DEPARTMENT OF PHYSICS & ASTRONOMY

UCL

2025 PHAS0052 Individual Report Cover Sheet

Identifier:	BB
Group:	20
Board Member:	Prof. Matthew Wing
Project Title:	Rediscovering the Higgs boson at the CMS experiment
Word Count:	1628

***Abstract* — In the project *Rediscovering Higgs Bosons*, I was responsible for analyzing the 2011, 2012, and 2015 $H \rightarrow ZZ^{(*)} \rightarrow 4l$ simulated data and a portion of the 2011 Double Electrons collision data (1.8 TiB out of 3.6 TiB). After deriving the four-momentum of each particle, the invariant mass of the final state in each event was calculated and further used to draw the mass histogram. Gaussian fitting was applied to simulated data, while Crystal Ball fitting was used for real data to determine the reconstructed Higgs boson mass. For 2011, M_H was calculated to be (124.95 $\pm$ 0.12) GeV, (123.76 $\pm$ 0.16) GeV, and (124.58 $\pm$ 0.18) GeV for channels $H \rightarrow ZZ \rightarrow 4\mu$, $H \rightarrow ZZ \rightarrow 4e$, and $H \rightarrow ZZ \rightarrow 2e2\mu$, respectively. For 2012, M_H was calculated to be (125.75 $\pm$ 0.60) GeV, (125.29 $\pm$ 0.28) GeV, and (126.74 $\pm$ 0.71) GeV for channels $H \rightarrow ZZ \rightarrow 4\mu$, $H \rightarrow ZZ \rightarrow 4e$, and $H \rightarrow ZZ \rightarrow 2e2\mu$, respectively. For 2015, M_H was calculated to be (124.99 ± 0.27) GeV. For 2011 real collision data, M_H was calculated to be (120.00 ± 0.66) GeV.**

I. INTRODUCTION

The discovery of the standard model (SM) Higgs boson is one of the primary scientific goals of the Large Hadron Collider (LHC). It was experimentally confirmed in 2012 by the ATLAS and CMS collaborations.

At the LHC, Higgs bosons are predominantly produced by proton-proton (pp) collisions, with the most common mechanism being gluon-gluon fusion (ggF), as shown in Fig.1. In this process, two gluons from the colliding protons interact via a virtual top-quark loop to produce Higgs boson.

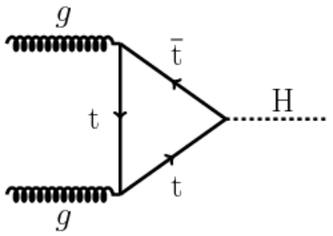


Figure.1. Feynman diagram of gluon-gluon fusion

When a Higgs boson is produced, it decays immediately into other particles. The decay $H \rightarrow ZZ^{(*)} \rightarrow 4l$ provides the cleanest signature.

The Higgs boson first decays into a pair of Z bosons, one of which is on-shell (real) and the other is off-shell (virtual). Then each Z boson decays into a pair of leptons (e^+e^- or $\mu^+\mu^-$) through Drell-Yan process, leading to the final state $ZZ^{(*)} \rightarrow 2l^+2l^-$ [13]. This results in a four-lepton final state.

$H \rightarrow ZZ^{(*)} \rightarrow 4l$ decay channel has several advantages. Since the final state is exclusively consist of electrons and muons, the resolution of their reconstructed parameters is high due to their efficient identification [12].

This report summarizes my personal contribution to the group project. I have focused on the $H \rightarrow ZZ^{(*)} \rightarrow 4l$ decay and analyzed the simulated data for year 2011, 2012, and 2015, along with real collision data from 2011. Also, I worked on building and creating containers in docker and programming to fit the mass spectrums. Therefore, the accurate data interpretation and computational efficiency could be further achieved by the whole group.

II. METHOD

A docker image with required python libraries, numpy, pandas, matplotlib, etc, was pulled, and the container 'cms_analysis' was created.

Then, the libraries-uproot and awkward-were installed through the terminal.

After the installation, the container was running and the jupyter notebook was started.

The methods to reconstruct mass of Higgs Bosons' candidates differ slightly between datasets due to the variations in naming conventions.

II.A Simulated 2011-2012 Higgs-to-four-leptons data

The simulated dataset ***Simulated dataset SMHiggsToZZTo4L_M-125_7TeV powheg15-JHUgenV3-pythia6*** (299683 events with 54.2GiB) [3] was chosen and the root files were opened via the uproot library.

