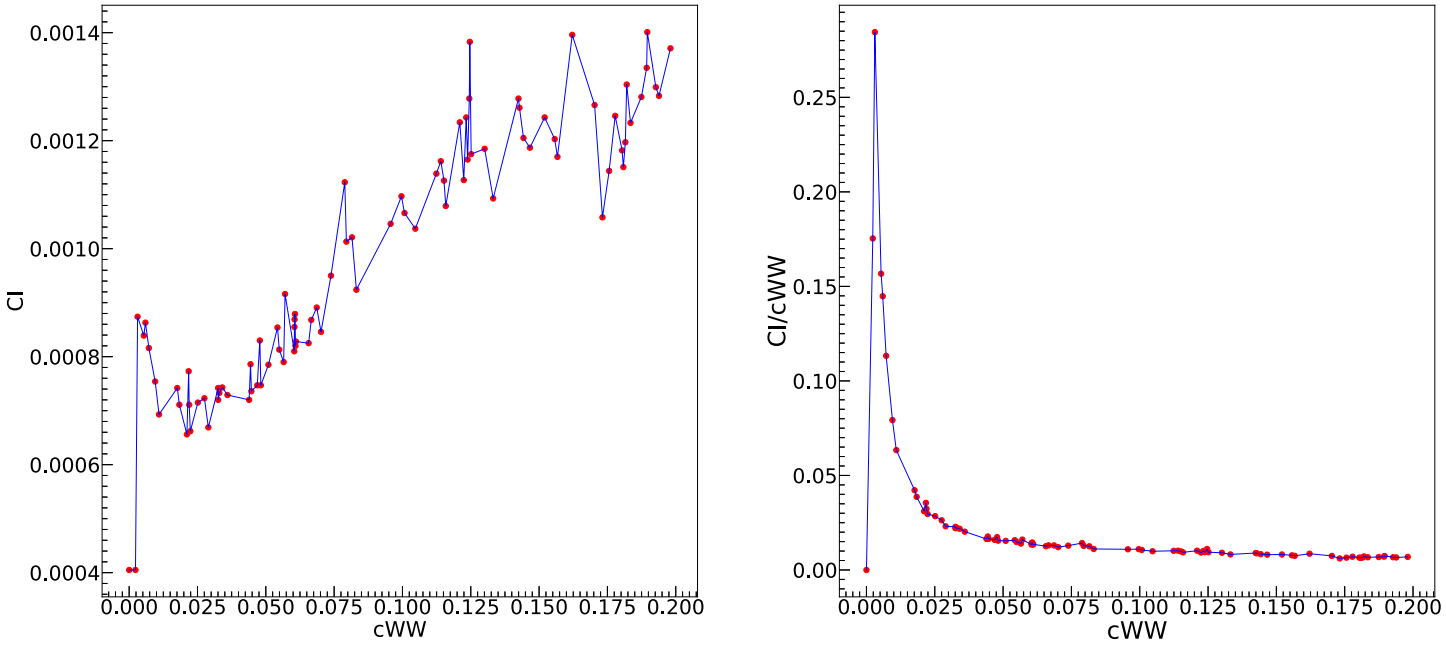


Figure 5.3: Variance plots for the measurement of confidence interval using the likelihood scans. Effectively variance increases with an increase in cWW. Normalised variance (right) is defined as the ratio of the CI calculated to the cWW parameter. So the interval gets normalised to units of cWW. The exact variance is to be considered to make a proper comparison.

The CI plots as shown in fig 5.3 define small intervals for the expected range of predictions. This is on account of the narrow likelihood for a large sample of approx 9k events. An interesting observation is a decrease in the precision with larger cWW. Although the predictions are fairly good due to small bias, but those predictions are less precise compared to smaller cWW values. This implies under 68% confidence, there are relatively more possible values for the parameter in large cWW case. This means the distributions are quite similar for a deviation of order of 0.01 in cWW.

The bias plots showed relatively large bias for small cWW values close to zero, and the sample size accounts for small CIs. The model bias needs to be reduced by more training to account for the very precise confidence intervals.

5.2 Comparative study of the confidence interval analysis

In sections 4.2 and 4.3, two approaches were defined for the measurement of confidence intervals, namely the Likelihood scan approach (**LSM**) as shown in equation 4.12 and the Fisher Information approach (**FIM**) as shown in equation 4.19. In order to compare the two approaches, the standard deviation for the MLE was estimated from the Fisher Information and the confidence interval obtained from the Likelihood scan. The sigmas calculated using both the approaches are plotted for similarities

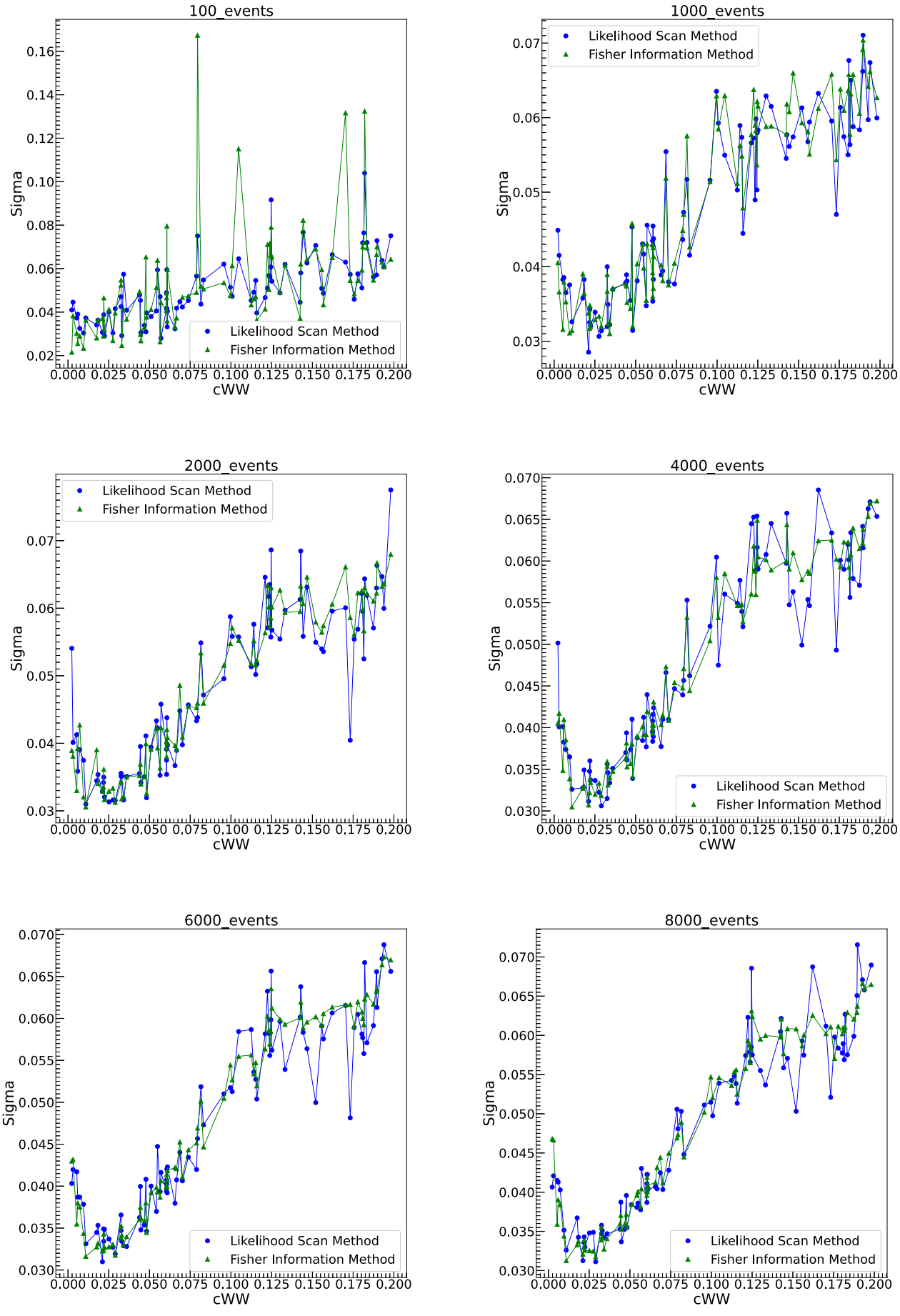


Figure 5.4: Values of sigma for different cWW datasets with varying sample sizes as shown. With increasing sample size, a small decrease in value of sigma is observed.

