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 the universe, but there are yet many questions left unanswered. The SM does not account for dark matter, gravity, 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 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 photon

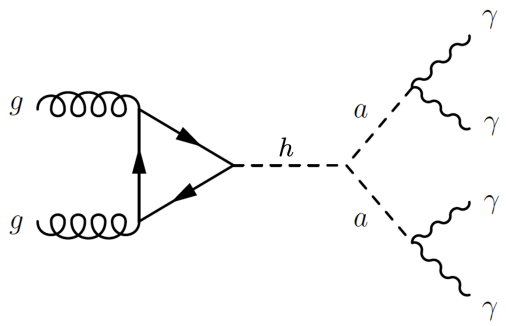

Figure 1: Tree-level Feynman diagram of $h \to aa \to \gamma\gamma\gamma\gamma$ for the gluon fusion Higgs production mode.

final state is shown in Fig 1. Given the current limits and exclusions, a large phase space 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. 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 $10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. 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 to even more. These particles first pass through the silicon tracker, which detects charged particles as tracks. Then the ECAL detects electromagnetically charged particles with lead tungstate crystals and is optimized for photons and electrons. 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, where η is the pseudorapidity. The $PbWO_4$ crystals measure the energy of any electromagnetically charged particle passing through them, albeit designed for optimal performance with electrons and photons. The crystals were chosen for their

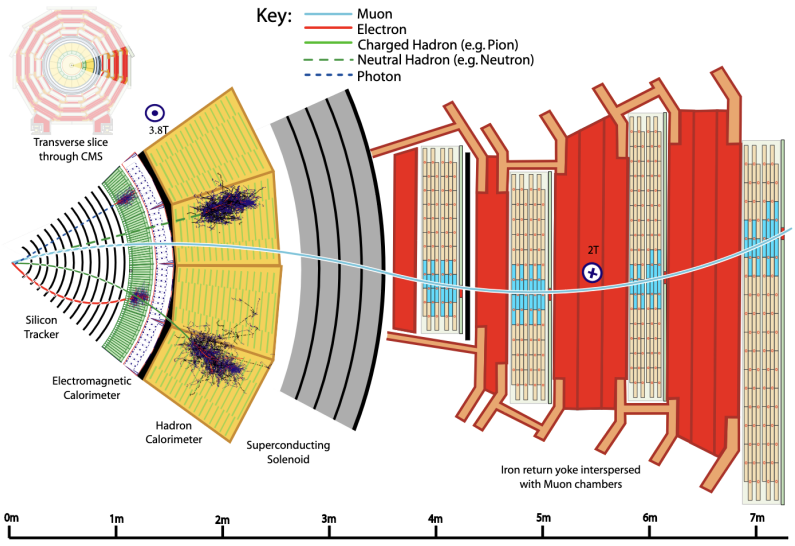

Figure 2: Slice showing CMS sub-detectors and how particles interact in them [2].

Molière radius and response time, i.e., a crystal contains 90% of a shower's deposition on average and the crystal is destructive to signals at a rate consistent with collisions. 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 have been done that include a Higgs boson decaying to two pseudoscalars, then each pseudoscalar decaying to two photons, as well as other final states. 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. 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 $50 - 100\%$ would allow for this analysis to set even stricter limits on the cross section times branching fraction, given by $\sigma(pp \to h) \times \text{Br}(h \to aa \to \gamma\gamma\gamma\gamma)$. 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 suggest an increase in statistics of about 50% [4]. The previous analysis reported an impact of $\sim$1% from systematic uncertainties, meaning that an increase in statistics would be an almost direct improvement to analysis

sensitivity [5]. 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 [6]. 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. These samples only consider the gluon fusion production mechanism of the Higgs boson and were generated using MADGRAPH5_aMC@NLO for the previous analysis. The 2018 signal samples were converted to nanoAOD in the manner prescribed above for the proposed analysis. 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 (QCD, $\gamma +$ jets, and $\gamma\gamma +$ jets) suffered from low statistics, and thus very large event weights, after requiring the presence of four photons in an event. 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 [7]. 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

