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 with extended Higgs sectors. Some models predict the existance of a pseudoscalar boson, e.g., models with at least one extra Higgs doublet (2HDM), an $SU(2)_L$ singlet (SM+S), or both (2HDM+S) [1]. In these models, the SM Higgs boson can decay to two pseudoscalars then to four photons. The relevant Feynman diagram is shown in Fig 1.

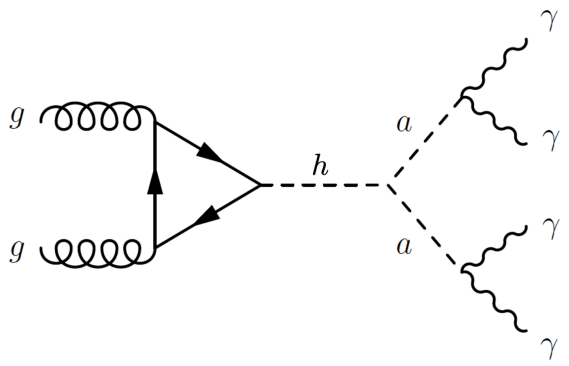

Figure 1: Tree-level Feynman diagram of $h \to aa \to \gamma\gamma\gamma\gamma$ for the gluon fusion Higgs production mode.

In 2HDM+S models, we have the SM-like Higgs boson, h, a heavy Higgs boson, H, two charged Higgs bosons, $H^{\pm}$, a scalar boson, s, and a pseudoscalar boson, a. 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 parits 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, enable a search for pseudoscalars with a mass range of $15 < m_a < 62$ GeV.

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 consists of 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, and then through the ECAL. 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, namely 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 any electromagnetically charged particle passing through them, albeit designed for optimal performance with electrons and photons. The crystals were chosen for their 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

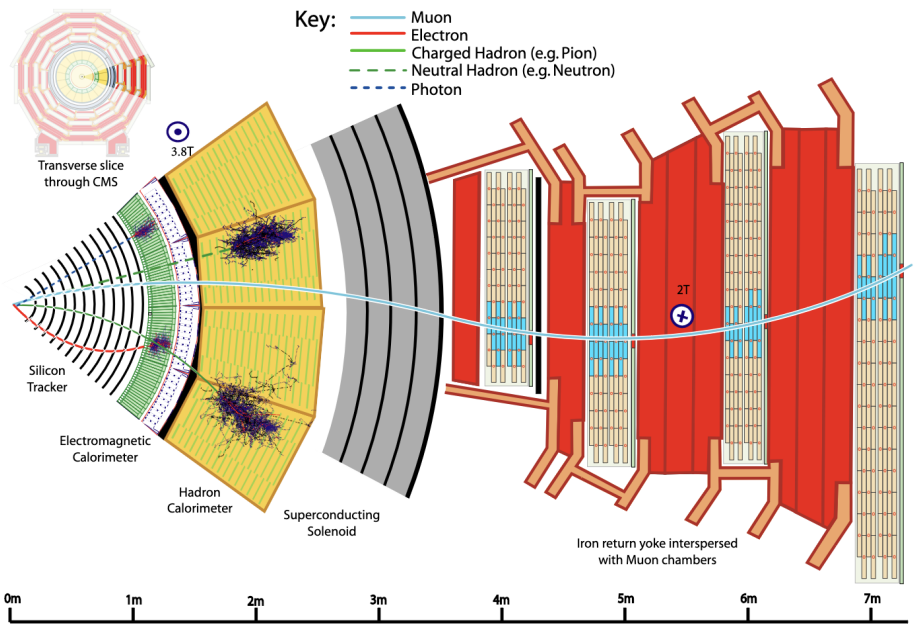

Figure 2: Slice showing CMS sub-detectors and how particles interact in them [2].

have an additional preshower component in two layers positioned in front of the crystals: lead and silicon strip detectors used to detect neutral pion decays.

Each crystal in the ECAL has photodetectors attached it: two avalanche photodiodes for crystals in the barrel and one vacuum phototriode for crystals in the endcap. These photodetectors are connected to the on-detector electronics with each trigger tower (corresponding to 5×5 crystals) consisting of five Very Front End boards, one Front End board, several Gigabit Optical Hybrids boards, one Low Voltage Regulator board, and a motherboard. Part of my technical contribution is working on the upgrade of the on/off-detector electronics for Phase II. These systems are also involved in calibrating the ECAL; there is a bias voltage over the photodetectors and measurements of this bias are used to calibrate them. There is also a laser system that is used to monitor crystal transparency due to radiation effects and to calibrate the crystal response.

