Search for an exotic Higgs boson decay to two pseudoscalar bosons with a four photon final state

Candidacy Proposal

Sergi Castells

Advisors: Colin Jessop and Nancy Marinelli

December 13, 2023

1 Introduction

The Standard Model (SM) is our most accurate and complete model of particle physics, describing three of the fundamental interactions, but there are yet many questions left unanswered. The discovery of the Higgs boson in 2012 at a mass of 125 GeV was a major milestone in showing the effectiveness of the theory. Yet, the SM does not account for dark matter, neutrino masses, and more. The search for Beyond the Standard Model (BSM) physics is an important facet of high energy physics that is motivated, in part, by these open questions. One set of interesting BSM phenomena to explore is exotic Higgs boson decays, which can be achieved in theories with extended Higgs sectors.

Some models predict the existance of a light, CP-odd pseudoscalar boson, e.g., models with at least one extra Higgs doublet – that is, an $\mathrm{SU}(2)_\mathrm{L}$ doublet [1]. An interesting process in many of these models is the 125 GeV SM Higgs boson decay to two pseudoscalars. The tree level Feynman diagram for a Higgs boson decaying to two pseudoscalars with a four

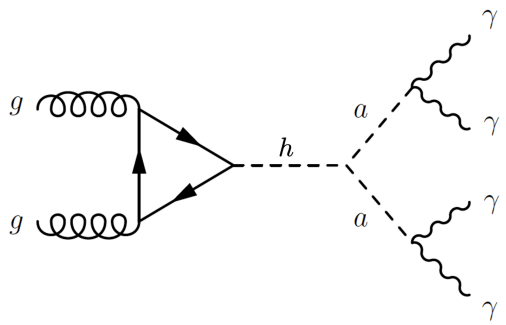

Figure 1: Tree-level Feynman diagram of $h \to aa \to \gamma\gamma\gamma\gamma$ for the gluon fusion Higgs production mode.

photon final state is shown in Fig 1. Given the current limits and exclusions, a large phase space still remains for these decays; thus, a search for this BSM process can be fairly model agnostic.

Depending on the mass of a, the decay $a \to \gamma\gamma$ may have three topologies. For $m_a > 15$ GeV, the photons are fully resolved, i.e., angular distance, $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$, is greater than 0.2, where η refers to pseudorapidity and ϕ refers to azimuthal coordinate in the detector. For $m_a < 15$ GeV, the opening angle of the photon pairs may be too small to resolve, yielding either a false two or three photon signal. In this analysis, we will be focusing on the fully resolved final state topology and only the gluon fusion Higgs production mode will be considered. These requirements, along with other exclusion limits to m_a, allow for a search for pseudoscalars with a mass range of $15 < m_a < 62$ GeV. A proposal for a study on the Higgs boson decay to pseudoscalars in the given mass range with the above final state topology will be put forth.

2 LHC and CMS

The Large Hadron Collider (LHC) is a proton-proton collider and the world's largest particle accelerator, located on the border of France and Switzerland. It is 27 kilometers in circumference and utilizes superconducting dipoles and quadrupoles, for bending and focusing the beams, respectively. The LHC currently operates at a center-of-mass energy of 13.6 TeV with a nominal instanteous luminosity of $10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ with a collision frequency of 40 MHz. There are two general purpose detectors at the LHC: CMS and ATLAS.

The Compact Muon Solenoid (CMS) detector is a hermetic, general purpose detector along the ring of the Large Hadron Collider (LHC). CMS is composed of several sub-detectors, each designed to collect data on a specific part of a proton-proton collision event, and a superconducting solenoid at a field strength of 3.8 T. A schematic of these sub-detectors is shown in Fig. 2. The main sub-detectors of CMS are, from the beam line outwards, the silicon tracker (pixel and strip), the electromagnetic calorimeter, the hadronic calorimeter, and the muon chambers [3].

Recording a collision event requires each of these subsystems to detect different parts of the event. Many particles are created during a collision with many decaying or hadronizing to produce even more particles. These particles first pass through the silicon tracker, which detects charged particles as tracks. Then the ECAL detects electromagnetically interacting particles with lead tungstate crystals. Photon and electron energy is completely measured, but charged hadrons, while they do produce a shower, pass through. The particles enter the HCAL, a sampling calorimeter, and produce showers. Charged and neutral hadrons are contained in the HCAL, but any remaining particles, i.e., muons, continue through the magnet to interact with the muon chambers.