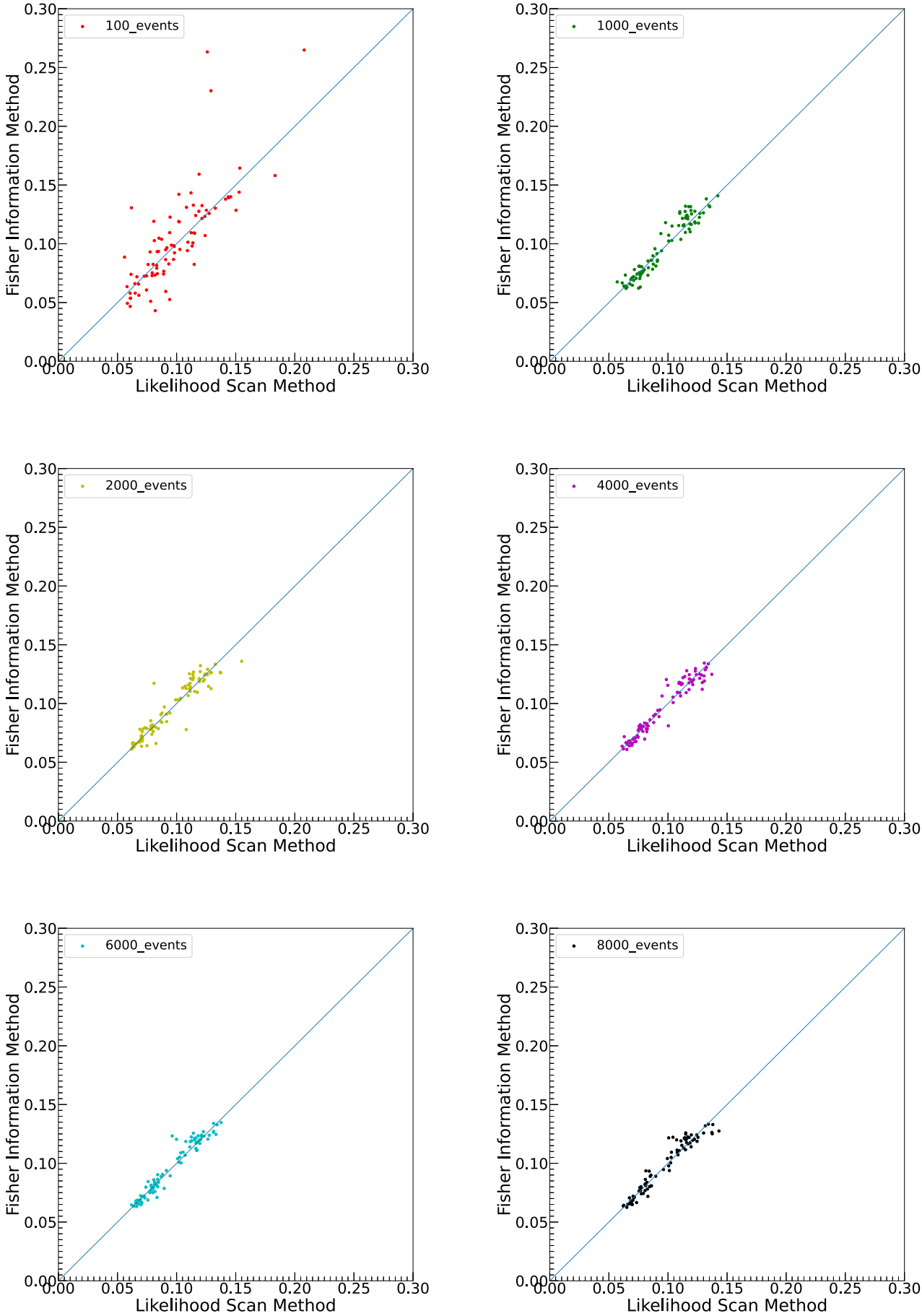


Figure 5.5: Scatter Variance Plots for samples of order 1k events. There is practically no difference in the behavior, apart from the small statistical deviations for some of the sample sizes.

Considering a confidence level of 68 %, for which the interval length is roughly 2σ, we have

$$2\sigma_{LSM} = \sqrt{n} \times CI_{LSM} \quad and \quad 2\sigma_{FIM} = \frac{2}{\sqrt{I(\theta_o)}} \tag{5.1}$$

Fig 5.4 and 5.5 shows the sigma values for different cWW and scatter plots to compare the two approaches for different sample sizes respectively. For large enough samples, both the approaches give equivalent interval lengths with small deviations. The FIM is meant for asymptotic sample limits, so it is accurate as large a sample is. The Wilks' method (LSM) also gets improved a lot with more statistics. As the sample size is decreased, the two approaches starts deviating, with scattering away from the slope 1 line is observed for more cWW samples with decreasing sample size. On average when one reaches a small sample as small as 100 events, the deviations are quite large, with cases having both approaches deviating by large amounts. The FIM fails for small samples because it heavily relies on the Gaussian approximation to the likelihood. The LSM, although gets better with Gaussian approximation does not requires the MLE to be Gaussian. So for such smaller samples also it can be used to measure approximate confidence intervals.

One point to be noted is that since FIM is not quite feasible for small samples, the LSM method takes less time for small samples so it can be used for evaluating intervals. When sample size is large enough, one can use the FIM approach to quickly evaluate approximate confidence intervals using the true cWW.

5.3 Model inference using signal and background events

So far the inference is being carried out without taking into account realistic physical scenarios as observed in LHC experiments. In this section, a prediction analysis is being carried out by considering EFT samples with mixed signal and background events for a luminosity of 35.9 fb^{-1}. For the analysis of mixed inference, the required samples would be passed with some weights w_i (and not unweighted as being done in until now). Two scenarios are being considered for the analysis at same luminosity:

- Weighting signal EFT samples with SM cross-section and background with SM cross-section.
- Weighting signal EFT samples with EFT cross-section and background with SM cross-section.

A small modification is done in the calculation of the log-likelihood. The effective negative log-likelihood is now given by

$$\ell_{eff}(\theta) = w_{sig}\ell_{sig}(\theta) + w_{back}\ell_{back}(\theta) \tag{5.2}$$

where w_{sig} and w_{back} are the signal and background weights evaluated at a given cross-section (SM or EFT) at 35.9 fb^{-1}. The weight calculation is explained in chapter 3. The negative likelihood ℓ is evaluated using equation 3.11 for signal and background samples. Because of the weights, the effective number of events would get reduced which would bring a change in the precision of the measurement(small sample effects).

From the prediction plot as shown in fig 5.6, the predictions did not change on changing the cross-section for pure signal events, this is expected because the only change that would come would be in the weights considered. However, deviations are observed when the model is inferred on the mixed sample of signal and background events. The deviation

from true parameter is small for small cWW values, this is because for small cWW the EFT resembles the SM case, and addition of background just scales the distributions by a multiplicative factor equally in all bins (see the distributions in chapter 3). Fig 5.7 shows the bias and confidence interval plots for the mix inference. From that also, one can see that both precision in the measurement and prediction worsens with increasing cWW.

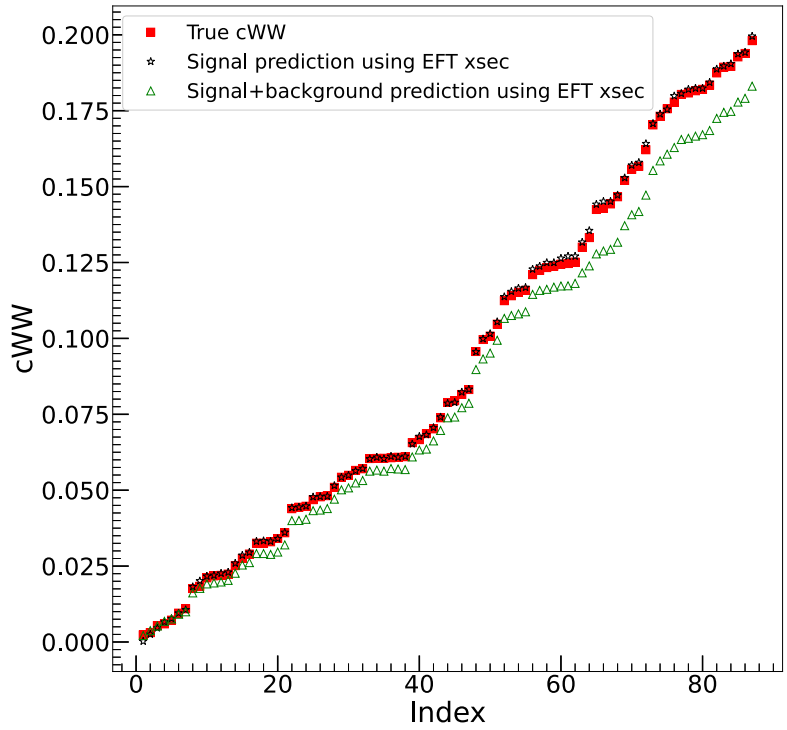


Figure 5.6: Prediction plot for the mixed signal and background inference for the model trained by the method described in chapter 4 for EFT cross-sections at a luminosity of 35.9 fb^{-1}.

However EFT signal starts deviating as cWW increases; but because the background is kept at the SM case the background mixing locally changes the distribution around the 90 GeV mass peak for the Z_1 distribution and for the Z_2 at around 30 GeV to 40 GeV. This change in local density pushes the net distribution to be similar to a EFT distribution of smaller cWW. This is because the off-shell condition grows with cWW, so a local increase in density is equivalent to reduction in the off-shell effect. This pushes the model to make predictions smaller than what is expected. From the bias and variance plots, we see that not only the prediction worsens by a large amount for small cWW but precision also gets worse. Larger samples are heavily affected by background due to this local density change, which the model has not been trained on to fit this local density change.

Hence, the model trained on a pure signal sample although performs really well for the pure signal scenario but the model needs to be retrained to make for precise predictions for the realistic mixed signal and background case.