2.2 Object Reconstruction

Translating the raw hits in the detector into usable objects is a crucial step in reconstructing a collision event. CMS uses an algorithm called *Particle Flow* (PF) to reconstruct objects for physics analysis [2]. Clustering and other reconstruction techniques create PF elements from the raw hits in the detector. These PF elements are run through a link algorithm that connects the PF elements from different sub-detectors while restricting the linking of PF elements to nearest neighbors in the (η, ϕ) plane, where η refers to pseudo-rapidity and ϕ refers to azimuthal coordinate in the detector.

The first step is reconstructing muons by matching tracks in the muon chambers with tracks in the inner tracker while imposing restrictions on calorimeter deposits. Once muons are reconstructed, all of the muon PF elements are masked for the next set of objects; a

similar masking step occurs for all PF elements after reconstruction. Electrons and isolated photons, i.e., photons without significant energy deposits in their vicinity, are then reconstructed by linking their tracks (or lack thereof), clusters in the ECAL, and lack of clusters in the HCAL; a Gaussian shower profile is used to assign the energy to clusters in the ECAL which is necessary for more complex event topologies with overlapping clusters. Proper understanding of shower shape in clusters is important for the ECAL since electrons and photons have a different shower shape than charged and neutral hadrons in the crystals [4]. A multivariate analysis technique (MVA) is used to distinguish prompt photons from jets using information regarding the photon's shower shape, isolation, energy, and η. A similar process is carried out for charged hadrons, with the addition of clusters in the HCAL and a calorimetric energy resolution cut on track p_T. Remaining PF elements are subject to the cross-identification of charged hadrons, neutral hadrons, and non-isolated photons/electrons, arising from parton fragmentation, hadronization, and decays in jets. These reconstructed objects are then used as inputs to jet reconstruction algorithms, e.g., anti-k_T.

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.

3.1 Advancements on Previous Studies

Advancements can be made relative to the Run 2 analysis in several ways. An increase in statistics of $50-100\%$ could allow for this analysis to set even stricter limits on the cross section and branching fraction, $\sigma(pp \to H) \times \text{Br}(H \to aa \to \gamma\gamma\gamma\gamma)$. The $132\,\text{fb}^{-1}$ of data analyzed in the previous analysis The previous analysis reported an impact of $\sim$1% from systematic uncertainties, meaning that an increase in statistics would be a direct improvement to the analysis sensitivity and systematic uncertainties would be of little hindrance [5]. The true increase in statistics is heavily dependent on both CMS and the LHC, but the projections for availability of 2022 and 2023 data by mid-2024 suggest an increase in statistics of about 50% [6]. It may also be possible to reduce the $\sim$1% impact from statistical uncertainties since Run 3 data is expecting a gain of $\sqrt{2}$ times precision on Run 2.

A new methodology may also be employed to increase sensitivity with respect to the previous analysis. It may be possible to utilize a different type of machine learning (ML) technique, previously a boosted decision tree (BDT), to train an event selection model; the new technique would likely use deep learning to construct the model. Deep learning models may require more events than are available to be viable as these models only surpass BDTs when trained with at least $\sim$100k events [7]. While the expected yield from $132\,\text{fb}^{-1}$ of data is nowhere near the $\sim$100k events required, the number of events used to train a ML model

is several orders of magnitude higher than the yield since training happens before several steps that involve cuts on events. Any lack of statistics required for a deep learning model can be alleviated by two things: generating additional signal samples (more computationally expense) and the inclusion of Run 3 data in the proposed analysis since the background for this analysis is data driven. It is yet unclear whether this analysis will be a partial Run 3 or a combination of the full Run 2 and partial Run 3 data; regardless, the increase in statistics will improve the viability of utilizing a deep learning model instead of a BDT.

A major difference between this proposed analysis and its inspiration, the Run 2 analysis, changes the tools necessary to search for this process. Arguably, the most significant change is the use of the newer nanoAOD data format as opposed to the use of the miniAOD data format. On one hand, it limits the use of certain event selection techniques and, on the other, requires an entirely new analysis framework. An analysis framework handles everything from loading samples to processing cuts and selections to applying systematic uncertainties to samples. The frameworks used in analyses with miniAOD and nanoAOD are common to the $H \to \gamma\gamma$ group: Flashgg for miniAOD, like the Run 2 analysis, and HiggsDNA for nanoAOD, like this proposed analysis [8, 9].

3.2 New Framework: HiggsDNA

The new analysis framework HiggsDNA, produced by the $H \to \gamma\gamma$ group, is built on Coffea, a tool for processing nanoAOD samples [10]. 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. Coffea processors enable the efficient use of cuts across an entire column (in the context of a flat tree), e.g., a simple p_T cut is now applied simultaneously to the entire column instead of looping through each event and applying the cut per event.