The branches 'recoMuons_muons__RECO.obj.pt_', 'recoMuons_muons__RECO.obj.eta_', 'recoMuons_muons__RECO.obj.phi_', 'recoMuons_muons__RECO.obj.mass_' were selected, representing muons' transverse momentum p_T, pseudorapidity η, azimuthal angle ϕ, and mass m, respectively.

The events with numbers of muons less than 4 were filtered out, and 49629 events left.

The four-momentum $\boldsymbol{p^\mu} = \left(E, p_x, p_y, p_z\right)$ for each particle was derived from the relativistic mechanics:

$$p_x = p_T cos(\phi) \quad (1)$$

$$p_y = p_T sin(\phi) \quad (2)$$

$$p_z = p_T sinh(\eta) \quad (3)$$

$$E = \sqrt{{p_x}^2 + {p_x}^2 + {p_x}^2 + m^2} \quad (4)$$

The invariant mass $M_{4\mu}$ of the final states in the $H \rightarrow ZZ \rightarrow 4\mu$ decay is given by the relativistic energy-momentum relation:

$$M = (\sum E_i)^2 - |\sum \boldsymbol{p_i}|^2 \qquad (5)$$

where E_i is the energy of each particle, $\boldsymbol{p_i} = (p_x, p_y, p_z)$ is the three momentums of each particle.

The branches 'recoGsfElectrons_gsfElectrons__RECO.obj.pt_', 'recoGsfElectrons_gsfElectrons__RECO.obj.eta_', 'recoGsfElectrons_gsfElectrons__RECO.obj.phi_', 'recoGsfElectrons_gsfElectrons__RECO.obj.mass_' were then selected.

The events with numbers of electrons less than 4 were filtered out, and 24729 events left.

The invariant mass M_{4e} of the final states in the $H \rightarrow ZZ \rightarrow 4e$ decay was derived through Eq.1 to Eq.5.

Then the events with numbers of electrons and muons both greater than two were selected, with 55808 events left, to calculate the invariant mass $M_{2e2\mu}$ of the final states in the $H \rightarrow ZZ \rightarrow 2e2\mu$ decay.

After deriving the invariant masses $M_{4\mu}$, M_{4e}, and $M_{2e2\mu}$ for the Higgs decay subchannels, the relevant background datasets [6][7][8] were chosen.

The same event selection and calculation mechanisms were implemented, and the invariant masses $M_{4\mu}^{ZZ}$, M_{4e}^{ZZ}, and $M_{2e2\mu}^{ZZ}$ were derived for the background mass spectrums $ZZ \rightarrow 4e$, $ZZ \rightarrow 4\mu$, and $ZZ \rightarrow 2e2\mu$.

The above procedures were repeated for the 2012 ***Simulated dataset SMHiggsToZZTo4L_M-125_8TeV-powheg15-JHUgenV3-pythia6 in AODSIM format for 2012 collision data*** (299973 events with 98.9GiB) [4] and background datasets[9][10][11], with remaining events 85445, 30952, 75622 for channels $H \rightarrow ZZ \rightarrow 4\mu$, $H \rightarrow ZZ \rightarrow 4e$, and $H \rightarrow ZZ \rightarrow 2e2\mu$ respectively.

II.B Simulated 2015 Higgs-to-four-leptons data

The dataset ***Simulated dataset ttH_HToZZ_4LFilter_M125_13TeV_powheg2_JHUgenV6_pythia8 in MINIAODSIM format for 2015 collision data*** (510647 events with 20.4GiB) [5] was chosen.

The branches 'LHEEventProduct_externalLHEProducer__LHE.obj.hepeup_.IDUP' and 'LHEEventProduct_externalLHEProducer__LHE.obj.hepeup_.PUP.x[5]' were selected, where IDUP represents Particle Data Group Identifier (PDG ID), see Appendix A of PDGID for common particles, and PUP represents the four-momentum and mass (p_x, p_y, p_z, E, m) of each particles.

Electrons and muons were identified by selecting IDUP values with ± 11 and ± 13.

After filtering out the events with lepton numbers less than 4, 152346 events remained, the invariant mass M_{4l} were calculated through Eq.5.

II.C 2011 Double Electron pp collision data

The dataset ***DoubleElectron primary dataset in AOD format from RunA of 2011*** (49693737 events with 6.1TiB) [2] was chosen.

The same branches for electrons in II.A were selected.

The electrons with $p_T < 7$ and $|\eta| > 2.5$ were filtered out [1], and the events with number of electrons less than 2 were filtered out.

