PHYSICAL REVIEW D **96,** 035022 (2017)

Can we discover double Higgs production at the LHC?

Alexandre Alves,[1,*] Tathagata Ghosh,[2,†] and Kuver Sinha[3,4,‡]

[1]*Departamento de Física, Universidade Federal de São Paulo, Diadema-SP, São Paulo 09972-270, Brazil*
[2]*Department of Physics and Oklahoma Center for High Energy Physics, Oklahoma State University, Stillwater, Oklahoma 74078-3072, USA*
[3]*Department of Physics and Astronomy, University of Oklahoma, Norman, Oklahoma 73019, USA*
[4]*Department of Physics and Astronomy, University of Utah, Salt Lake City, Utah 84112, USA*

(Received 7 June 2017; published 24 August 2017)

We explore double Higgs production via gluon fusion in the $b\bar{b}\gamma\gamma$ channel at the high-luminosity LHC using machine learning tools. We first propose a Bayesian optimization approach to select cuts on kinematic variables, obtaining a 30%–50% increase in the significance compared to current results in the literature. We show that this improvement persists once systematic uncertainties are taken into account. We next use boosted decision trees (BDT) to further discriminate signal and background events. Our analysis shows that a joint optimization of kinematic cuts and BDT hyperparameters results in an appreciable improvement in the significance. Finally, we perform a multivariate analysis of the output scores of the BDT. We find that assuming a very low level of systematics, the techniques proposed here will be able to confirm the production of a pair of standard model Higgs bosons at 5σ level with 3 ab^{-1} of data. Assuming a more realistic projection of the level of systematics, around 10%, the optimization of cuts to train BDTs combined with a multivariate analysis delivers a respectable significance of 4.6σ. Even assuming large systematics of 20%, our analysis predicts a 3.6σ significance, which represents at least strong evidence in favor of double Higgs production. We carefully incorporate background contributions coming from light flavor jets or c jets being misidentified as b jets and jets being misidentified as photons in our analysis.

DOI: 10.1103/PhysRevD.96.035022

I. INTRODUCTION

Measuring possible deviations of the triple Higgs coupling λ_3 from its predicted standard model (SM) value is a key goal of future colliders. This has implications for a whole range of new physics scenarios, such as supersymmetry and other extensions of the SM with two Higgs doublets. The cosmological implications are also profound, since λ_3 is related to the strength of the electroweak phase transition which is critical for understanding electroweak baryogenesis, for example.

The triple Higgs coupling can be probed by Higgs pair production processes, which have been extensively studied in the context of the high-luminosity LHC and future hadron colliders. Higgs pair production occurs dominantly via gluon fusion, with other production processes being more than an order of magnitude smaller. Final states that have been studied, in the context of di-Higgs production at the LHC, include $b\bar{b}\gamma\gamma$ [1–7], $b\bar{b}\tau^+\tau^-$ [8,9], $b\bar{b}W^+W^-$ [10], and $b\bar{b}b\bar{b}$ [11–13].

The purpose of this paper is to investigate the prospects of Higgs pair production at the LHC in the $b\bar{b}\gamma\gamma$ channel. Our analysis builds on previous studies in two ways: we use tools from the machine learning (ML) literature in our analysis, and we carefully account for background contributions coming from light flavor jets (j) or c jets being misidentified as b jets and electrons or jets being misidentified as photons. With regard to the use of ML tools, we note that this is somewhat hostile terrain for theorists. However, the comparative gains in discovery prospects over other methods, which we discuss at length, will hopefully convince the reader that planning for future colliders should exploit state of the art data analysis tools to ensure that projections are reasonable.

For the benefit of the reader, we chart out the steps in our analysis and the main results of each step. We present the details of our signal and background simulation in Sec. II. We provide a brief discussion on previous studies in Sec. III.

In Sec. IV, we ask the question: given an event topology and a set of kinematic observables, is there a systematic and computationally feasible method to obtain the most optimal selections that maximize the significance? We show that Bayesian optimization, as described in Refs. [14,15], performs better than selections currently proposed in the literature, and is computationally much more tractable than a brute force multivariable scan. We demonstrate our results with the Python algorithm `Hyperopt` [16]. Our main results of this section are presented in Fig. 2, and we find that there is a 30%–50% increase in the significance metric

*aalves@unifesp.br
†tghosh@okstate.edu
‡kuver.sinha@ou.edu

$S/\sqrt{B}$ compared to current results in the literature. Moreover, this relative improvement persisted after incorporating systematic uncertainties on the background rate, as demonstrated in Fig. 3.