2.1 ECAL

The ECAL is a homogeneous calorimeter and composed of $\sim$75,000 scintillating $PbWO_4$ crystals in two parts: the barrel (EB) and the endcaps (EE), in η ranges $|\eta| < 1.48$ and $1.48 < |\eta| < 3.00$, respectively. The $PbWO_4$ crystals measure the energy of electrons and

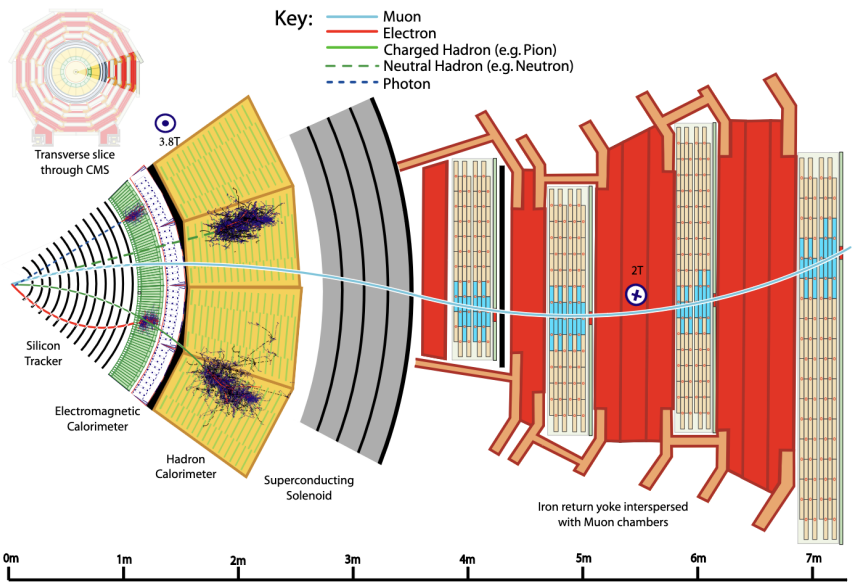

Figure 2: Slice showing CMS sub-detectors and how particles interact in them [2].

photons passing through them by collecting the energy of their electromagnetic showers [4]. The crystals were chosen for their Molière radius, i.e., a crystal contains 90% of a shower's deposition on average, and response time [5]. Unlike the barrel section, the endcaps have an additional preshower component in two layers positioned in front of the crystals: lead and silicon, where the silicon strip detectors are used to detect neutral pion decays.

3 Previous Studies

Previous studies of a Higgs boson decaying to two pseudoscalars, then each pseudoscalar decaying to two photons, have been performed [6]. The precursor to this proposed analysis is the most recent study of the same process, i.e., the CMS Run 2 analysis of $h \to aa \to \gamma\gamma\gamma\gamma$, where the photons are fully resolved and the pseudoscalar mass range is $15 < m_a < 62$ GeV. The aforementioned Run 2 analysis utilized $132\,\text{fb}^{-1}$ of data from the CMS detector and set the strictest limits on the process' production cross section to date. The limit on cross section times branching fraction from the Run 2 analysis, given by $\sigma(pp \to h) \times \text{Br}(h \to aa \to \gamma\gamma\gamma\gamma)$, was set at 0.80 fb for $m_a = 15$ GeV and 0.26 fb for $m_a = 62$ GeV [6]. For reference, the Higgs production cross section for all channels combined is 52 pb. Run 2 lasted from late 2015 through 2018, although the nominal dataset contains data from 2016–2018.

3.1 Advancements on Previous Studies

Advancements can be made relative to the Run 2 analysis in several ways. An increase in statistics of 50% could allow for this analysis to produce a discovery. However, in the case that we merely set limits, this increase in statistics would allow for even stricter limits on the production cross section for this process. The true increase in statistics is heavily dependent on both CMS and LHC schedules, but the availability of 2022 and 2023 data by

early-2024 suggests a match in statistics with Run 2 and increase in raw statistics of about 50% by the end of the 2024 data-taking period [7]. The previous analysis reported an impact of ∼1% from systematic uncertainties, meaning that an increase in statistics would be an almost direct improvement to analysis sensitivity [6]. However, it may yet be possible to reduce the ∼1% impact from systematic uncertainties since Run 3 data is expecting a gain of $\sqrt{2}$ times precision on Run 2.