The invariant mass M_{4l} were calculated through Eq.1 and Eq.2.

III. Results & Data Analysis

In the analysis of simulated 2011, 2012, and 2015 Higgs-to-four-leptons data, a Gaussian distribution was employed to fit the $M_{4\mu}$, M_{4e}, $M_{2e2\mu}$ within the mass range of 80GeV to 150 GeV. The Gaussian function used is given by:

$$\boldsymbol{f(x) = Ae^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}} \qquad (3)$$

where A is the normalization factor, μ is the mean of the distribution, σ is the standard deviation, and x corresponds to invariant mass of the four-lepton system.

To perform the Gaussian fit, the curve_fit function from the scipy.optimize library was utilized, the full code was provided in Appendix B.

The distribution of four-lepton invariant masses for the 2011 simulated data are shown in Fig.2, Fig.3, and Fig.4, and to make the distribution more realistic, the background from Standard Model ZZ^* process is represented by shaded histograms.

The red lines of the figures depict distinct peaks around 125GeV, align with the simulated mass value. The reconstructed masses of Higgs Boson were determined using Gaussian fitting, as summarized in Table.1.

The deviations of the measured masses, $M_{4\mu}$, M_{4e}, $M_{2e2\mu}$, from the expected value are quantified by 0.41 σ, 7.75 σ, and 2.33 σ, respectively.

Since the filled histograms were the invariant masses of the final states in $ZZ^* \rightarrow 4\mu$, $ZZ^* \rightarrow 4e$, and $ZZ^* \rightarrow 2e2\mu$ decays, the distinct peaks around 90GeV align with the mass of Z boson, 91GeV.