In Sec. V, we build on the Bayesian optimization of kinematic cuts, and show that training a Boosted Decision Tree (BDT) algorithm to better classify signal and backgrounds events, in addition to the procedure of using optimal cuts to select the best volume of the features space for the BDT training, increases the discovery prospects dramatically. For our calculations, we use the `XGBoost` [17] implementation of BDTs for Python. We present our results in three stages. In Sec. V A, we first introduce the kinematic observables used in the BDT analysis, and provide a discussion of the interplay between BDT classifiers and cut selections, without addressing the question of cut optimization. In Sec. V B, we sequentially optimize the cuts on the kinematic observables using `Hyperopt`, and then optimize the BDT hyperparameters. Finally, in Sec. V C, we perform a joint optimization of the kinematic cuts and the BDT hyperparameters.

Our results from this stage of the analysis are summarized in Table V and Fig. 8. We find that the use of BDT enhances the significance irrespective of the kinematic cuts used. The largest enhancement, however, occurs with cuts optimized using `Hyperopt`, and we reach a significance of 3.88 for 3000 fb^{-1} of data.

In Sec. VI, we focus on the statistical side of the analysis by estimating the log-likelihood ratio statistics from the output scores of the BDTs provided by `XGBoost`, following [18].

The final results of our paper are presented in Table VII. We find that assuming a very low level of systematics, the techniques proposed here will be able to confirm the production of a pair of SM Higgs bosons at 5σ level. Assuming a more realistic projection of the level of systematics, around 10%, the optimization of cuts to train BDTs combined with a multivariate analysis delivers a respectable significance of 4.6σ. This is the largest significance achieved so far in the $b\bar{b}\gamma\gamma$ channel with realistic assumptions concerning backgrounds and systematic uncertainties at the 14 TeV LHC. Even assuming large systematics of 20%, our analysis predicts a 3.6σ significance, which represents at least strong evidence in favor of double SM Higgs production.

We pause to make a few comments about signal and background event rate estimation before proceeding with our analysis. There has been considerable disagreement about this in the literature, with some of the older studies giving optimistic results due to an underestimation of background. We discuss these issues in Sec. III, where we compare and summarize previous studies. Throughout this work, we will take the background and signal event rates of Azatov *et al.*, Ref. [4], which we consider robust, as a reference point. However, we are also careful to incorporate the backgrounds $c\bar{c}\gamma\gamma$, $b\bar{b}\gamma j$ and $c\bar{c}\gamma j$, whose importance has been highlighted by ATLAS Ref. [5].

In Appendix A, we briefly comment about the metrics used to compute the statistical significances and, in Appendix B, we show a Python snippet of a simple code to implement the selection cuts optimization based on `Hyperopt`.

II. DETAILS OF $pp \to b\bar{b}\gamma\gamma$ SIMULATIONS

The details of the signal and background simulation will be presented in this section.

Instead of reevaluating all cross sections for the process of interest, the strategy we will pursue in this work is to assume the production rates presented in Ref. [4]. In our opinion, the calculations performed by Azatov *et al.* are reliable enough to be used as a starting point, especially given that we are interested in a close comparison of our results with those previously obtained in the literature. We will use events generated only as a means of estimating the kinematic distributions germane to the cut-and-count analysis and to train our ML algorithms. We do, however, take into account three additional subdominant backgrounds beyond those of Ref. [4].

A. Higgs pair production

For the simulation of the signal and backgrounds events, we use `MadGraph5_aMC@NLO_v2.3.3` [19] with the CTEQ5L [20] and CTEQ6L [21] parton distribution functions, respectively. At the leading order (LO), there are two one-loop diagrams that contribute to the process $pp \to hh$ [22–24] and they interfere destructively. While the triangle diagram is sensitive to the Higgs trilinear coupling, λ_3, the box diagram is not. The simulation of our signal includes the effect of both these diagrams. However, over the last 30 years significant improvement on the theoretical calculation of this process to higher orders [25–31] has taken place.

In Ref. [4], the signal cross section at the 14 TeV LHC was calculated at LO with `MadGraph5_aMC@NLO_v2.1.1` and then multiplied by the partial NNLO K-factor of 2.27 [27], calculated in the large quark mass limit. The resulting production cross section is 36.8 fb. The combined branching ratio of the $b\bar{b}\gamma\gamma$ channel is small, only 0.264%. The number of signal events after 3000 fb^{-1}, before cuts and efficiencies, is around 290.