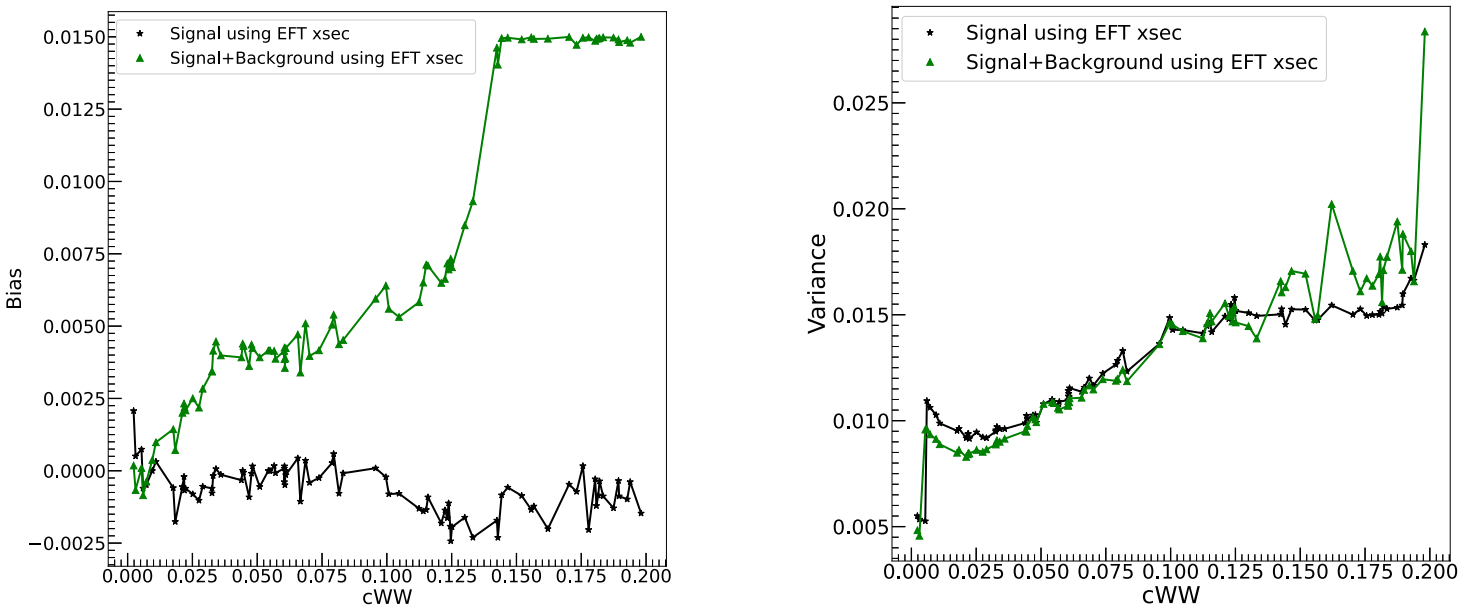


Figure 5.7: Bias (left) and variance (right) plots for the mixed signal and background inference for luminosity of 35.9 fb^{-1} for EFT cross-sections.

So in order to train the model with both signal and background events in the samples, the following modifications would be done in the model:

- Events from both signal and background would be mixed taking into account the signal to background ratio. Due to lack of sufficient number of background events, weights as calculated in chapter 2 would be considered to achieve this task.

- EFT samples used for training and validation would be mixed with same background sample. The main focus of this training would be to model the effect of cWW on background and use it to modify the mapping on the prior gaussian distribution.

- The loss function is modified, because now we have to take into consideration the signal to background ratio between EFT signal and SM background events, as mentioned in chapter 2. So the loss function for this model gets modified inspired from equation 5.2 as

$$\mathscr{L}_{mix} = -\frac{\sum_{i=1}^{n} w_{sig}^{i} \times L_{sig}^{i} + w_{back}^{i} \times L_{back}^{i}}{\sum_{i=1}^{n} w_{sig}^{i} \times N_{sig}^{i} + w_{back}^{i} \times N_{back}^{i}} \tag{5.3}$$

  where summation is carried out over corresponding cWW datasets in the training sample (or the validation sample). w corresponds to the weights for the given signal or background event, and L corresponds to the negative log-likelihood corresponding to the signal or background event calculated using equation 3.11. N corresponds to the number of events in the corresponding sample considered.

- Most of the model architecture as described in subsection 3.4.4 is kept the same, except changes are made in the learning rate scheduler to ensure as stable of training as possible.

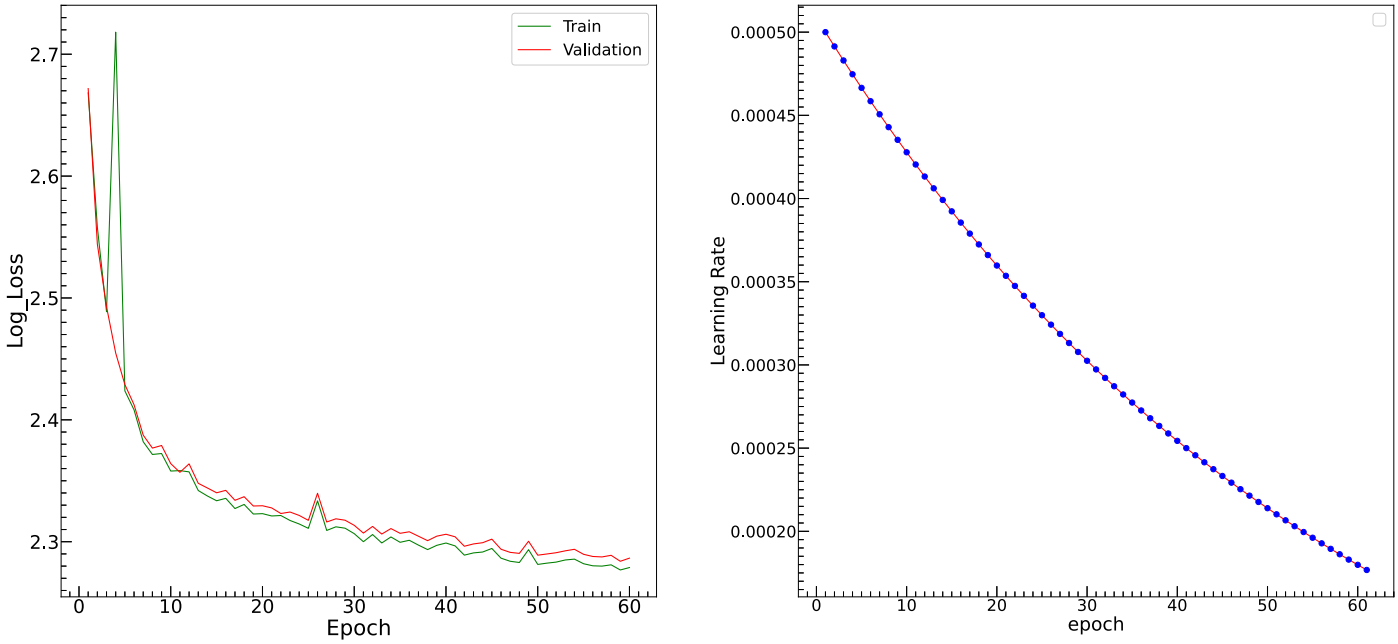
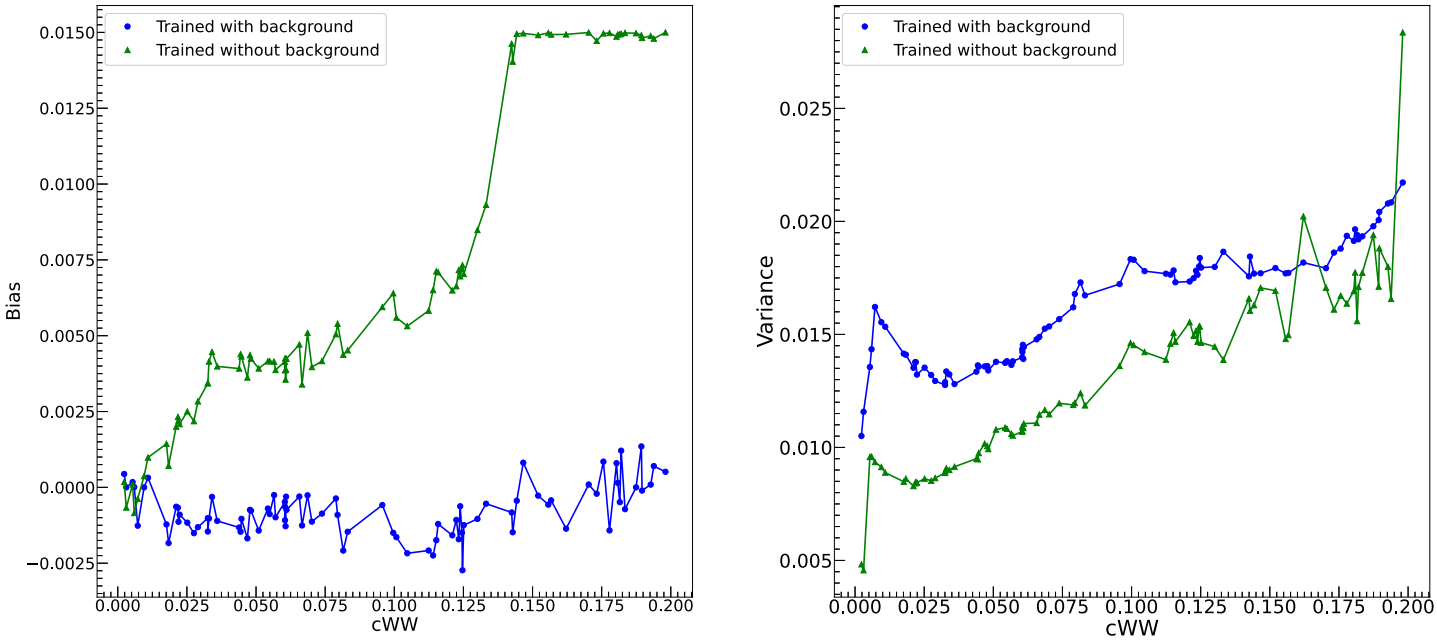


Figure 5.8: (a) Average loss plot with epoch with loss calculated using equation 5.3 (b) Learning rate scheduler for training of the model.

.



(a) **(b)**

Figure 5.9: Bias and variance plots for mix inference comparing two models with different training conditions. Samples weighted using EFT cross-sections at a luminosity of 35.9 fb^{-1}.