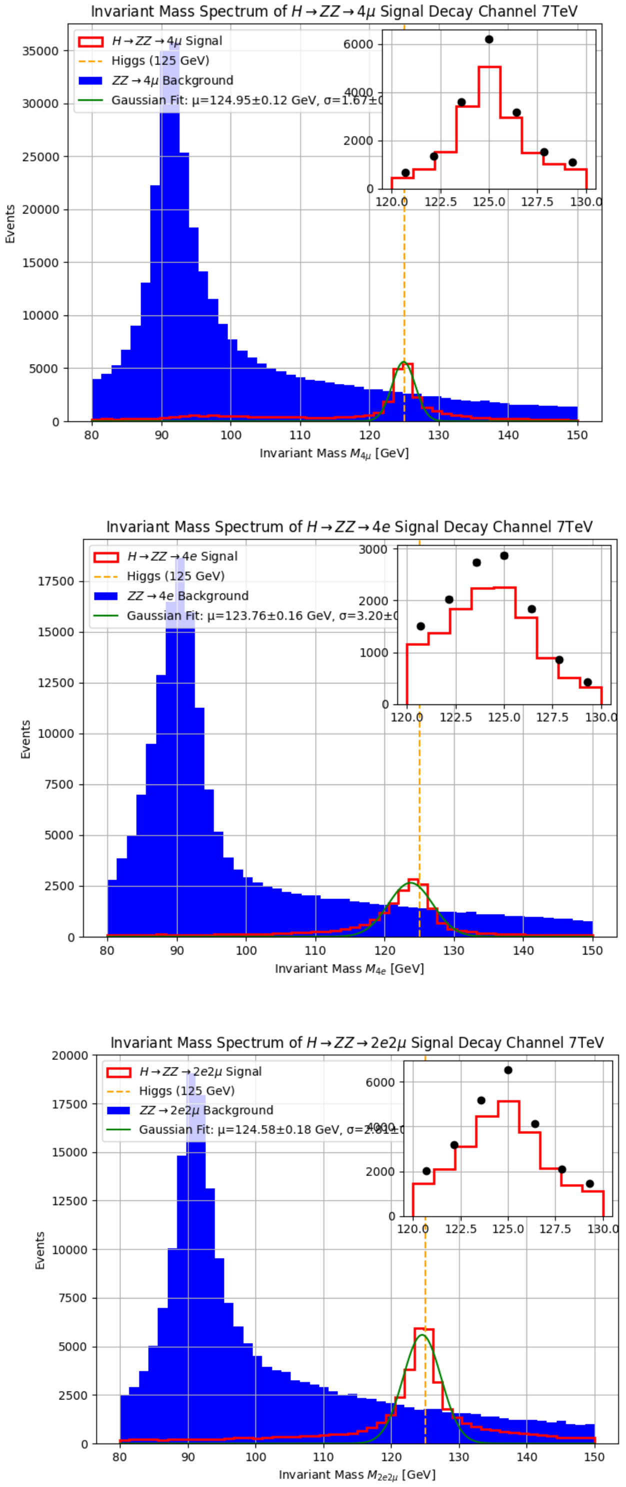


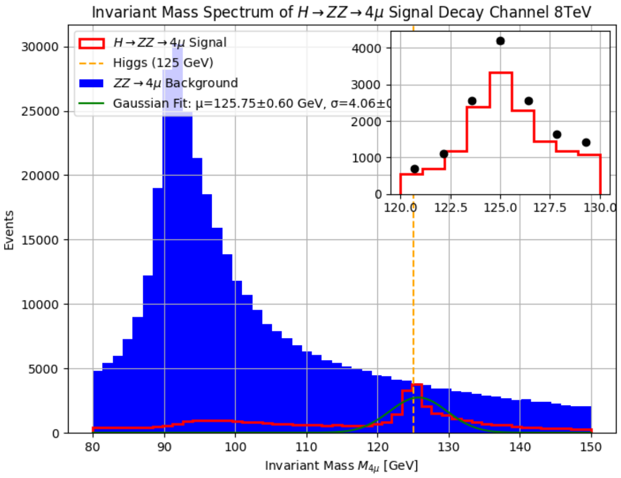
Figure.2, 3, and 4. The invariant masses, $M_{4\mu}, M_{4e}, M_{2e2\mu}$, distribution. The red lines depict the signals extracted from the 2011 Higgs Bosons decays; the filled histogram represents the corresponding backgrounds from 2011 SM ZZ^* process.

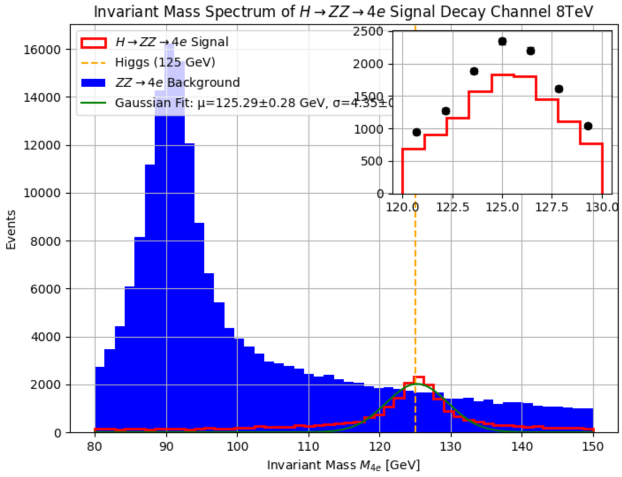
Channels	M_H (GeV)	Events
$H \rightarrow ZZ \rightarrow 4\mu$	124.95 ± 0.12	46923
$H \rightarrow ZZ \rightarrow 4e$	123.76 ± 0.16	24729
$H \rightarrow ZZ \rightarrow 2e2\mu$	124.58 ± 0.18	55808

Table 1. Reconstructed masses of Higgs Bosons for 2011 simulated dataset

The distribution of four-lepton invariant masses for the 2012 simulated data are shown in Fig.5, Fig.6, and Fig.7.

The deviations of the measured masses, $M_{4\mu}$, M_{4e}, $M_{2e2\mu}$, from the expected value are quantified by 1.25 σ, 1.04 σ, and 1.04 σ, respectively.





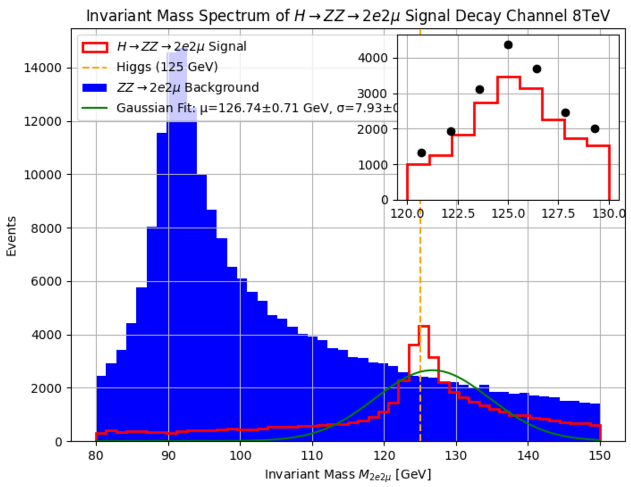


Figure.5, 6, and 7. The invariant masses, $M_{4\mu}, M_{4e}, M_{2e2\mu}$, distribution. The red lines depict the signals extracted from the 2012 Higgs Bosons decays; the filled histogram represents the corresponding backgrounds from 2012 SM ZZ process*