3.2 New Framework: HiggsDNA

The new analysis framework HiggsDNA (Higgs to Diphoton NanoAOD Framework), produced by the $h \to \gamma\gamma$ group, is built on Coffea, a tool for processing nanoAOD samples [8]. HiggsDNA utilizes Coffea processors and uses a columnar-based approach rather than an event-based approach to accessing data; current columnar-based analysis is significantly faster than its event-based counterpart. A proof of concept is needed before applying this new HiggsDNA framework to a Run 3 analysis so a partial recreation of the Run 2 analysis is underway. The recreation is for 2018 only and there is an attempt to keep any necessary modifications to a minimum; all cuts, BDT input variables, and other Run 2 specific parameters are unchanged. This recreation should confirm the efficacy of the new HiggsDNA framework and allow for a smooth transition to new data.

4 Signal and Background

The signal samples are simulated $h \to aa \to \gamma\gamma\gamma\gamma$ Monte Carlo (MC) with a pseudoscalar mass range of $15 < m_a < 60$ GeV in steps of 5 GeV and a Higgs boson mass of 125 GeV. As stated previously, only the gluon fusion production mode of the Higgs boson is considered, and this is reflected in these MC samples. The 2018 signal samples were converted to nanoAOD with additional branches to enable the use of certain cuts. New signal MC will be needed for the Run 3 analysis and are expected to be generated in a substantially similar manner.

The Run 2 analysis has shown that a MC-based background was not viable since the MC available from the standard $h \to \gamma\gamma$ analysis suffered from low statistics, and thus very large event weights, after requiring the presence of four photons in an event. The main backgrounds for the $h \to aa \to \gamma\gamma\gamma\gamma$ final state are either well-isolated, prompt photons or isolated photons reconstructed due to very collimated decays fragmented from jets, i.e. fake photons. A data-driven background will be used in this analysis to avoid the large event weight problem. The data driven background utilizes a technique similar to hemisphere mixing to remove the presence of any signal and artificially create the background shape [9]. This technique is called event mixing and is a simplification of the hemisphere mixing procedure. Event mixing shuffles the photons in an event with the photons from consecutive events, modulo the number of events.

5 Event Selection

Events are chosen from data with the L1 trigger and the high level trigger (HLT) selecting interesting events with selections specifically optimized for the low mass diphoton search. Then selections are applied to photon candidates in these events to ensure that the candidates are indeed photons. When a sufficient set of selections have been imposed on photon candidates, additional selections designed to ensure optimal reconstruction of pseudoscalar candidates are imposed. Furthermore, a Boosted Decision Tree (BDT) is used to more effectively discern signal-like and background-like events. It is important to note that the triggers and selections will all be evaluated, and potentially changed, for the Run 3 analysis. All of the listed criteria for the 2018 proof-of-concept study are taken from the Run 2 analysis, where optimization of selections has already been performed.

Selections are applied to a photon candidates' energy deposition in the ECAL, isolation from other particle candidates, and to other particle candidate tracks to remove fake photons. Photon candidates are required to have a minimum transverse energy (E_T), concentration of energy in a 3×3 crystal matrix relative to its supercluster (R_9), energy weighted standard deviation of single crystal η within its supercluster ($\sigma_{i\eta i\eta}$), and ratio of energy deposited in nearest HCAL tower within a cone of $\Delta R = 0.15$ centered on photon candidate direction and energy deposited in its supercluster. A supercluster is defined as a 5×5 crystal matrix centered around the crystal of greatest energy deposition. These requirements ensure that a photon candidate has the correct shower shape, enough energy to be efficiently reconstructed, and is produced from a decay.

Photon candidates are also required to pass isolation requirements. To ensure that there is sufficiently little energy deposition in the area around the photon candidate, a maximum value of the following quantities is imposed: the sum of E_T of all Particle Flow photon candidates within a cone of $\Delta R = 0.3$ excluding the candidate photon, sum of of track p_T within a cone of $\Delta R = 0.3$, and sum of p_T of all Particle Flow charged hadron candidates within a cone of $\Delta R = 0.3$. These isolation requirements ensure that the photon candidate is not a very collimated decay fragmented from a jet, i.e., a fake photon. It is also important to ensure that a photon candidate's supercluster does not match with an electron candidate hit in the pixel tracker. This is known as an electron veto.