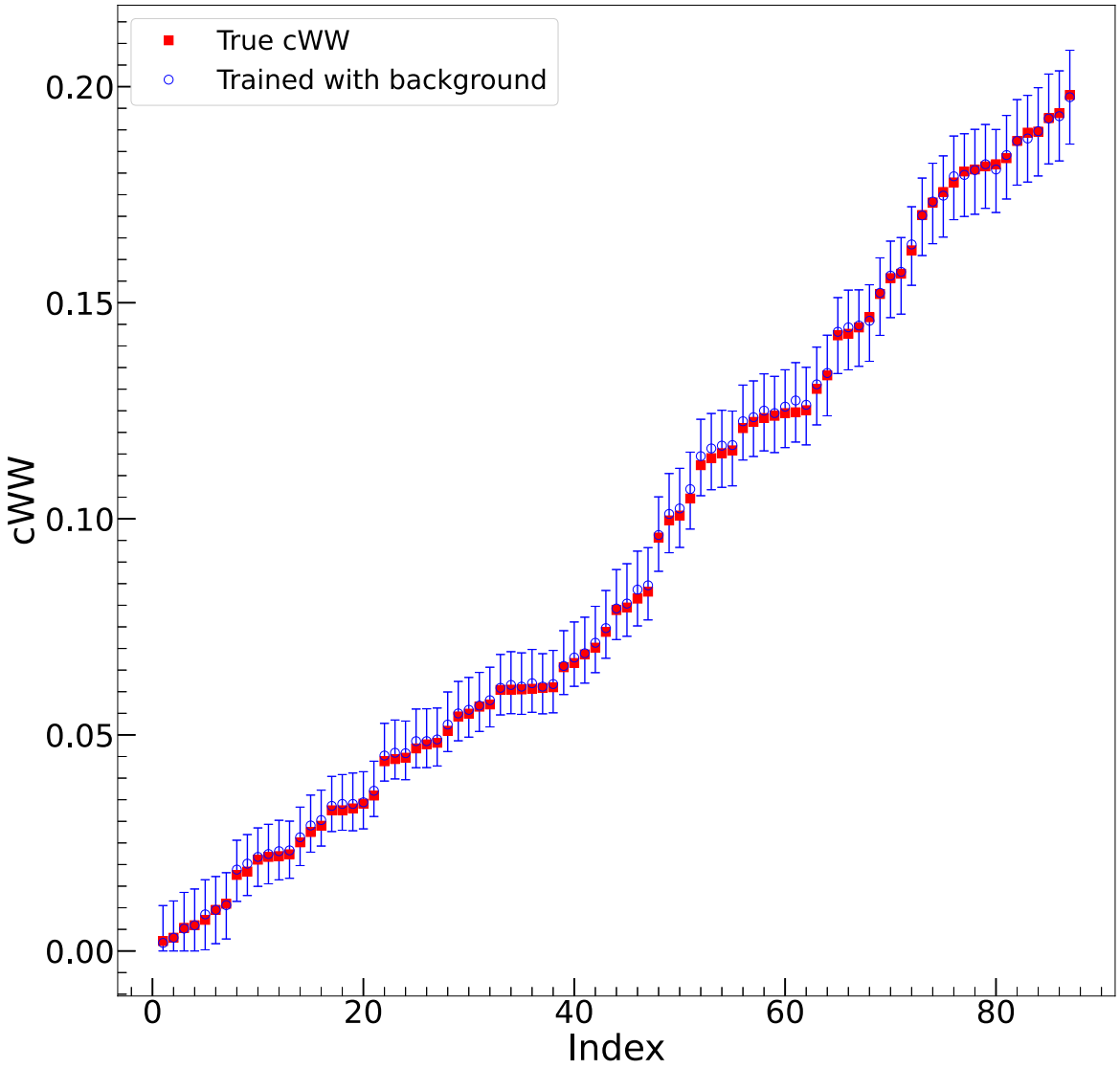


Figure 5.10: Prediction plot with error bars representing the uncertainties in predictions (black bars are for model trained with background and orange for not trained with background). Samples weighted using EFT cross-sections at a luminosity of 35.9 fb^{-1}.

Fig 5.8 shows the loss plot and learning rate scheduler used for training. The training was still a bit noisy, but apart from that there are no signs of abnormal fittings, so it can be used for inference. Fig 5.9 shows the bias and variance plots comparing the two models with different training conditions. From fig 5.9a, prediction improves when model is trained with background with cost of slight increase in interval length as shown in fig 5.9b. Fig 5.10 shows the prediction plots for all cWW with interval lengths as error bars. One can see that as long as small cWW parameters are concerned, both models perform in a similar fashion. When trained with background, it also is able to make good and precise predictions for large cWW (compare blue circles of fig 5.10 with green triangles

of fig 5.6). Lastly from fig 5.11, we see that indeed the predictions made are within 1σ deviation from the MLE for the model trained with background.

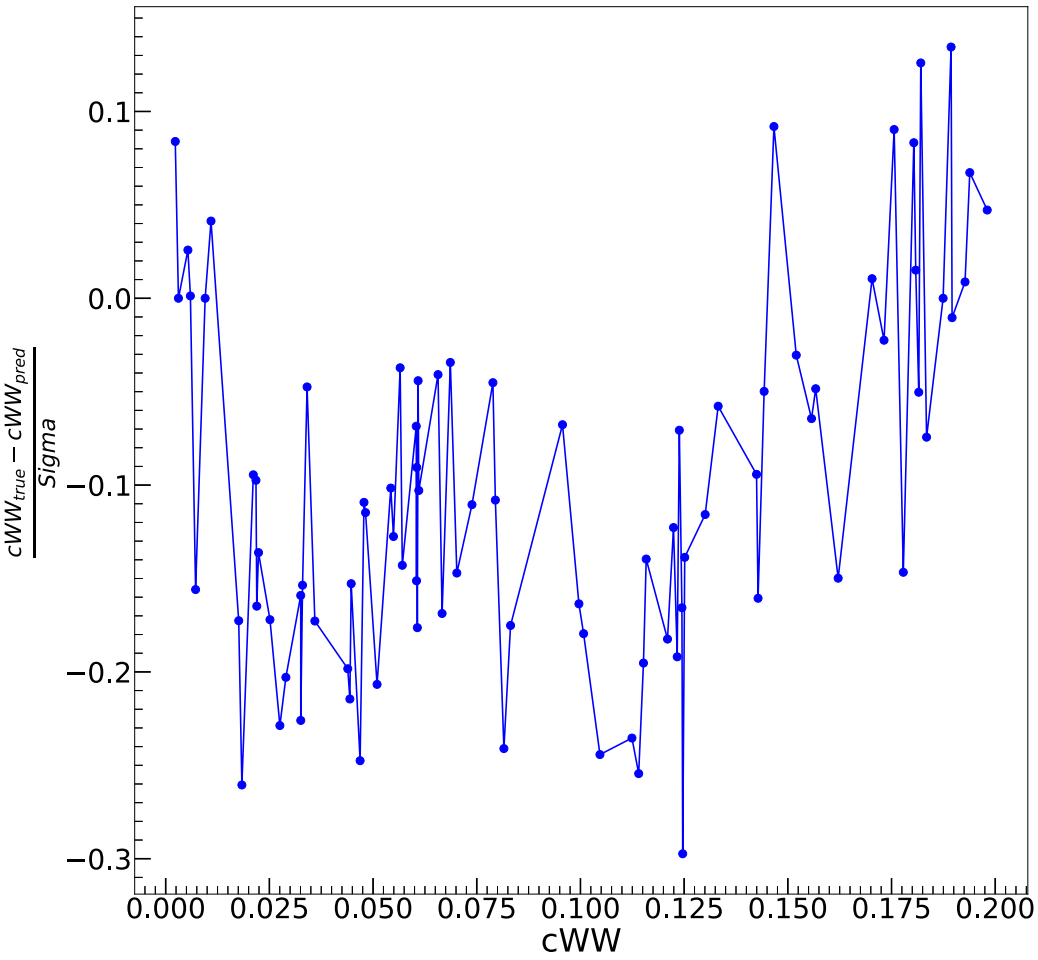
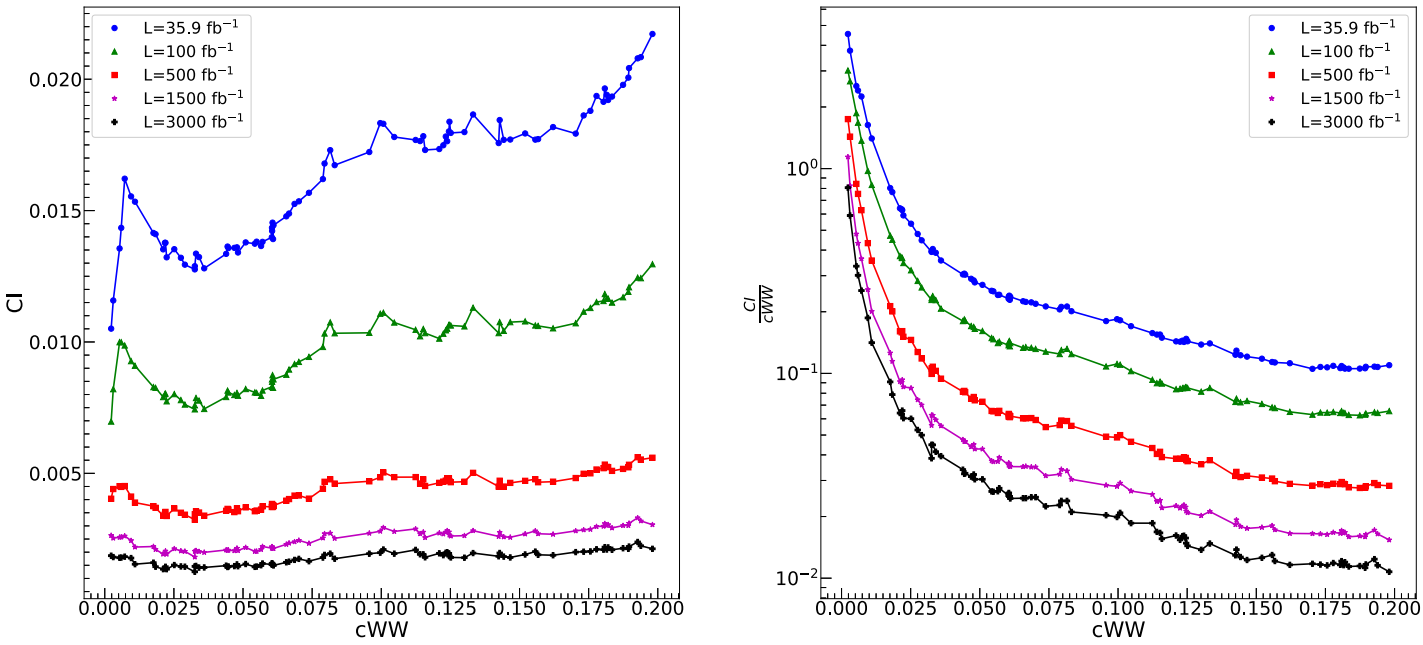