Channels	M_H (GeV)	Events
$H\ \to ZZ\ \to\ 4\mu$	125.75 ± 0.60	85448
$H\ \to ZZ\ \to\ 4e$	125.29 ± 0.28	30952
$H\ \to ZZ\ \to\ 2e2\mu$	126.74 ± 0.71	75622

Table 2. Reconstructed masses of Higgs bosons for 2012 simulated dataset

Since only simulated $H\ \to\ 4l$ decay dataset without background was found for 2015, the histogram in Fig.7 appears overly simplistic. The final reconstructed mass of Higgs boson in 2015 is $124.99\ \pm\ 0.27$ GeV, with the theoretical mass falling within 0.04 σ.

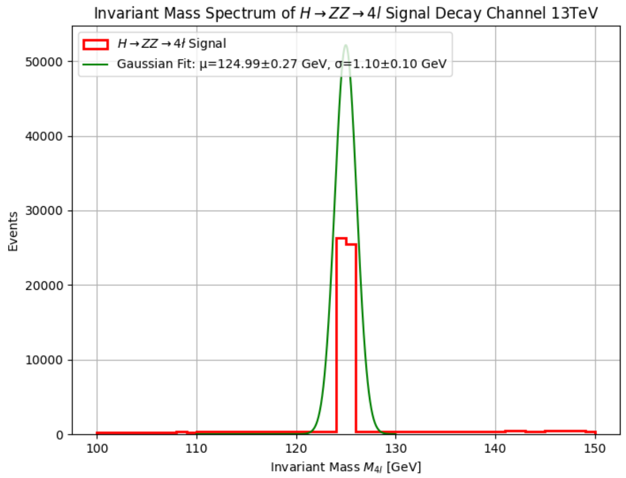


Figure.7. Invariant mass, M_{4l}, distribution. The red line depicts the signal extracted from the 2015 Higgs Bosons decay.

In the analysis of 2011 collision data, the crystal ball function, combining a Gaussian core and a power-tail, was used to model the reconstructed candidate events. The mathematical definition for the function is shown in Appendix C.

The invariant mass distribution for Double-Electron dataset is shown in Fig.8. The blue line uses polynomial function to fit the background, and the red line is the combination of background and signal fitting. There is a bump around 120 GeV, a slightly excess over the background.

Fig.9 depicts the invariant mass distribution where the background is subtracted, and now the red line simply depicts the signal.

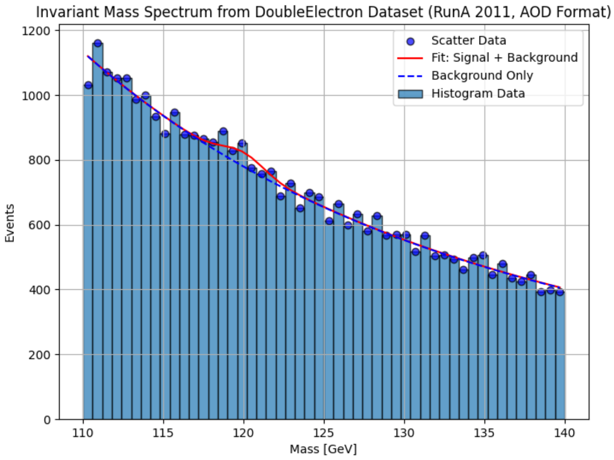


Figure.8. The shaded histogram represents invariant mass distribution of 2011 Double-Electron collision dataset. The blue line is the fitting background, and the red lines is the combination of signal and background. (This figure derived together with DD)

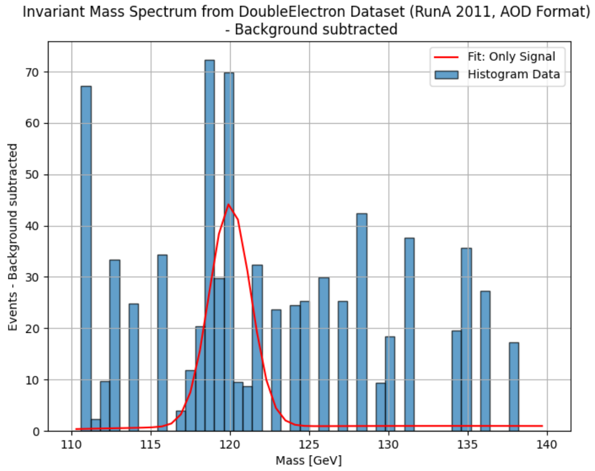


Figure.9. The same dataset with Fig.8. where the background is now subtracted. The red line depicts the fitting signal. (This figure derived together with DD)

IV. Conclusion

The reconstructed mass of the Higgs boson, derived from the 2011 and 2012 simulated data, exhibits strong alignment with the expected value of 125 GeV.