Effects of the finite top quark mass to the NLO QCD cross section of Higgs pair production has been taken into account in Refs. [32,33]. The full mass dependence diminishes the NLO prediction by 14% compared to the large top quark mass approximation, however approximated NNLO effects increase the NLO predictions by ~20% according to Ref. [34], therefore, the K-factor adopted by Azatov *et al.* constitutes a fair approximation to the total rate.

Hard jet radiation and finite top quark mass effects are also expected to change the shape of distributions involving the four-momenta of the reconstructed Higgs bosons at higher orders as shown in Refs. [32–35].

In order to obtain the distributions of the kinematic variables of interest, we pass our simulated events to PYTHIA_v6.4 [36] for showering and hadronization. Finally, these events are passed to DELPHES_v3.3 [37] for detector simulation. For the signal, the Higgs bosons are decayed into bottom quarks and photons with the MadSpin module of MadGraph5. In contrast, for the relevant backgrounds which contain a Higgs in the final state, the Higgs boson has been decayed within PYTHIA. Photon isolation criteria and jet clustering are similar of those of Azatov *et al.* who found that their results do not differ much from other works with somewhat different criteria.

Both signals and backgrounds were required to pass the following minimal selection criteria

$$p_T(j) > 20\ \text{GeV}, \qquad p_T(\gamma) > 20\ \text{GeV},$$
$$|\eta(j)| < 2.5, \qquad |\eta(\gamma)| < 2.5 \tag{2.1}$$

$$100\ \text{GeV} < |M_{jj}| < 150\ \text{GeV},$$
$$100\ \text{GeV} < |M_{\gamma\gamma}| < 150\ \text{GeV}. \tag{2.2}$$

In the next section, we comment about the backgrounds and give further details of the computations.

It is important to stress that a better estimation of production rates and invariant mass distributions would mainly require including the effects of the finite top quark mass and higher-order corrections. That, however, is beyond the scope of this work.

B. Backgrounds

We have evaluated the backgrounds to $(h \to b\bar{b}) + (h \to \gamma\gamma)$ signal from multiple SM processes:

(1) $b\bar{b}\gamma\gamma$;
(2) Zh, $Z \to b\bar{b}$ and $h \to \gamma\gamma$;
(3) $b\bar{b}h$, $h \to \gamma\gamma$;
(4) $t\bar{t}h \to b\bar{b} + \gamma\gamma + X$;
(5) $jj\gamma\gamma$, where the light-jets jj are mistaken for a b-jet pair in the detector;
(6) $b\bar{b}jj$, where the light-jets jj are mistaken for a photon pair in the detector;
(7) $c\bar{c}\gamma\gamma$, where a c jet is mistagged as a b jet;
(8) $b\bar{b}\gamma j$, one light-jet is mistaken for a photon;
(9) $c\bar{c}\gamma j$, the c jets are mistagged as bottom jets and the light-jet as a photon.

The cross section normalizations for the backgrounds from 1 to 5 are taken from Ref. [4]. In that work, the continuum $b\bar{b}\gamma\gamma$ is computed at LO with one extra jet radiation and a K-factor of 2 is estimated for the NLO QCD corrections.This large K-factor for the dominant background has been neglected in many previous studies in this channel. The backgrounds Zh and $b\bar{b}h$ were also evaluated with one extra jet radiation to estimate the higher-order QCD corrections. The $t\bar{t}h$ K-factor was taken from [38] and it is small. The signal and backgrounds estimates of Azatov *et al.* are found to agree reasonably well of the Snowmass group report of Ref. [39].

Our backgrounds events (1–4) are also generated with 1 extra parton radiation in order to better simulate the kinematic distributions. MLM scheme [40] of jet-parton matching has been utilized to avoid double counting. The extra hard jet was included in the $b\bar{b}\gamma\gamma$ background once it is the dominant one. The reason for including the extra QCD radiation in the ressonant backgrounds $t\bar{t}h$, Zh and $b\bar{b}h$ is that the Higgs boson recoils against the extra hard jets which is important to obtain the $M_{b\bar{b}\gamma\gamma}$ invariant mass distribution. Unfortunately, it is computationally too expensive to simulate the signals in the same way, and beyond our means.

The $t\bar{t}h$ background is simulated in the inclusive way. Events with hard charged leptons are easily classified as backgrounds events however and efficiently discarded as we are going to see.

Background processes with light jets are important when a jet radiates a hard photon which is mistaken for an isolated photon in the detector. This is the case of the backgrounds $b\bar{b}jj$, $b\bar{b}\gamma j$ and $c\bar{c}\gamma j$. All the backgrounds from 5 to 9 in the above list were simulated with MadGraph5_aMC@NLO_v2.3.3 at LO and multiplied by the NLO QCD K-factors presented in Ref. [41].