Figure 5.11: Ratio of bias to σ for model trained with background. σ is evaluated from CI using the fact that the interval length corresponds to 2σ. The deviations from true parameter lie within 1σ range, implying good prediction by the model. Samples weighted using EFT cross-sections at a luminosity of 35.9 fb^{-1}.

It should be noted that the mixed model inference done in this section is approximate because the samples are weighted using EFT weights. In reality, the events are to be passed unweighted with event sizes modelled by weights instead of supplying weights to the log-likelihoods. Based on this approach, the effect of weights will not be observed in the MLE predictions (as long as the parameter space is not changed) with luminosity but the length of CI would decrease increasing precision in measurement with luminosity. This is the effect of increase in effective sample size with luminosity.

Fig 5.12 shows the variation of the confidence intervals with luminosity, which represents the effect of the sample size. One can see that with increase in luminosity, the statistical error decreases. The decrease is inversely proportional to $\sqrt{N_{eff}}$ with N_{eff} being the effective sample size of the mixed signal and background sample. From figures



(a) **(b)**

Figure 5.12: (a) Variation of confidence interval with luminosity (b) Confidence interval normalised with respect to cWW measured. The precision in measurement increases with luminosity as expected, due to narrowing of the log-likelihoods evaluated on account of large sample size

5.12a and 5.12b, one can see that although the variance increases with increase in cWW, the normalised variance with respect to the cWW decreases with luminosity. This implies that the probability of discarding higher cWW values from measurement increases with luminosity, because the small interval makes these measurements distinguishable and distinct from SM. Due to the resemblance with SM and large intervals, the small cWW values cannot be discarded off. Fig 5.13 shows the number of signal and background events associated to a given luminosity.

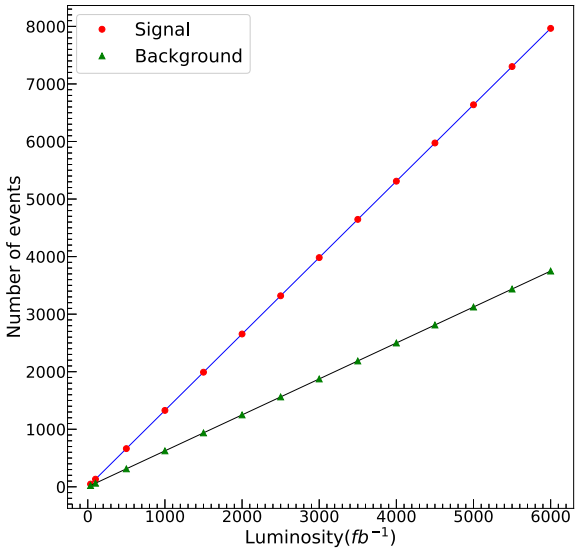


Figure 5.13: Expected number of events as function of luminosity. The slopes for the signal and background samples were evaluated using the event yields for luminosity of 35.9 fb^{-1} described in chapter 2.

Chapter 6

Conclusion

Effective field theories are theoretical extensions to the Standard Model in an attempt to understand the interactions between the particles in more detail. After the discovery of the Higgs boson, attempts were made to study its production and decay modes in more detail via EFT approaches. The EFTs developed are constrained using the experimental data which also became a benchmark for various BSM models.

In this thesis, we analysed the Higgs effective lagrangian focussing on the modifications in the decay width of Higgs decay to Z bosons via cWW EFT parameter. For that we developed a machine learning approach which can help in estimating the probability densities of the invariant masses of the Higgs boson and the intermediate Z bosons parametrised by the cWW parameter. Estimation of parameters was performed by a maximum likelihood estimation and a confidence interval measurement was performed via likelihood ratio tests for 68% confidence

. Interval measurements are important to ensure the true measurements to lie within the estimated interval. We saw from the invariant mass distributions and the likelihood scans that the model very well predicts the cWW parameter given the kinematic distributions with good precision.

Since the model was trained on pure signal samples, it showed deviations when tested on a mixed signal and background sample. So we retrained the model with background and observed an improvement in the predictions. However there are two things to be considered here:

- Based on the measurements for cWW performed as shown in [22, 23], the cWW parameter lies close to zero. So comparing the two models, both of them could be used to constrain cWW from the data in a precise manner.

- The main purpose of developing a methodology to have a trained model for mixed signal and background event is the uncertainty of how small the other parameters would be compared to zero. In that respect, it is better to have the analysis done considering background interference as well.

When trained with background, we see that the measurement lies within 1σ interval for 68% confidence for all cWW parameters. This was not the case when tested with a model having zero background knowledge. Based on the results, the methodology developed in this thesis for Conditional RealNVP model can be extended to include measurement of more than one parameter, for which the training of the model has to be performed with appropriate hyperparameters. One can also use this model to constrain some other

EFT parameters, for which required kinematic variables as features need to be carefully selected.

Chapter 7

Appendix

7.1 Technical details of the computational tools

7.1.1 Python packages for machine learning

The following packages were used to prepare the machine learning model (respective versions are in brackets) :

- Python (3.8.8)
- Tensorflow (2.4.1)
- Keras (2.4.3)
- Tensorflow probability (0.12.2)
- Numpy (1.20.1)
- Matplotlib (3.4.2)
- Scikit-learn (0.24.1)

7.1.2 Analysis of event samples

For analysing the full simulation samples and Delphes samples, the CERN analysis software ROOT (v 6.24/02) is used.

7.2 Kinematic distributions

7.2.1 Kinematics of EFT Higgs signal over SM ZZ events

Figures 7.1 and 7.2 shows the stacked distributions for the invariant masses $m_{4\ell}$ (left column), m_{Z1} (middle column) and m_{Z2} (right column) for EFT Higgs signal over SM ZZ background.

7.2.2 Forward flow model inference

Figures 7.3 and 7.4 shows the comparison of distributions of true data (blue) and model predictions (green) for the invariant masses $m_{4\ell}$ (left column), m_{Z1} (middle column) and m_{Z2} (right column) for EFT Higgs signal.