Also, both 2011 and 2012 simulated data manifests that the events for $H \rightarrow ZZ \rightarrow 4e$ decay is less than 50% of other two decay channels. This may be due to electrons interact strongly with materials in detector, leading to a reduction in electromagnetic calorimeter's efficiency in measuring their energies.

The reconstructed mass of the Higgs boson in 2011 collision data is (120.00 ± 0.66) GeV, which deviates from expected value by 7.58 σ. This might be due to several reasons.

Firstly, the dataset is limited. The dataset 2011 Double-Electron collision dataset contains events where at least two high-energy electrons in the events. Therefore, while it can capture the events with two Z bosons decaying into 4 electrons, the $H \rightarrow ZZ \rightarrow 4\mu$ and $H \rightarrow ZZ \rightarrow 2e2\mu$ channels are lost.

Secondly, the event selection requires further improvements. Due to the computational limitations of our laptops, only 3.6TiB of real data were analyzed in total. Therefore, it's difficult to balance between the event selection strictness and the amount of data remains. In the future, a Boosted Decision Tree (BDT) [1] could be applied and trained to identify Higgs Boson-related events by leveraging results and patterns derived from simulated data.

V. References

[1] Chatrchyan. S., et al. (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, [online] 716(1), pp.30–61. doi: https://doi.org/10.1016/j.physletb.2012.08.021.

[2]CMS collaboration (2025) */DoubleElectron/Run2011A-12Oct2013-v1/AOD*, *Cern.ch*. CERN Open Data Portal. Available at: https://opendata.cern.ch/record/16 (Accessed: 10 March 2025).

[3]CMS collaboration (2025). */SMHiggsToZZTo4L_M-125_7TeV-powheg15-JHUgenV3-pythia6/Summer11LegDR-PU_S13_START53_LV6-v1/AODSIM*. [online] Cern.ch. Available at: https://opendata.cern.ch/record/1507 (Accessed 10 Mar. 2025).

[4]CMS collaboration (2025). */SMHiggsToZZTo4L_M-125_8TeV-powheg15-JHUgenV3-pythia6/Summer12_DR53X-PU_S10_START53_V19-v1/AODSIM*. [online] Cern.ch. Available at: https://opendata.cern.ch/record/9356 (Accessed 10 Mar. 2025).

[5] CMS Collaboration (2025). */ttH_HToZZ_4LFilter_M125_13TeV_powheg2_JHUgenV6_pythia8/RunIIFall15MiniAODv2-PU25nsData2015v1_76X_mcRun2_asymptotic_v12-v1/MINIAODSIM*. [online] Cern.ch. Available at: https://opendata.cern.ch/record/22285 (Accessed 10 Mar. 2025).

[6]CMS collaboration (2025) */ZZTo4mu_mll4_7TeV-powheg-pythia6/Summer11LegDR-PU_S13_START53_LV6-v1/AODSIM*, *Cern.ch*. CERN Open Data Portal. Available at: https://opendata.cern.ch/record/1651 (Accessed: 10 March 2025).

[7] CMS collaboration (2025) */ZZTo4e_mll4_7TeV-powheg-pythia6/Summer11LegDR-PU_S13_START53_LV6-v1/AODSIM*, *Cern.ch*. CERN Open Data Portal. Available at: https://opendata.cern.ch/record/1648 (Accessed: 10 March 2025).

[8]CMS collaboration (2025) */ZZTo2e2mu_mll4_7TeV-powheg-pythia6/Summer11LegDR-PU_S13_START53_LV6-v1/AODSIM*, *Cern.ch*. CERN Open Data Portal. Available at: https://opendata.cern.ch/record/1382 (Accessed: 10 March 2025).

[9] CMS collaboration (2025) */ZZTo4mu_8TeV-powheg-pythia6/Summer12_DR53X-PU_RD1_START53_V7N-v1/AODSIM*, *Cern.ch*. CERN Open Data Portal. Available at: https://opendata.cern.ch/record/10071 (Accessed: 10 March 2025).