Following previous studies [1–6], we adopt the probability of 1.2×10^{-4} for a light-jet to be mistagged as a photon. However, in the presence of pileup events this value might be an underestimate [7]. Nevertheless, the $b\bar{b}jj$ background was found to be negligible after imposing cuts and mistagging factors.

Finally, for $c\bar{c}\gamma\gamma$ backgrounds where a c jet is mistagged as a b jet, the $b-$ and c-tagging, and also the light-jet mistagging are parametrized according to the jet's transverse momentum and rapidity as implemented in Delphes, specifically as the default simulation of the CMS detector. The Delphes parametrization assumes that a 70% b-tagging efficiency is reached for $p_T > 100$ GeV at the cost of a 20(5)% mistagging factor for $c(j)$ jets. These subdominant backgrounds $b\bar{b}\gamma j$, $c\bar{c}\gamma\gamma$ and $c\bar{c}\gamma j$ were not taken into account in the majority of the previous studies we are considering in this work for comparisons, except for [3,5,7]. All the uncertainties in the backgrounds rates are taken into account in this work as systematic uncertainties in the calculation of the signal significances.

The numbers of backgrounds events after imposing the basic cuts of Eq. (2.2) for 3 ab^{-1} of integrated luminosity is shown in Table I.

In the next section, we will investigate a method to optimize the cut-and-count analysis, instead of manual

TABLE I. The number of signal and the various types of backgrounds considered in this work after imposing the basic cuts of Eq. (2.2) for 3 ab^{-1} of data. We found $b\bar{b}jj$ negligible after cuts and estimating the probability of the jet pair fakes a photon pair.

Signal	$b\bar{b}\gamma\gamma$	$c\bar{c}\gamma\gamma$	$jj\gamma\gamma$	$b\bar{b}\gamma j$	$t\bar{t}h$	$c\bar{c}\gamma j$	$b\bar{b}h$	Zh	Total backgrounds
42.6	1594.5	447.7	160.3	137	101.1	38.2	2.4	1.8	2483

tuning of cut thresholds as is commonly done. This requires us to plant ourselves on a set of baseline results and cut strategies, but also to adopt the signal and background normalizations of this baseline work. We chose to adopt the results, cuts and normalization of Ref. [4] as our baseline due their careful treatment of signals and backgrounds concerning QCD higher-order effects. As we will show, this work also presents the best cut strategy when compared to other theoretical and experimental works. On the other hand, we go beyond that work by including the subdominant backgrounds $b\bar{b}\gamma j$, $c\bar{c}\gamma\gamma$ and $c\bar{c}\gamma j$. Our simulations for these backgrounds agree reasonably well with those from [3,5,7].

III. COMPARISON AND SUMMARY OF PREVIOUS STUDIES

In this section, we present a summary of previous studies of double Higgs production at the LHC, taking Refs. [1–6]