Fundamentally, the values of these cuts are determined by the physical properties of the detector. There are also selections that are motivated by the geometry of the detector, however. Selections also enforce that photon candidates lie within a range of η such that they are within the tracker fiducial region, i.e., the η region ecompassing the entire tracker, and not within the gap between the ECAL Barrel and Endcaps, where photon reconstruction is suboptimal. A selection that is entirely motivated by the search however is the mass window of the four photon candidate. This requirement is motivated by the decay width of the Higgs boson, which is ideally represented by the four photon object. At a minimum, a requirement of at least four photons in an event and at least one diphoton candidate is necesssary.

The pseudoscalars then need to be reconstructed from the four photons in an event. In the case of an event with more than four photons, the four with the highest p_T are chosen. To reconstruct the pseudoscalar candidate, a technique called mass mixing is used. All combinations of the pseudoscalar candidate pairs, which each consist of two photons, are computed. To determine which set of pseudoscalar candidates is the best match, the

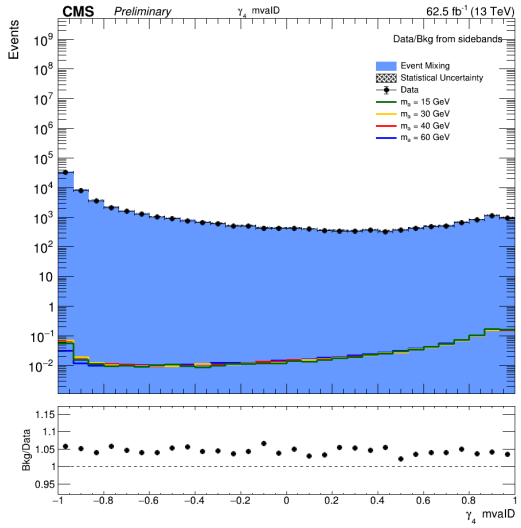

Figure 3: Example of MVA ID for γ_4, calculated by the $h \to \gamma\gamma$ group, is the strongest discriminating variable for the event selection BDT. Signal MC for four nominal pseudoscalar mass points, event mixed background, and data are shown with selections applied.

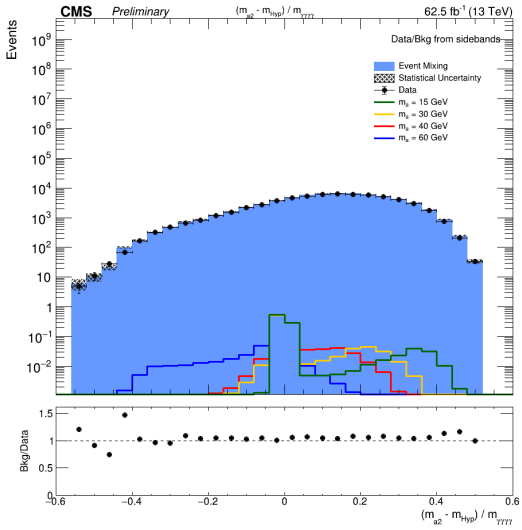

Figure 4: $(m_{a2\,RECO} - m_{a\,Hyp})/m_{\gamma\gamma\gamma\gamma}$ is the second strongest discriminating variable for the event selection BDT. Signal MC for four nominal pseudoscalar mass points, event mixed background, and data are shown with selections applied.

quantity $\Delta M = |m_{a_{\gamma_a\gamma_b}} - m_{a_{\gamma_c\gamma_d}}|$, where $a, b, c, d \in \{1, 2, 3, 4\}$, is calculated. This ΔM tells us which two sets of photons have the lowest absolute difference in invariant mass, i.e., the two sets with the most optimal photon pairing. The pair of pseudoscalar candidates with the lowest ΔM are chosen for each event.

5.1 Event Selection BDT

The event selection BDT is a 4-photon event classifier used to distinguish between signal-like and background-like events. It functions to improve analysis sensitivity relative to a purely cut-based approach. Training samples are produced using signal MC and data. Data samples are processed to produce event mixed background samples then signal MC and event mixed background samples have all analysis selections applied. An additional step of adding a variable, m_{Hyp}, corresponding to the hypothesis mass point is done such that only one event selection BDT model is necessary for all pseudoscalar mass points. Fig. 5 shows the constructed distribution for m_{Hyp}. This variable, and dependent training variables, can be recalculated as the model is applied for various pseudoscalar mass points.