[10] CMS collaboration (2025) */ZZTo4e_8TeV-powheg-pythia6/Summer12_DR53X-PU_RD1_START53_V7N-v2/AODSIM*, *Cern.ch*. CERN Open Data Portal. Available at: https://opendata.cern.ch/record/10065 (Accessed: 10 March 2025).

[11] CMS collaboration (2025) */ZZTo2e2mu_8TeV-powheg-pythia6/Summer12_DR53X-PU_RD1_START53_V7N-v2/AODSIM*, *Cern.ch*. CERN Open Data Portal. Available at: https://opendata.cern.ch/record/10054 (Accessed: 10 March 2025).

[12] Duarte, L. *et al.* (2023) 'Exclusive doubly charged Higgs boson pair production in pp collisions at the LHC', *Physical review. D/Physical review. D.*,

107(3). doi: https://doi.org/10.1103/physrevd.107.035010.

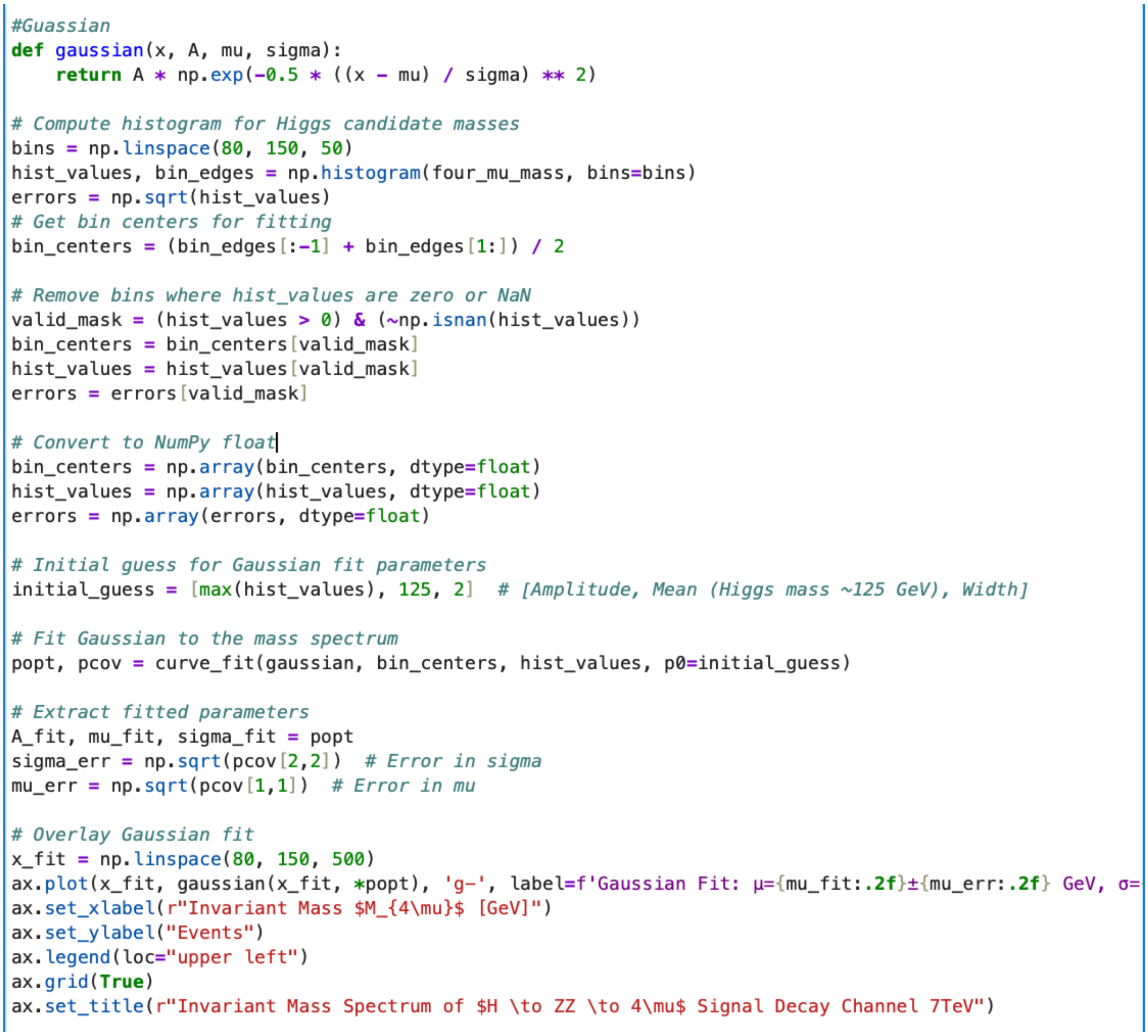
[13] Zagoskin, T.V. and Korchin, A.Y. (2017). The Higgs boson decay into *ZZ* decaying to identical fermion pairs. *International Journal of Modern Physics A*, [online] 32(27), pp.1750166–1750166. doi: https://doi.org/10.1142/s0217751x17501664.