A proof of concept is needed before applying HiggsDNA to a Run 3 analysis so a partial recreation of the Run 2 analysis is underway. The partial recreation is for 2018 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. It should be noted that all samples for the year 2018 from the Run 2 analysis have been converted from miniAOD to nanoAOD with some additional, non-standard branches that are used by the $H \to \gamma\gamma$ group; the subset of these branches that are required for HLT mimicking cuts are *chargedHadronIso*, *trkSumPtHollowConeDR03*, and *pfPhoIso03*, and *fixedGridRhoAll* for pileup corrections. These branches are included in newer versions of nanoAOD, but are not included in the 2018 Ultra Legacy (UL) samples since they are in nanoAODv9. In the case of 2022-onwards data, the centrally produced nanoAOD samples should be nanoAODv12, which has all but *chargedHadronIso*. Unfortunately, the use of even this modified nanoAOD makes it impossible to utilize a BDT to determine the primary vertex, which can occasionally differs from the standard CMS primary vertex ($\max \Sigma p_T$), as part of the event selection without significant effort. The $H \to \gamma\gamma$ group has a tool to include the above branches but adding those necessary for a vertex BDT is a major challenge. The lack of a vertex BDT additional step in the event selection process reduces the Higgs mass resolution by about 3%. Lastly, scaling and smearing corrections for photons have not been applied in this analysis thus far as there seem to be some bugs in the HiggsDNA framework.

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, γ + 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. The same technique will be used in this analysis to avoid ths same problems. The data driven background utilizes a technique similar to hemisphere mixing to remove the presence of any signal and artificially create the background shape [11]. This technique is called *event mixing* and is a simplification of the hemisphere mixing procedure. The methodology of event mixing varies between columnar-based analysis and event-based analysis; columnar-based event mixing is significantly less computationally expensive. Event mixing shuffles the photons in an event with the photons from proceeding events, modulo the number of events. Considering photons in descending order by p_T for an event N, the first photon is unchanged in event N, the second photon is replaced with the photon from event $N+1$, the third is replaced with the photon from event $N+2$, and the forth is replaced by the photon from event $N+3$.

The event mixed background samples were generated in several steps. 2018UL data samples were converted from miniAOD to nanoAOD with the necessary extra branches, thus diverging slightly from the central 2018UL data samples and the 2018 non-UL samples used in the full Run 2 analysis. The resulting UL samples then underwent the event mixing procedure before being processed in the event selection step. Raw event mixed background samples were not stored as a space saving measure. Bias studies will be performed on the newly generated background samples as was done in the Run 2 analysis.

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 for 2018 are shown in Table 1. 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 and are shown in Table 2. 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 in this analysis.

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

Table 2: Additional cuts applied to diphoton candidates.

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 cuts, 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. No cuts are applied to directly to the pseudoscalar candidates.

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.

Table 3: Variables used to train the 4-photon event selection BDT.

The result 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 shown in Table 3. There is some disagreement between data and background distributions of the training variables so an N-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 are considered, both training and testing sets used in the previous step. A categorization optimization procedure is applied on the predicition 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.

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

In order to minimize large statistical fluctuations with very fine binning but retain the high granularity, the event mixed background predicition 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. Only one category was necessary in the Run 2 analysis, but that is not the case currently with this analysis. The current idea is that this is due to a combination of factors, namely a difference in the quality of the event selection BDT, the lack of scaling and smearing corrections, and the lack of a vertex BDT.

Finally, these category cuts are applied to the signal and event mixed background such that there is a set of samples for each category. Only training (even) signal events are kept after categorization for the signal modelling step.

6 Signal Modelling

A model of the signal shape for each nominal pseudoscalar mass point is needed. After all selections and categorization cuts are applied, 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. Note that due to current issues with the BDT, the categorization step is imperfect and thus, the signal modelling step suffers as well. Models are fit with a Double-sided Crystal Ball and with the sum of Gaussians however, the Gaussian is shown to fit the best. The proper number of Gaussians used to fit the signal $m_{\gamma\gamma\gamma\gamma}$ distribution is reached when $-2\Delta\log(\mathcal{L}) < 0.05$, where $\mathcal{L} = L_{N+1}/L_N$, L_N is the likelihood, and N is the number of Gaussians. The likelihood L is computed as the probability of a given χ^2 for a number of degrees of freedom, using ROOT's *TMath::Prob*. The ideal number of Gaussians to fit is found to be two.

7 Summary

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

[7] S. May, "Machine learning in CMS", International Journal of Modern Physics A **37**, 2240020 (2022) 10.1142/S0217751X22400206, https://doi.org/10.1142/S0217751X22400206.

[8] *Flashgg*, version 10.6.29, 2023, https://github.com/cms-analysis/flashgg.

[9] *HiggsDNA*, 2023, https://gitlab.cern.ch/HiggsDNA-project/HiggsDNA.

[10] 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.

[11] 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.