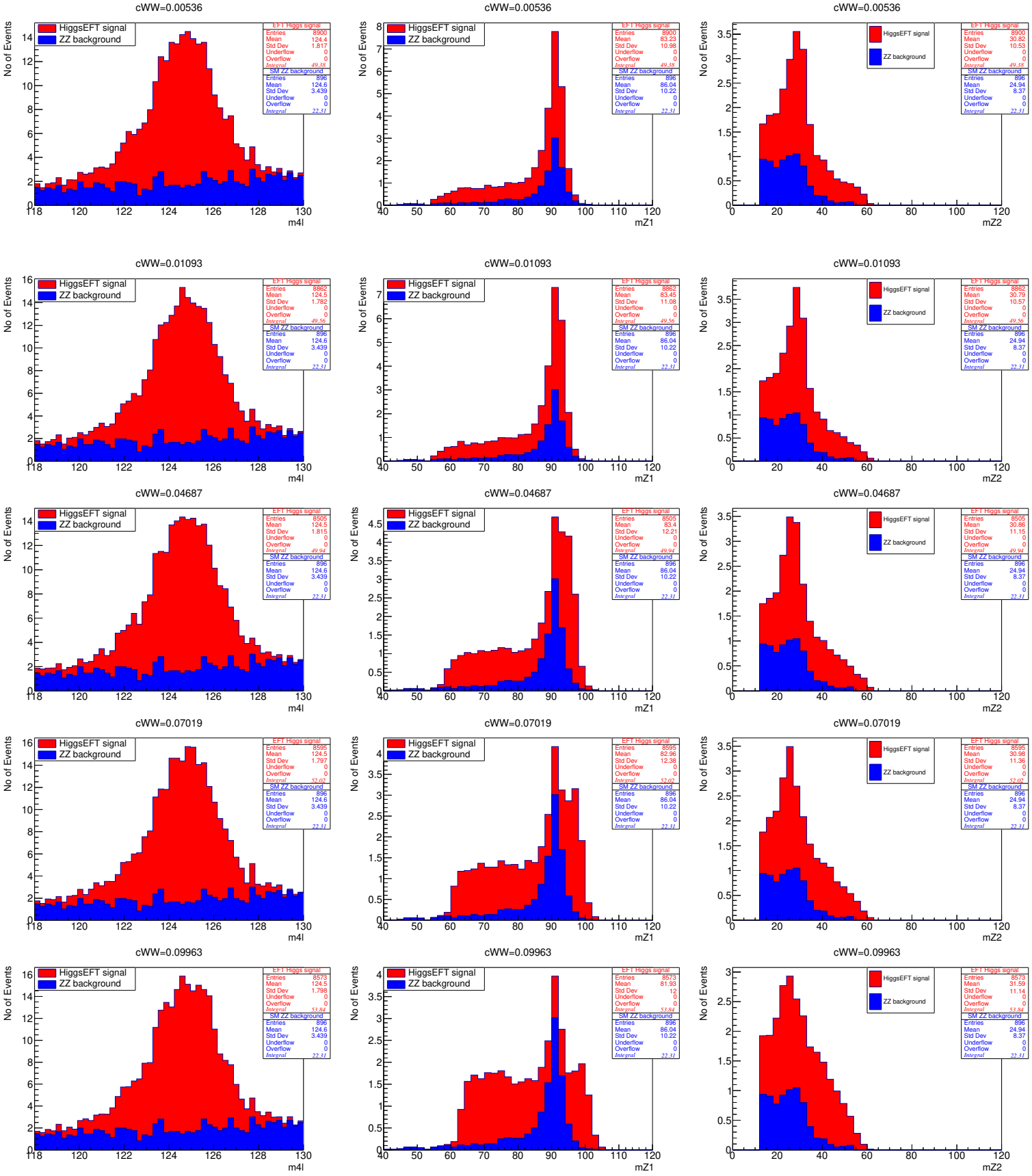


Figure 7.1: Kinematic distributions of invariant masses for EFT Higgs signal for cWW < 0.1 for luminosity of 35.9 fb^{-1}

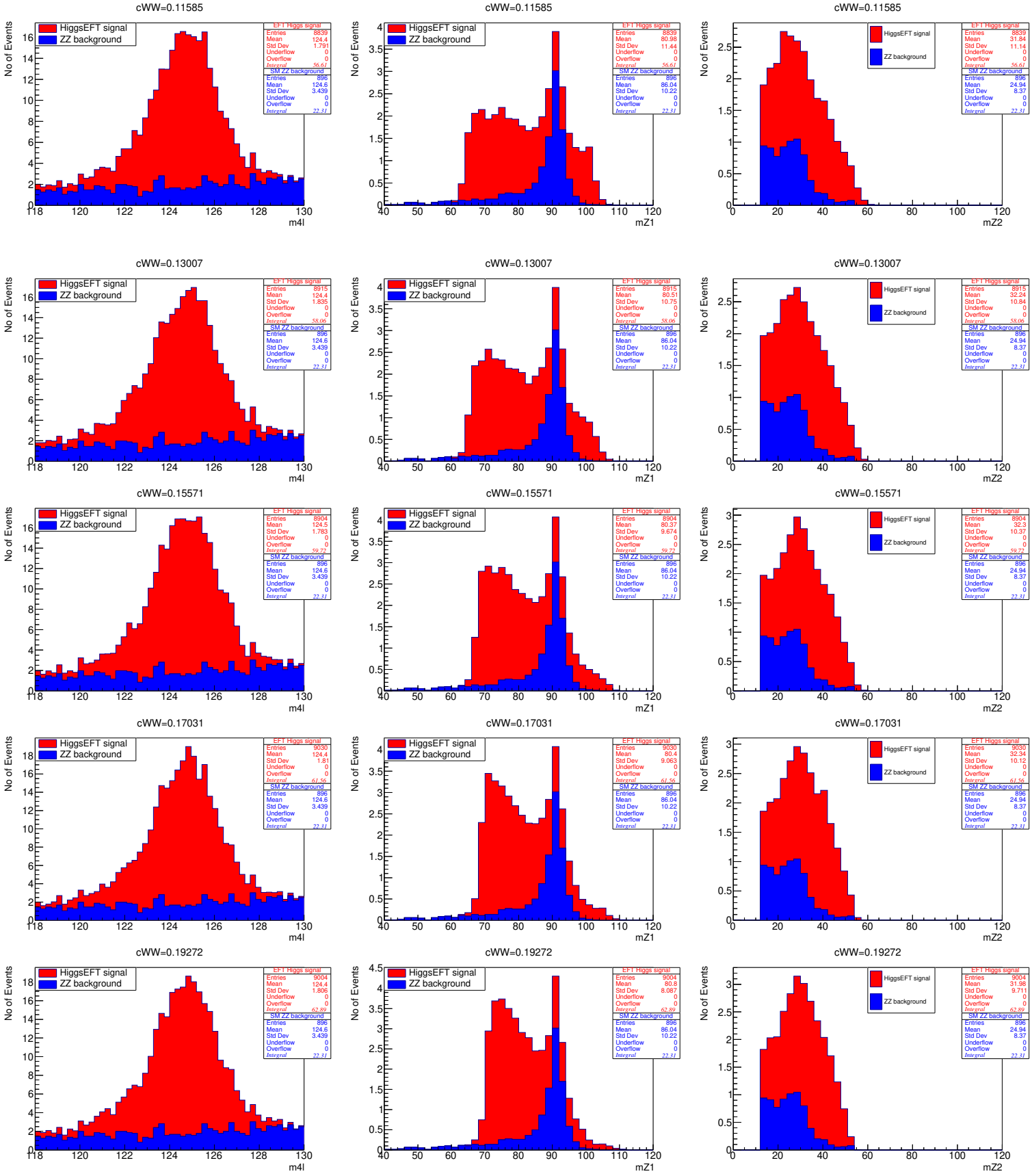


Figure 7.2: Kinematic distributions of invariant masses for EFT Higgs signal for cWW > 0.1 for luminosity of 35.9 fb^{-1}

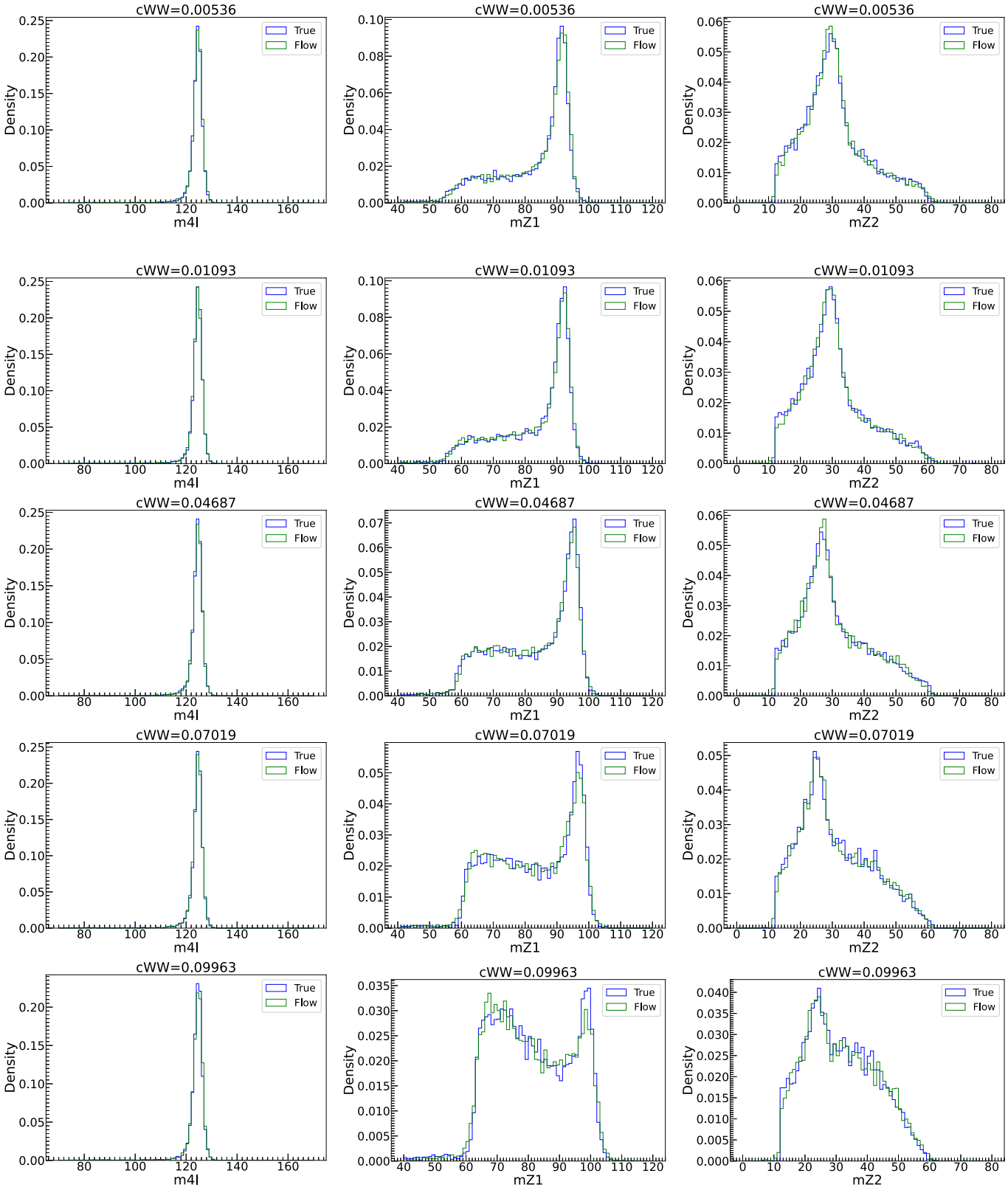


Figure 7.3: Kinematic distributions for the invariant masses for true data (blue) vs the one predicted by the model (green) for cWW < 0.1