The result of applying these cuts is samples of pure signal events and (ideally) pure background events. These signal and background samples are used as inputs to the 4-photon BDT, where a subset of relevant variables are used to train the model. Only one model is trained for all mass points; the m_{Hyp} variable, and other variables dependant on it, allow the model to distinguish the hypothesis mass point when computing predictions. The training variables are listed as follows: γ_{1-4} MVA ID; $a_{1,2}$ p_T; $\Delta R(a_1, a_2)/m_{\gamma\gamma\gamma\gamma}$; $m_{a1} - m_{a2}$; $\cos\theta_{a\gamma}$, where $\theta_{a\gamma}$ is defined as the angle between the leading photon coming from

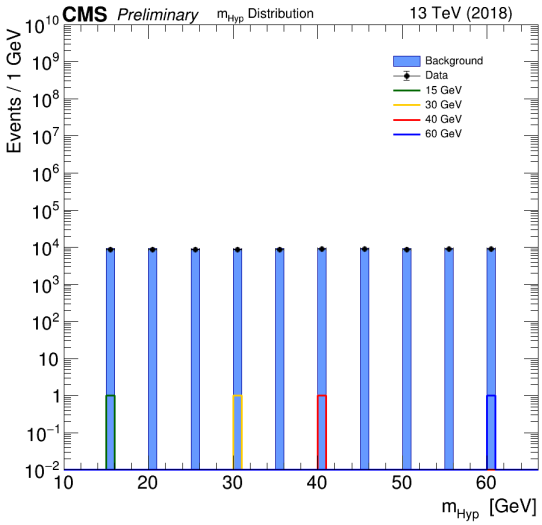

Figure 5: The m_{Hyp} distribution of signal MC for several pseudoscalar mass points, event mixed background, and data. m_{Hyp} is defined as the hypothesis mass for a given pseudoscalar mass point and as a flat function with discrete peaks for event mixed background and data, during event selection BDT training. m_{Hyp} and associated variables are recalculated for the appropriate pseudoscalar mass point when applying the model.

the leading pseudoscalar and the direction of $a \to \gamma\gamma$; $(m_{a1\,RECO} - m_{a\,Hyp})/m_{\gamma\gamma\gamma\gamma}$; and $(m_{a2\,RECO} - m_{a\,Hyp})/m_{\gamma\gamma\gamma\gamma}$. Two input variables for the training, with cuts applied, can be seen in Fig. 3 and Fig. 4 for various pseudoscalar mass points.

There is some disagreement between data and background distributions of the training variables; thus, a multi-dimensional, per-event reweighting is performed on the entire $m_{\gamma\gamma\gamma\gamma}$ range. The reweighting is calculated using the ratio of event mixed background and data in the $m_{\gamma\gamma\gamma\gamma}$ sideband region. The variables $\Delta\mathrm{R}(a_1, a_2)$, $a_{1,2}\ p_T$, and $m_{a1} - m_{a2}$ are used to calculate the per-event weights. The reweighted samples are used as input to train and test the BDT. A split on even/odd events is performed such that the training set consists of odd events and testing events consists of even events.

5.2 Categorization

Once the model is trained, it is applied to signal MC and event mixed background. All events from the training and testing set used in the previous step are considered. A categorization optimization procedure is applied on the prediction distributions to maximize the significance for each category for a minimal number of categories. The Approximate Mean Signficance is defined in Eq 1. It is known that the total significance increases with the number of categories; however, the number of categories is optimized such that a minimal number are created, which increases the number of signal events per category.

$$\mathrm{AMS} = \sqrt{2\left[(S+B)\ln\left(1+\frac{S}{B}\right) - S\right]} \tag{1}$$

To better emulate the prediction distribution of real data, a 1-dimensional reweighting

is performed. The new event weights are generated from the ratio of background and data in the sideband region. In order to minimize large statistical fluctuations with very fine binning but retain the high granularity, the event mixed background prediction distribution is smoothed using the *SmoothSuper* function in the *TGraphSmooth* class from ROOT. Before the smoothing, a cut on the BDT score is applied at ~ 0.1, which removes the background-like peak from the distribution, in order to improve the smoothing procedure.

The optimization procedure is applied and the category boundaries are generated. A requirement of at least 8 events in the signal region of the data samples is required when calculating category boundaries. Finally, these BDT score cuts are applied to signal and data samples such that there is a set of samples for each category. It is important to note that both the N-dimensional and 1-dimensional reweighting procedures do not affect real data in any way. The reweighting is used only as a tool to improve the effectiveness of the event selection BDT and the categorization optimization.