VI. Appendix

A. Table of Some common PDGID

Quarks		Leptons		Gauge and Higgs bosons	
Symbol	PDGID	Symbol	PDGID	Symbol	PDGID
d	1	e^-	11	g	9 and 21
u	2	v_e	12	γ	22
s	3	μ^-	13	Z	23
c	4	v_μ	14	W^+	24
b	5	τ^-	15	h^0 or H_1^0	25
t	6	v_τ	16	H^+	37
b'	7	τ'^-	17		
t'	8	$v_{\tau'}$	18		

B. Full code for Guassian Fitting



```
#Guassian
def gaussian(x, A, mu, sigma):
    return A * np.exp(-0.5 * ((x - mu) / sigma) ** 2)

# Compute histogram for Higgs candidate masses
bins = np.linspace(80, 150, 50)
hist_values, bin_edges = np.histogram(four_mu_mass, bins=bins)
errors = np.sqrt(hist_values)
# Get bin centers for fitting
bin_centers = (bin_edges[:-1] + bin_edges[1:]) / 2

# Remove bins where hist_values are zero or NaN
valid_mask = (hist_values > 0) & (~np.isnan(hist_values))
bin_centers = bin_centers[valid_mask]
hist_values = hist_values[valid_mask]
errors = errors[valid_mask]

# Convert to NumPy float
bin_centers = np.array(bin_centers, dtype=float)
hist_values = np.array(hist_values, dtype=float)
errors = np.array(errors, dtype=float)

# Initial guess for Gaussian fit parameters
initial_guess = [max(hist_values), 125, 2]  # [Amplitude, Mean (Higgs mass ~125 GeV), Width]

# Fit Gaussian to the mass spectrum
popt, pcov = curve_fit(gaussian, bin_centers, hist_values, p0=initial_guess)

# Extract fitted parameters
A_fit, mu_fit, sigma_fit = popt
sigma_err = np.sqrt(pcov[2,2])  # Error in sigma
mu_err = np.sqrt(pcov[1,1])  # Error in mu

# Overlay Gaussian fit
x_fit = np.linspace(80, 150, 500)
ax.plot(x_fit, gaussian(x_fit, *popt), 'g-', label=f'Gaussian Fit: μ={mu_fit:.2f}±{mu_err:.2f} GeV, σ=
ax.set_xlabel(r"Invariant Mass $M_{4\mu}$ [GeV]")
ax.set_ylabel("Events")
ax.legend(loc="upper left")
ax.grid(True)
ax.set_title(r"Invariant Mass Spectrum of $H \to ZZ \to 4\mu$ Signal Decay Channel 7TeV")
```

C. Mathmetical definition of Crystall Ball function

The crystal ball function is mathematically defined as:

$$f(x) \quad = N \cdot \begin{cases} exp(-\frac{(x-\bar{x})^2}{2\sigma^2}), & \frac{x-\bar{x}}{\sigma} > -\alpha \\ A \cdot \left(B - \frac{x-\bar{x}}{\sigma}\right)^{-n}, & \frac{x-\bar{x}}{\sigma} \leq -\alpha \end{cases} \qquad (4)$$

$$A = \left(\frac{n}{|\alpha|}\right)^n \cdot exp\left(-\frac{|\alpha|^2}{2}\right) \qquad (5)$$

$$B = \frac{n}{|\alpha|} - |\alpha| \qquad (6)$$

D. Reflections on the project

From personal perspective, the project Rediscovering Higgs Bosons not only give me theoretical knowledge about particle physics, but also let me learn lots of computational skills like Linux, ROOT, Docker, etc.

Being responsible of data analysis in the group project, I provided reconstructed results for both simulated data and real data. Also, I had learned how to collaborate with members to work as a team.

However, due to limitations on computational ability of our laptops and slightly uneven distributions of tasks within the groups, the final result is not very comprehensive.

In the future, access to more powerful computing resources and a more efficient task allocation strategy would enable us to analyze larger datasets and implement a more sophisticated event selection mechanism, leading to more refined results.