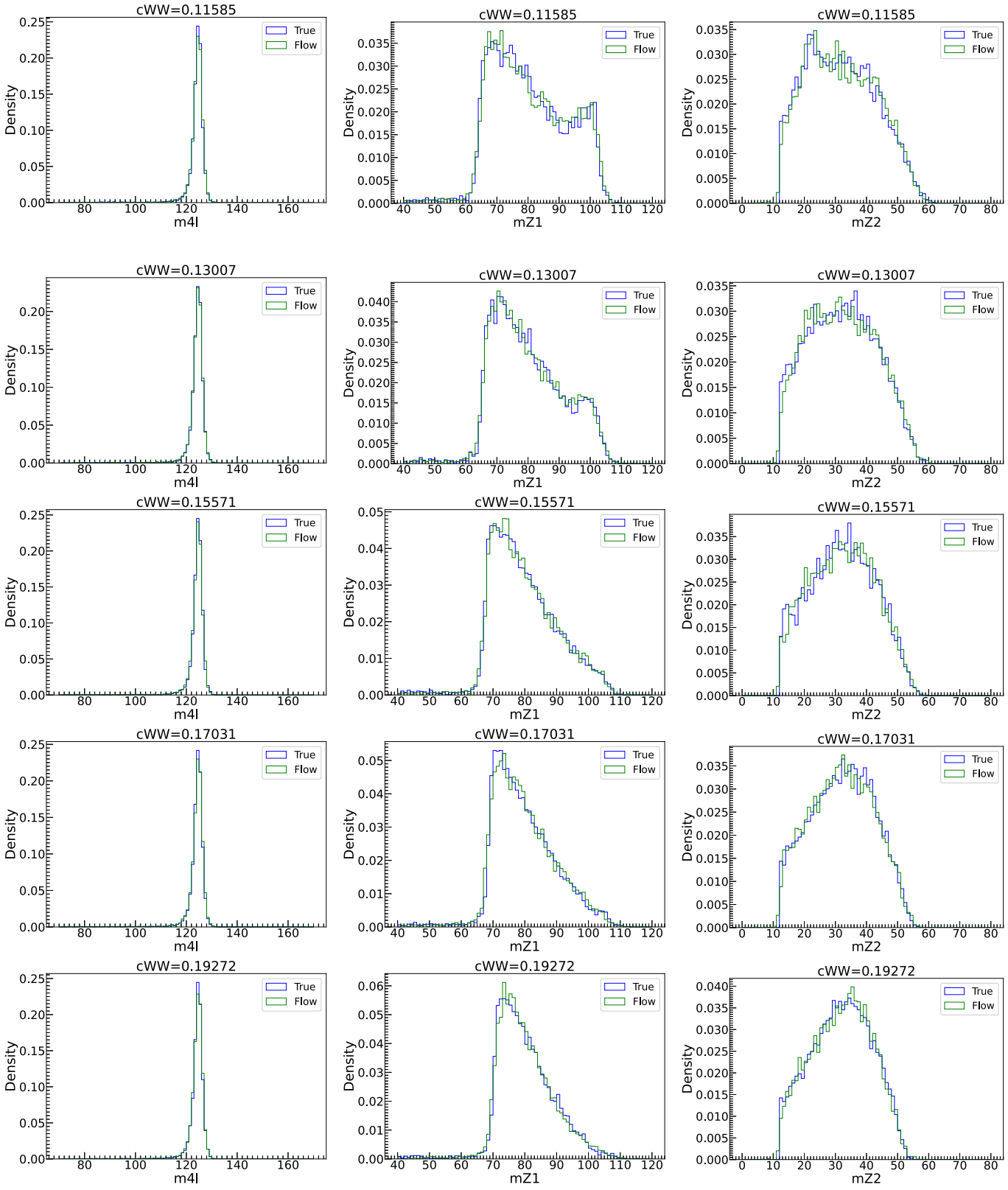
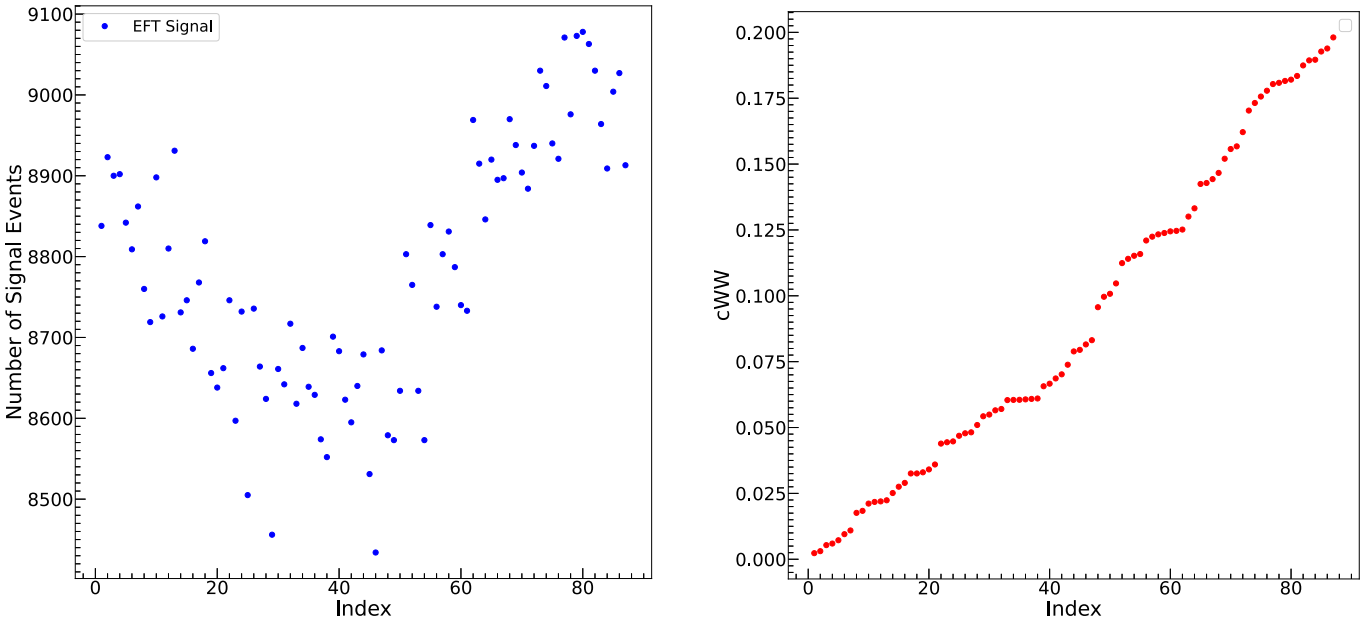


Figure 7.4: Kinematic distributions for the invariant masses for true data (blue) vs the one predicted by the model (green) for cWW > 0.1

7.3 Sample statistics

Fig 7.5 shows the number of events in the EFT signal used for training the model as described in section 3.4.4 and model inference as described in sections 3.5 and 5.1, showing the sample size of cWW datasets. Table 7.1 shows the cWW values whose datasets were used for training (green) and validation (blue).



(a) **(b)**

Figure 7.5: Number of events for the model inference using full EFT sample

0.00231	0.00307	0.00536	0.00596	0.00721	0.00952	0.01093	0.01761
0.01837	0.02113	0.02176	0.02196	0.02238	0.02514	0.02753	0.02898
0.03256	0.03257	0.033	0.03409	0.03597	0.0439	0.04441	0.04473
0.04687	0.04781	0.04818	0.05096	0.05427	0.05492	0.05653	0.05704
0.06041	0.06048	0.06052	0.06068	0.06086	0.06102	0.06566	0.06664
0.06863	0.07019	0.07384	0.07889	0.07948	0.08154	0.08315	0.09567
0.09963	0.10076	0.1047	0.11241	0.11404	0.11519	0.11585	0.12099
0.12244	0.12332	0.12384	0.12446	0.12465	0.12514	0.13007	0.1332
0.14245	0.14282	0.14429	0.14665	0.15203	0.15571	0.15672	0.16213
0.17031	0.17319	0.17563	0.17784	0.18037	0.18084	0.18156	0.18205
0.18345	0.18744	0.18936	0.1896	0.19272	0.19387	0.19807	

Table 7.1: The non-zero cWW values considered for the EFT samples.

Bibliography

[1] S. Novaes, "Standard Model: An Introduction", **2000**, DOI https://doi.org/10.48550/arXiv.hep-ph/0001283.

[2] H. Haber, Properties of the Gell-Mann matrices, http://scipp.ucsc.edu/~haber/ph251/gellmann17.pdf.

[3] D. Tong, Lectures on Quantum Field Theory, https://www.damtp.cam.ac.uk/user/tong/qft/four.pdf.

[4] F.Englert, R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons", *Phys. Rev. Lett. 13 321* **1964**, DOI http://dx.doi.org/10.1103/PhysRevLett.13.321.

[5] P.W.Higgs, "Broken Symmetries and the Masses of Gauge Bosons", *Phys. Rev. Lett. 13 508* **1964**, DOI http://dx.doi.org/10.1103/PhysRevLett.13.508.

[6] G.Guralnik, C.Hagen, T.Kibble, "Global Conservation Laws and Massless Particles", *Phys. Rev. Lett. 13 585* **1964**, DOI http://dx.doi.org/10.1103/PhysRevLett.13.585.

[7] https://cds.cern.ch/record/2012465/plots.

[8] I. van Vulpen, The Standard Model Higgs Boson, https://dokumen.tips/documents/the-standard-model-higgs-boson-nikhef-ivovhiggslecturenotepdfthe-standard-model.html?page=1.