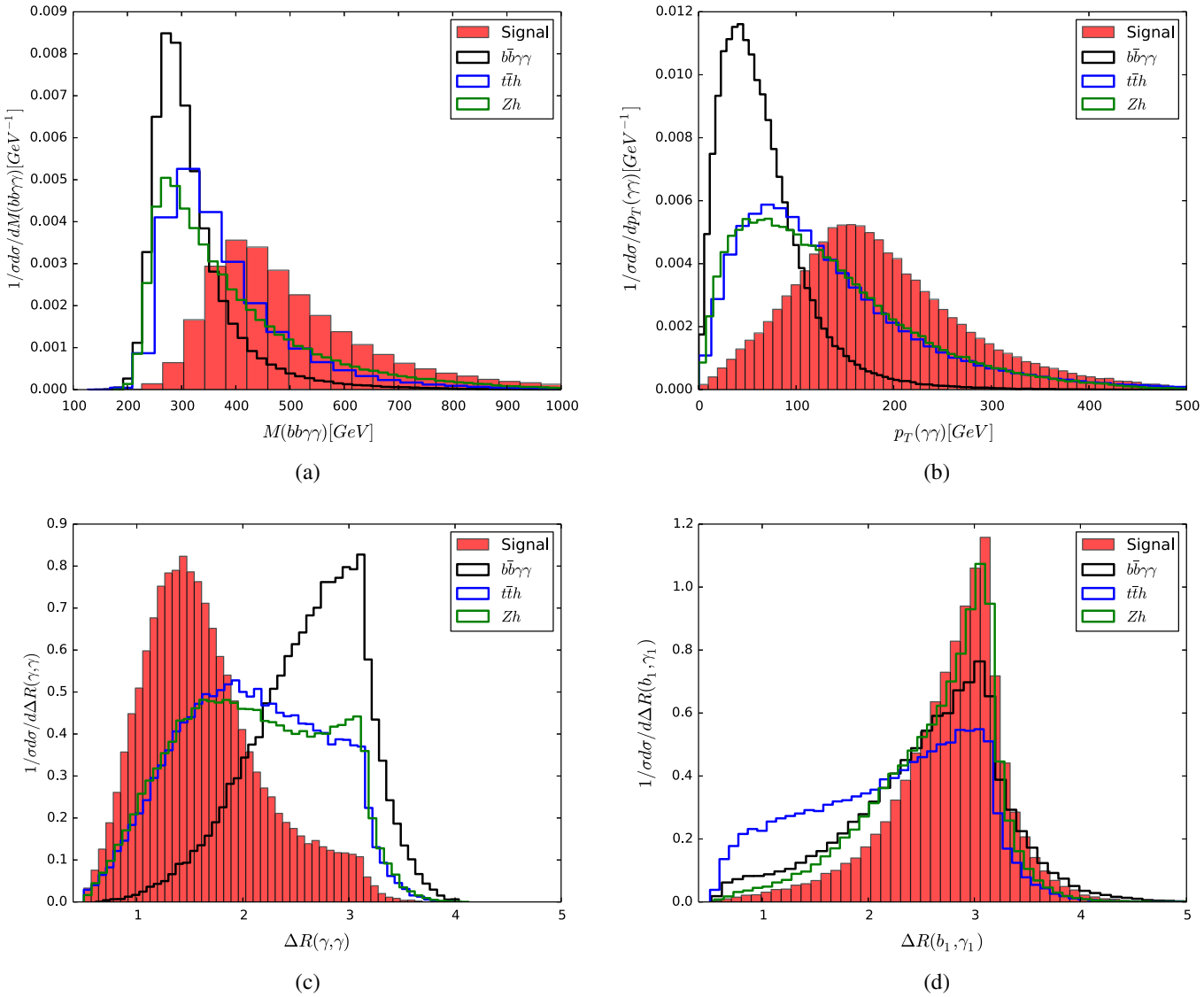

FIG. 1. Kinematic distributions of the signal (shaded red), and the backgrounds $b\bar{b}\gamma\gamma$ (black), $t\bar{t}h$ (blue) and Zh (green) are displayed. In (a), we show the invariant mass of two b jets and two photons. In (b), we show the transverse momentum of a pair of photons. In (c) and (d), we show the distance ΔR between a pair of photons, and between the hardest photon and the hardest b jet, respectively.

as representatives. Our main goal here is to show that despite the varying levels of rigor in terms of calculating signal and background event rates, and the differences in selection strategies, the cut and count analyses employed in these disparate studies yield similar significances.

The process $pp \to hh \to b\bar{b}\gamma\gamma$, with final states containing two b jets and two hard photons, presents many features which make it possible to employ a large variety of kinematic variables and selection strategies. We describe the most pertinent ones below:

(1) transverse momentum of b jets and photons: $p_T(b)$ and $p_T(\gamma)$
(2) $b\bar{b}$ and $\gamma\gamma$ invariant masses: M_{bb} and $M_{\gamma\gamma}$, where signal events exhibit resonance peaks at m_h
(3) transverse momentum of $b\bar{b}$ and $\gamma\gamma$: $p_T(bb)$ and $p_T(\gamma\gamma)$
(4) invariant mass of two b jets and two photons: $M_{bb\gamma\gamma}$
(5) distance between pairs of b jets and photons: $\Delta R(bb)$, $\Delta R(\gamma\gamma)$ and $\Delta R(b\gamma)$, where $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ in the pseudo-rapidity and azimuthal angle plane (η, ϕ)
(6) the fraction $E_T/M_{\gamma\gamma}$ for the two hardest photons in the event; these are variables used in experimental searches as in Ref. [42,43]

Some of these kinematic distributions have been presented in Fig. 1 for the signal, and continuum $b\bar{b}\gamma\gamma$, $t\bar{t}h$ and Zh backgrounds. In panel (a), we show the invariant mass of two b jets and two photons. In panel (b), we show the transverse momentum of a pair of photons. In panels (c) and (d), we show the distance ΔR between a pair of photons, and between the hardest photon and the hardest b jet, respectively.