Signal events are extracted from data via high level trigger (HLT) mimicking cuts, additional cuts and selections on the pseudoscalars and the four photon object, a BDT model to discriminate signal/background, and a categorization procedure based on BDT score. Events are selected using an HLT path optimized for the low mass diphoton Higgs boson search. The HLT path and associated mimicking cuts applied to 2018 UL samples are shown in Table 1. It is important to note that these trigger paths, cuts, and selections will all be evaluated, and potentially changed, for the Run 3 analysis. All of the listed criteria for the

	E_T	R_9	H/E	$\sigma_{i\eta i\eta}$	PF Pho Iso	Tracker Iso
EB; $R_9 > 0.85$	15.0	> 0.5	< 0.08	< 0.015	< 4.0	< 6.0
EB; $R_9 \leq 0.85$	15.0	> 0.5	< 0.08	< 0.015	< 4.0	< 6.0
EE; $R_9 > 0.9$	15.0	> 0.8	< 0.08	< 0.035	< 4.0	< 6.0

Table 1: Cuts mimicking the low mass diphoton Higgs boson HLT path: *HLT_Diphoton30_18R9IdL_AND_HE_AN_IsoCaloId_NoPixelVeto*.

2018 proof-of-concept study are taken from the Run 2 analysis, where optimization of cuts has already been performed.

The HLT is applied to both simulated and data samples. Events are also required to contain at least one diphoton candidate. Additional, more strict cuts are then applied to the photons from the diphoton candidate. Both photons must pass either $R_9 > 0.8$, Charged Hadron Isolation < 20 GeV, or (Charged Hadron Isolation)$/p_T < 0.3$ if $p_T > 14$ GeV and $H/E < 0.15$. Photons must also pass an electron veto; a photon candidate is rejected if its supercluster has a matching electron track with at least one hit in the pixel tracker. The cuts on diphoton candidates will be referred to as pre-selections.

Events that consist of at least one diphoton which pass the pre-selection requirements are subject to additional selections relating to the pseudoscalars and their daughter photons. The events must contain at least four photons where, in descending order of p_T, the leading two photons must pass γ_1 $p_T > 30$ GeV and γ_2 $p_T > 18$ GeV while the remaing two photons must pass a minimum p_T threshold of $p_T > 15$ GeV. There is some overlap with the pre-selections, however the additional cut of 15 GeV on $\gamma_{3/4}$ is necessary to achieve maximal performance from the event selection BDT. All photons must also fall within the tracker fiducial region, i.e., $|\eta| < 2.5$, and not within the EB-EE gap, $1.442 < |\eta| < 1.552$, where photon reconstruction is suboptimal. Finally, there is a restriction on the mass of the four photon object of $110 < m_{\gamma\gamma\gamma\gamma} < 180$ GeV. The signal region in this analysis is defined as $115 < m_{\gamma\gamma\gamma\gamma} < 135$ GeV while the sideband region is defined as the union of $110 < m_{\gamma\gamma\gamma\gamma} < 115$ GeV and $135 < m_{\gamma\gamma\gamma\gamma} < 180$ GeV.

After applying pre-selections and photon selections, the pseudoscalars 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 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. 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 for both samples. For the event mixed background sample, m_{Hyp} is a flat distribution with discrete peaks at each pseudoscalar mass point, by construction. When using the trained model to make predictions, m_{Hyp} and any variables that depend on it are recalculated as a discrete peak at the hypothesis mass point.

The result of applying these cuts is samples of pure signal events and 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 the leading pseudoscalar and the direction of $a \to \gamma\gamma$; $(m_{a1\,RECO} - m_{a\,Hyp})/m_{\gamma\gamma\gamma\gamma}$; and $(m_{a2\,RECO} - m_{a\,Hyp})/m_{\gamma\gamma\gamma\gamma}$.

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 know that the total significance increases with the number of categories, thus 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 data samples for the same bin is required when calculating category boundaries. Finally, these category cuts are applied to the signal and event mixed background such that there is a set of samples for each category. 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 training (even) signal events 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.

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 [5]. 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

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 [8, 9]. 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.

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 limits, regardless of whether this analysis utilizes purely Run 3 data or combines Run 2 and 3. 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. Some work is still required to complete the 2018 proof-of-concept study, which 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] R. S. Maurizio Pierini, *Physics Coordination Report*, URL: https://indico.cern.ch/event/1180058/contributions/5560408/, Sept. 2023.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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