[9] S. Regnard, Measurements of Higgs boson properties in the four-lepton final state at $\sqrt{s}$ = 13 TeV with the CMS experiment at the LHC. https://inspirehep.net/files/dd62de3eda9ff12f31020a6f11ae09a7.

[10] K. Heeger, "Evidence for Neutrino Mass : A Decade of Discovery", *nternational Conference on the Seesaw Mechanism and Neutrino Mass Paris France 10-11 June 2004*, DOI https://doi.org/10.1142/9789812702210_0005.

[11] A. Djouadi, "The Anatomy of Electro-Weak Symmetry Breaking. I: The Higgs boson in the Standard Model", *Phys.Rept.457:1-2162008* **2008**.

[12] C Grojean, "Higgs Physics", *Proceedings of the 2015 CERN-Latin-American School of High-Energy Physics Ibarra Ecuador 4 - 17 March 2015*, DOI https://doi.org/10.5170/CERN-2016-005.143.

[13] https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHWG#SM_Higgs.

[14] LHC Higgs Cross Section Working Group, "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector", *CERN Yellow Reports: Monographs Volume 2/2017 (CERN–2017–002-M)* **2017**, DOI https://doi.org/10.23731/CYRM-2017-002.

[15] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", *Phys. Lett. B 716 (2012) 30* **2012**, DOI https://doi.org/10.48550/arXiv.1207.7235.

[16] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Phys.Lett. B716 (2012) 1-29* **2012**, DOI https://doi.org/10.1016/j.physletb.2012.08.020.

[17] CMS collaboration, "Study of the mass and spin-parity of the Higgs boson candidate via its decays to Z boson pairs", *Phys. Rev. Lett. 110 (2013) 081803* **2013**, DOI https://doi.org/10.1103/PhysRevLett.110.081803.

[18] https://twiki.cern.ch/twiki/bin/view/AtlasPublic.

[19] https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResults.

[20] B. Grzadkowski, M. Iskrzynski, M. Misiak, J. Rosiek, "Dimension-Six Terms in the Standard Model Lagrangian", *J. High Energ. Phys. 2010 85 (2010)* **2010**, DOI https://doi.org/10.1007/JHEP10(2010)085.

[21] B. Henning, X. Lu, T. Melia, H. Murayama, "2, 84, 30, 993, 560, 15456, 11962, 261485, ...: Higher dimension operators in the SM EFT", **2019**, DOI https://doi.org/10.48550/arXiv.1512.03433.

[22] A. Alloul, B. Fuks, V. Sanz, "Phenomenology of the Higgs Effective Lagrangian via FeynRules", *JHEP 1404 (2014) 110* **2014**, DOI https://doi.org/10.48550/arXiv.1310.5150.

[23] C. Hays, V. Sanz, G. Zemaityte, Constraining EFT parameters using simplified template cross sections, https://cds.cern.ch/record/2673969?ln=en.

[24] M. E. Peskin, T. Takeuchi, "Estimation of oblique electroweak corrections", *Phys. Rev. D 46 381* **1992**, DOI https://doi.org/10.1103/PhysRevD.46.381.

[25] The CMS Collaboration, "Measurement of the properties of a Higgs boson in the four-lepton final state", *Phys. Rev. D 89 092007* **2014**, DOI https://doi.org/10.1103/PhysRevD.89.092007.

[26] The CMS Collaboration, "Measurements of properties of the Higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s}$ = 13 TeV", *Journal of High Energy Physics* **2017**, DOI https://doi.org/10.1007/JHEP11(2017)047.

[27] CMS Collaboration, CMS-AN-2016-448-v2 internal note,Measurements of properties of the Higgs boson in the four-lepton final state at $\sqrt{s}$ = 13 TeV, https://cms.cern.ch/iCMS/jsp/db_notes/noteInfo.jsp?cmsnoteid=CMS\%20AN-2016/442.

[28] J. Brehmer, K. Cranmer, G. Louppe, J. Pavez, "A Guide to Constraining Effective Field Theories with Machine Learning", *Phys. Rev. D 98 052004* **2018**, DOI https://doi.org/10.48550/arXiv.1805.00020.

[29] I. Kobyzev, S. J. Prince, M. A. Brubaker, "Normalizing Flows: An Introduction and Review of Current Methods", *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2020**, DOI https://doi.org/10.48550/arXiv.1908.09257.

[30] https://cms.cern/detector.

[31] The CMS Collaboration, "The CMS experiment at CERN", *JINST 3 S08004* **2008**, DOI https://iopscience.iop.org/article/10.1088/1748-0221/3/08/S08004/pdf.

[32] S. Agostinelli et al., "GEANT4–a simulation toolkit", *Nucl. Instrum. Meth. A* **2003**, DOI 10.1016/S0168-9002(03)01368-8.

[33] V. L. S. Ovyn, X. Rouby, "Delphes, a framework for fast simulation of a generic collider experiment", *Computer Physics Communications* **2010**, DOI https://doi.org/10.48550/arXiv.0903.2225.

[34] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, M. Selvaggi, "DELPHES 3 A modular framework for fast simulation of a generic collider experiment", *J. High Energ. Phys. 2014 57 (2014)* **2014**, DOI https://doi.org/10.1007/JHEP02(2014)057.

[35] C. Oleari, "The POWHEG-BOX", *Nucl.Phys.Proc.Suppl.205-206:36-412010* **2010**, DOI https://doi.org/10.1016/j.nuclphysbps.2010.08.016.

[36] Y. Gao, A. V. Gritsan, Z. Guo, K. Melnikov, M. Schulze, N. V. Tran, "Spin determination of single-produced resonances at hadron colliders", *Phys. Rev. D 81 075022* **2010**, DOI https://doi.org/10.1103/PhysRevD.81.075022.

[37] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, P. Z. Skands, "An Introduction to PYTHIA 8.2", *CERN-PH-TH-2014-190* **2014**.

[38] NNPDF Collaboration, "Recent progress on NNPDF for LHC", **2008**, DOI https://doi.org/10.48550/arXiv.0805.3100.

[39] C. Degrande, C. Duhr, B. Fuks, D. Grellscheid, O. Mattelaer, T. Reiter, "UFO - The Universal FeynRules Output", *Computer Physics Communications Volume 183 Issue 6* **2012**, DOI https://doi.org/10.1016/j.cpc.2012.01.022.

[40] CMS Collaboration, "Particle-flow reconstruction and global event description with the CMS detector", *JINST 12 (2017) P10003* **2017**, DOI https://doi.org/10.1088/1748-0221/12/10/P10003.

[41] https://cds.cern.ch/record/290480/files/ppe-95-147.pdf.

[42] D. J. Griffiths, *Introduction to Elementary Particles*, Wiley publications.

[43] C. Degrande, B. Fuks, K. Mawatari, K. Mimasu, V. Sanz, "Electroweak Higgs boson production in the standard model effective field theory beyond leading order in QCD", *Eur.Phys.J. C77 (2017) no.4 262* **2017**, DOI https://doi.org/10.48550/arXiv.1609.04833.

[44] A. Roy, An Introduction to Gradient Descent and Backpropagation, https://towardsdatascience.com/an-introduction-to-gradient-descent-and-backpropagation-81648bdb19b2.

[45] N. Mcculum, Deep Learning Neural Networks Explained in Plain English, https://www.freecodecamp.org/news/deep-learning-neural-networks-explained-in-plain-english/.

[46] L. Weng, Flow-based Deep Generative Models, https://lilianweng.github.io/posts/2018-10-13-flow-models.

[47] L. Dinh, J. Sohl-Dickstein, S. Bengio, "Density Estimation using RealNVP", *International Conference of Learning* **2017**, DOI https://doi.org/10.48550/arXiv.1605.08803.

[48] http://buzzard.ups.edu/courses/2007spring/projects/million-paper.pdf.

[49] J. Brownlee, Train-Test Split for Evaluating Machine Learning Algorithms, https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/.

[50] Andrew NG, Machine Learning Lectures of StanFord University, https://www.youtube.com/watch?v=jGwO_UgTS7I&list=PLoROMvodv4rMiGQp3WXShtMGgzqpfVfbU.

[51] Y. Verma, A Complete Understanding of Dense Layers in Neural Networks, https://analyticsindiamag.com/a-complete-understanding-of-dense-layers-in-neural-networks/.

[52] J. Brownlee, A Gentle Introduction to the Rectified Linear Unit(ReLU), https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/.

[53] S. Patrikar, Batch, Mini Batch & Stochastic Gradient Descent, https://towardsdatascience.com/batch-mini-batch-stochastic-gradient-descent-7a62ecba642a.

[54] J. Brownlee, A Gentle Introduction to Mini-Batch Gradient Descent and How to Configure Batch Size, https://machinelearningmastery.com/gentle-introduction-mini-batch-gradient-descent-configure-batch-size/.

[55] S. Kullback, R. A. Leibler, "On Information and Sufficiency", *The Annals of Mathematical Statistics Mar. 1951 Vol. 22 No. 1 (Mar. 1951) pp. 79-86* **1951**, DOI https://www.jstor.org/stable/2236703.

[56] Kullback-Leibler Divergence Explained, https://www.countbayesie.com/blog/2017/5/9/kullback-leibler-divergence-explained.

[57] Particle Data Group, PDG Review Statistics, **2019**, https://pdg.lbl.gov/2020/reviews/rpp2020-rev-statistics.pdf.

[58] I. M. John E Freund, M. Miller, *Mathematical Statistics with applications*, Pearson.

[59] https://web.stat.tamu.edu/~debdeep/LRTests.pdf.

[60] J. A. Rice, *Mathematical Statistics and Data Analysis*.

[61] D. Pati, Likelihood Ratio tests, http://www.acclab.helsinki.fi/~knordlun/mc/mc5nc.pdf.