In Table II, we display the analyses performed by the representative theory groups, along with the ATLAS study [5], which is shown in the last row. The first column gives the relevant reference, while the second column gives the kinematic variables and selections that were applied in the corresponding paper. The different groups made very different signal and background estimates, and we refer to Ref. [4] for a detailed discussion of these differences. For the significance calculations shown in the final column, we take all signal and background cross sections to be normalized to the values obtained by Ref. [4], which, in our opinion, is the most robust theory study. However, we also take into account the backgrounds $c\bar{c}\gamma\gamma$, $b\bar{b}\gamma j$ and $c\bar{c}\gamma j$ which were not taken into account in Ref. [4].

The final column of Table II thus shows the performance that each group would have had with its selection strategies, if all cross sections had been normalized by the ones of Ref. [4] and if $c\bar{c}\gamma\gamma$, $b\bar{b}\gamma j$ and $c\bar{c}\gamma j$ backgrounds had been taken into account. The statistical significance for each study is calculated with the naive metrics of Eq. (A1), for 3 ab^{-1} of data with no systematic uncertainties. The numbers inside parenthesis denote the S/B ratio of each study.

Our main message from Table II is that the different search strategies employed by the groups yield similar significances, once signal and background cross sections are normalized to the proper value. In other words, the selections and cut and count analysis of any particular group does not radically outperform that of any other.

We now discuss the studies conducted by the different groups in more detail.

TABLE II. In the first column at left, we show the literature references of each cut strategy displayed at the second column. In the last column, we compute the signal significance with the number of signal and background events estimated in this work. The number inside parenthesis in the last columns are the signal-to-background ratios. We took the $c\bar{c}\gamma\gamma$, $b\bar{b}\gamma j$ and $c\bar{c}\gamma j$ backgrounds into account but no systematics. The approximated mean significance (AMS) function significance is that of Eq. (A3).

Reference	Kinematic cuts	AMS(σ) (S/B)
(A) [1]	$p_{T_{\gamma(b)}} > 20(45)$ GeV, $\lvert\eta_{b,\gamma}\rvert < 2.5$ $\lvert M_{bb} - m_h\rvert < 20$ GeV, $\lvert M_{\gamma\gamma} - m_h\rvert < 2.3$ GeV $\Delta R_{b\gamma} > 1.0$, $\Delta R(\gamma\gamma) < 2.0$	1.54(0.30)
(B) [2]	$p_{T_{b,\gamma}} > 50$ GeV, $\lvert\eta_{b,\gamma}\rvert < 2.5$, $\Delta R_{b\gamma} > 0.4$, $\Delta R(bb) < 2.5$ $110 < M_{bb} < 135$ GeV, $\lvert M_{\gamma\gamma} - m_h\rvert < 5$ GeV, $M_{bb\gamma\gamma} > 350$ GeV $\lvert\eta_H\rvert < 2$, $P_{T_H} > 100$ GeV	1.33(0.39)
(C) [3]	$p_{T_{b,\gamma}} > 30$ GeV, $\lvert\eta_{b,\gamma}\rvert < 2.5$ $\lvert M_{bb} - m_h\rvert < 12.5$ GeV, $\lvert M_{\gamma\gamma} - m_h\rvert < 5$ GeV $M_{bb\gamma\gamma} > 350$ GeV	1.51(0.17)
(D) [4]	$p_{T_{1(2)}} > 30(50)$ GeV, $\lvert\eta_{b,\gamma}\rvert < 2.4$ $\Delta R_{b\gamma} > 1.5$, $\Delta R(bb,\gamma\gamma) < 2$ $\lvert M_{bb} - m_h\rvert < 20$ GeV, $\lvert M_{\gamma\gamma} - m_h\rvert < 5$ GeV	1.76(0.27)
ATLAS [5]	$p_{T_\gamma} > 30(30)$ GeV, $p_{T_\gamma} > 40(25)$ GeV, $\lvert\eta_{b,\gamma}\rvert < 2.4$ $\Delta R_{b\gamma} > 0.4$, $\Delta R(bb,\gamma\gamma) < 2$, $p_{T_{bb,\gamma\gamma}} > 110$ GeV $\lvert M_{bb} - m_h\rvert < 25$ GeV, $123 < M_{\gamma\gamma} < 128$ GeV	1.73(0.28)