6 Statistical Analysis

6.1 Signal Modelling

A model of the signal shape for each nominal pseudoscalar mass point is needed. After all selections and categorization cuts are applied, only even signal events, used for testing the BDT, are kept for the signal modelling step. A signal model is computed for the $m_{\gamma\gamma\gamma\gamma}$ distribution from each signal sample. A separate signal model is generated for each category. There was a choice to fit models with a double-sided Crystal Ball function or the sum of Gaussians however, the Gaussian is shown to fit the best. Parametric signal models for $m_a = 15$ GeV and $m_a = 50$ GeV pseudoscalar mass points are shown in Fig. 6.

After signal models for nominal mass points are generated, models for intermediate mass points, e.g., 47 GeV, are required. It is known that the $m_{\gamma\gamma\gamma\gamma}$ resolution varies smoothly as a function of m_a [6]. Should the variation of mass resolution between each nominal mass point be fairly small, the parameters of the nominal mass points may be used. Normalization may also be interpolated from the smooth distribution of detector efficiency $\times$ analysis acceptance. This process for generating these intermediate mass point models is currently in progress.

6.2 Background Modelling

Although the background modelling step is incomplete, a short treatment will be provided. A background model is used to quantify the shape of $m_{\gamma\gamma\gamma\gamma}$ resulting from non-signal events. The background shape is modelled as one of the following functions: exponentials, Bernstein polynomials, Laurent series, and power law functions. The choice of function is treated as a nuisance parameter in the likelihood fit to data. Uncertainties are accounted for with the discrete profiling method, performed by the Higgs Combine statistical analysis software [10, 11]. The fits will be computed over the entire $110 < m_{\gamma\gamma\gamma\gamma} < 180$ GeV range. A background model will be generated for each pseudoscalar mass point.

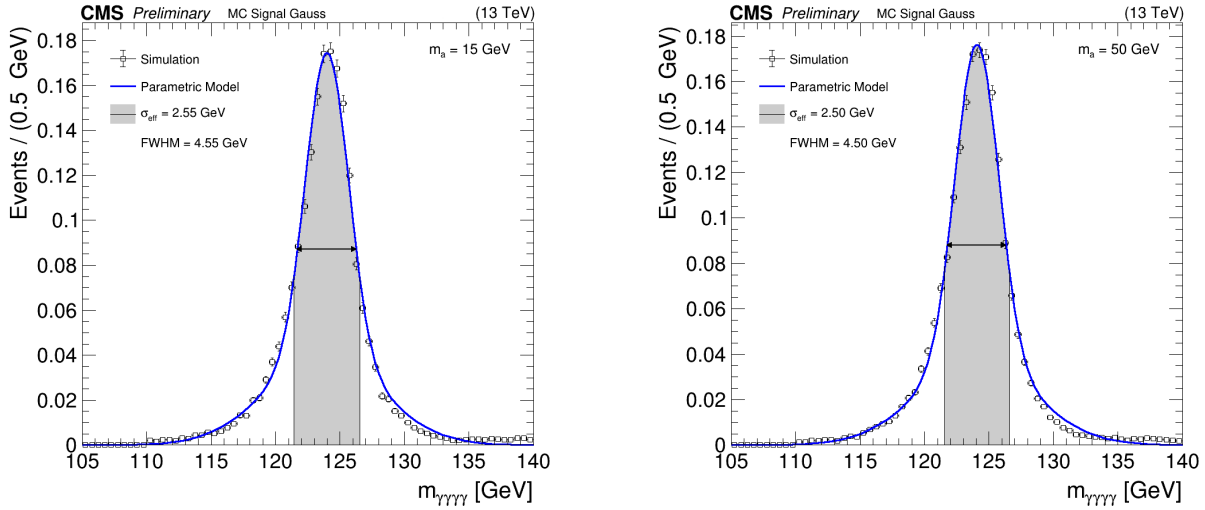

Figure 6: Examples of signal models of $m_{\gamma\gamma\gamma\gamma}$ for $m_a = 15$ GeV and $m_a = 50$ GeV pseudoscalar mass points. The parametric models are built from the sume of 2 Gaussians. The Full Width Half Maximum (FWHM) and σ_{eff}, defined as the region containing 68% of signal events, are calculated based on the parametric model. Normalization of the signal models is arbitrary.