The sets (A) and (D), from Refs. [1,4], displayed in the first and fourth rows of Table II, respectively, rely on the very distinctive shapes of the ΔR_{bb}, $\Delta R_{\gamma\gamma}$ and $\Delta R_{b\gamma}$ distributions to reduce background events. In plot (c) of Fig. 1, we show the $\Delta R_{\gamma\gamma}$ distribution for the signal and the main backgrounds. For the signal, photons come from the decay of a heavy particle and are more collimated with diminished distance in the (η, ϕ) plane. On the other hand, the photons and b jets from the $b\bar{b}\gamma\gamma$ continuum originate from QED and QCD radiation, respectively, and are thus less collimated. The $t\bar{t}h$ and Zh backgrounds resemble the signal as they contain a Higgs boson. The same occurs for the ΔR_{bb} distribution except for $t\bar{t}h$ as the b jets come from different top decays. The $\Delta R_{b\gamma}$ distribution between the hardest b jet and photon, shown in Fig. 1(d), is more useful to reduce the $t\bar{t}h$ backgrounds since the bottoms from top decays and the radiated photons from them tend to get more collimated.

The sets (B) and (C), from Refs. [2,3], displayed in the second and third rows of Table II, respectively, take advantage of the fact that the signal events feature a harder spectrum of the $b\bar{b}\gamma\gamma$ invariant mass and the transverse momentum of the b jet and photon pair distributions, $p_T(bb)$ and $p_T(\gamma\gamma)$. This is evident from panels (a) and (b) of Fig. 1. We note that these strategies, however, do not reach a higher efficiency compared to those based solely on ΔR distributions. Moreover, the S/B ratio also does not differ significantly. The set (B) is able to reach almost 0.4, but at the expense of accepting more backgrounds which decreases the significance with no systematics compared to the other analyses. This conclusion may be somewhat modified if systematic uncertainties are incorporated.

The set of cuts from ATLAS combines selections across all the theoretical studies, as can be seen from the last row of Table II. A signal significance of $\sim 1.73\sigma$, very similar to that of set (D) from Ref. [4], is obtained.

It is interesting to compare the signal and background yields obtained in our work to those of the ATLAS paper, Ref. [5]. Adopting the cuts of the last row of Table II, we found 11.8 and 41.9 events for signal and backgrounds, respectively, compared to 8.4 for the signal and 47.1 for backgrounds quoted in Ref. [5]. The $S/\sqrt{B}$ significance of the ATLAS paper, with 7.5% systematics in the background rate quoted in that study, is 1.3σ, against 1.6σ from our results. This cross check gives us confidence in our signal and background estimates and reassure the importance of including systematic uncertainties. The discrepancy between our estimates and those of ATLAS may in part be explained by the fact that we do not discard photons which hit the barrel/end-cap transition region, $1.37 < |\eta| < 1.52$, and are totally inclusive in the number of jets accepted. The ATLAS study, on the other hand, included events up to five jets with $p_T > 25$ GeV. These somewhat looser criteria might explain part of the discrepancy between our estimates.

More recently, the ATLAS Collaboration updated the prospects for this channel in Ref. [7] taking pileup effects and some other subdominant backgrounds, such as $Z(\to b\bar{b})\gamma\gamma$ and $t\bar{t}\gamma$, into account. Pileup effects were shown to have moderate influence in the discovery prospects, but the backgrounds were found to be somewhat larger than before. The signal significance is estimated to be approximately 1σ for around 8% systematics in the background rate with the $S/\sqrt{B}$ metrics. The major discrepancy compared to the previous ATLAS study of Ref. [5] and other works is in the number of $b\bar{b}\gamma j$ events, which was estimated to be almost as large as $b\bar{b}\gamma\gamma$ due to an estimated probability for a jet to fake photons that was four times larger than that assumed in previous studies. Since we do not take into account the effect of pileup, we keep comparing our results against those of Ref. [5].

The cut strategy in this new ATLAS study Ref. [7] followed the previous study of Ref. [5] closely. The main difference was a softening of the $p_T(b\bar{b}, \gamma\gamma)$ cut by vetoing events where this variable is less than 80 GeV. The significance obtained after applying these cuts with our extended backgrounds, assuming no systematics and using the AMS metric, is 1.76σ and $S/B = 0.26$. This is very similar to the results of the last row of Table II.

Finally, since the subsequent sections will be devoted to applications of ML algorithms to the question of Higgs pair production, we note that in Ref. [6], a likelihood function-type discriminator was built to better discriminate between signal and background events with a large improvement in the signal significance. In that work, however, an underestimation of backgrounds led to a large significance not confirmed in subsequent analyses.

IV. OPTIMAL SELECTION OF KINEMATIC CUTS

In the previous section, we discussed the analysis performed by several theory groups, as well as an ATLAS study. The summary is provided in Table II, where it is evident that once signal and background cross sections are properly accounted for, the studies are similar in their performances.