7 Summary

A search for the SM Higgs boson decaying to pseudoscalars with a fully resolved, four photon final state is proposed. Many of the tools needed to perform the full analysis have been created – adjustments for new data notwithstanding – including a full event selection with an associated BDT, category optimization tools, and signal modelling tools for nominal mass points. The aim of a Run 3 study is to improve upon the existing Run 2 analysis sensistivity, regardless of whether this analysis utilizes purely Run 3 data or combines Run 2 and 3. Ideally, a Run 3 study could produce a discovery given better analysis sensistivity from increased statistics, reduced systematic error due to better precision in Run 3, better estimation of background, and a more optimized selection procedure via improved machine learning techniques. The plans for this analysis are to reach a level of acceptance of the recreation of the 2018 subset of the Run 2 analysis, achieving the previously set limits at a minimum, with the HiggsDNA framework such that we can begin working with Run 3 data. More work is still required to complete the 2018 proof-of-concept study. However, this study should enable a fast and smooth transistion to working with Run 3 data.

References

[1] D. Curtin, R. Essig, S. Gori, P. Jaiswal, A. Katz, T. Liu, Z. Liu, D. McKeen, J. Shelton, M. Strassler, Z. Surujon, B. Tweedie, and Y.-M. Zhong, "Exotic decays of the 125 GeV Higgs boson", Physical Review D **90**, 10.1103/physrevd.90.075004 (2014) 10.1103/physrevd.90.075004, https://doi.org/10.1103%2Fphysrevd.90.075004.

[2] A. M. Sirunyan et al. (CMS), "Particle-flow reconstruction and global event description with the CMS detector", JINST **12**, P10003 (2017) 10.1088/1748-0221/12/10/P10003.

[3] S. Chatrchyan et al. (CMS Collaboration), "The CMS experiment at the CERN LHC", Journal of Instrumentation **3**, S08004 (2008) 10.1088/1748-0221/3/08/S08004, https://dx.doi.org/10.1088/1748-0221/3/08/S08004.

[4] A. Sirunyan et al. (CMS Collaboration), "Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC", Journal of Instrumentation **16**, P05014 (2021) 10.1088/1748-0221/16/05/P05014, https://dx.doi.org/10.1088/1748-0221/16/05/P05014.

[5] "The CMS electromagnetic calorimeter project: Technical Design Report", (1997).

[6] A. Tumasyan et al. (CMS Collaboration), "Search for the exotic decay of the Higgs boson into two light pseudoscalars with four photons in the final state in proton-proton collisions at $\sqrt{s} = 13$ TeV", Journal of High Energy Physics **2023**, 10.1007/jhep07(2023)148 (2023) 10.1007/jhep07(2023)148, https://doi.org/10.1007%2Fjhep07%282023%29148.

[7] R. S. Maurizio Pierini, *Physics Coordination Report*, URL: https://indico.cern.ch/event/1180058/contributions/5560408/, Sept. 2023.

[8] L. Gray, N. Smith, B. Tovar, Y.-M. ". Chen, A. Novak, J. Chakraborty, N. Hartmann, P. Fackeldey, I. Krommydas, G. Watts, D. Thain, G. Stark, BenGalewsky, J. Rübenach, B. Fischer, D. Taylor, M. Aly, D. Kondratyev, P. Gessinger, J. Pata, A. Woodard, A. Potrebko, M. R., slehti, Z. Surma, C. Papageorgakis, K. Pedro, and dnoonan08, *CoffeaTeam/coffea: v0.7.22*, version v0.7.22, Oct. 2023, 10.5281/zenodo.8408347, https://doi.org/10.5281/zenodo.8408347.

[9] P. de Castro Manzano, M. Dall'Osso, T. Dorigo, L. Finos, G. Kotkowski, G. Menardi, and B. Scarpa, "Hemisphere Mixing: A Fully Data-Driven Model Of QCD Multijet Backgrounds For LHC Searches", PoS **EPS-HEP2017**, 370 (2017) 10.22323/1.314.0370.

[10] P. Dauncey, M. Kenzie, N. Wardle, and G. Davies, "Handling uncertainties in background shapes: the discrete profiling method", Journal of Instrumentation **10**, P04015 (2015) 10.1088/1748-0221/10/04/P04015, https://dx.doi.org/10.1088/1748-0221/10/04/P04015.

[11] C. H. C. Team, *CMS Higgs Combine: v9.1.0*, version v9.1.0, Mar. 2023, https://github.com/cms-analysis/HiggsAnalysis-CombinedLimit.