The similarity among the performances of Refs. [1–5] shown in Table II suggests that the quest for superior performance in cut and count analyses is largely based on previous results and well known variables proposed in the literature. Sometimes, new variables are found to exhibit good discriminative power, such as the ratio $E_T(\gamma)/M_{\gamma\gamma}$ proposed in Ref. [43]. Of course, there is a lot of variation in the way different groups design their cuts, the extent to which they experiment with old and new variables, and the methods they employ to estimate the boundary of the chosen kinematic variables.

A. Bayesian optimization of kinematic cuts

Given an event topology and a set of kinematic observables, is there a way to systematically obtain the most optimal cuts on the kinematic observables, so as to maximize

the significance? Our purpose in this section is to probe this question, and we shall see that Bayesian optimization offers a pathway.

A typical cut analysis consists in finding a set of kinematic variables thresholds $\{x_k^c, k = 1, \ldots, n\}$ such that the number of signal or background events is given by

$$S, B(x_1^c, \ldots, x_n^c) = L \times \sigma_{S,B}(pp \to X) \times \varepsilon_{\text{eff}} \times \prod_{k=1}^{n} H(\mathcal{O}_k(x_k, x_k^c)) \tag{4.1}$$

where L is the integrated luminosity, $\sigma_{S,B}$ is the signal or background production cross section of X, ε_{eff} a factor that accounts for detection efficiencies, and H is the Heaviside step function. The functions $\mathcal{O}(x, y)$ relate a kinematic variable x and its cut x^c according to one of the following alternatives in this work: $x - x^c$, $x^c - x$, and $|x - M| - x^c$. The goal of our phenomenological analysis is of maximizing a signal significance metric, such as $S/\sqrt{B}$, by retaining the largest possible number of signal events while rejecting the largest amount of background events by finding an optimal set of cuts $\{x_k^c, k = 1, \ldots, n\}$.

When a ML algorithm is trained to better classify the signal and background events, it may be asked to return the probability of a given event to be a signal event. We will call this an *"output score"*. In this way, we can construct distributions of scores for signal and background events and then apply another cut on this distribution. In this case, Eq. (4.1) is modified by multiplying it by another unit step function $H(\mathcal{O}_{\text{ML}}(x_{\text{ML}}, x_{\text{ML}}^c))$. The ML scores x_{ML} may themselves depend on other specific parameters θ_{ML} and must also be adjusted for a good performance. We discuss this in Sec. V.

The most brute force method to obtain the optimal set of cuts, a multivariable scan, is also the one that is the least pragmatic. For example, the ATLAS [5] study makes use of more than 10 kinematic variables. A hypercube in this space with just a tenfold division in each direction represents 10^{10} different cut strategies. To cite another example, one can consider the search for single-top production at the Tevatron [44], which trained neural networks with up to 30 variables that could be used in a cut analysis. It is evident that large grids are unfeasible without large computational facilities.

The situation becomes even more untenable when ML algorithms are used to enhance the collider searches, since they add a much longer time of computation in the analysis chain. A deep neural network, for example, might take from several minutes to several hours to train, depending on the computational resources and the size of the training/testing samples. On the other hand, selection cuts may have a significant effect on the kinematic variables (features) which are used to train ML algorithms. These effects are often neglected but may significantly impact the performance of discrimination tools.

Intuitively, one expects that requiring hard cuts to clean up samples would force one into a small corner of feature space where signal and background events present little distinction. This degrades the ML performance. In other words, hard cuts introduce biases which make signal and background distributions indistinguishable. Loosening the cuts reduces bias, but the gain in performance of the ML discrimination may not compensate for the increased number of background events. This, too, may lead to a degraded performance, especially when systematic uncertainties are taken into account.

The maximum significance achievable must, therefore, be a trade-off between cuts on the kinematic variables and ML performance. We note that ML classification can be performed in two ways: (1) by generating a new distribution with the ML output classification ranking of signal and background events, where a good discriminator should give the majority of signal (backgrounds) events a score close to 1(0), for example, and subsequently using this distribution to place another cut as discussed above, and (2) using the output distributions in a multivariate statistical analysis (MVA) based on the likelihood ratio statistic for the final discrimination.

The solution to avoid expensive grid searches can be found in the data science literature itself. The most powerful ML algorithms, such as neural networks and decision trees, have a large number of parameters (called hyperparameters) which control their performance. Adjusting hyperparameters to achieve a high classification accuracy is an important goal in ML, and avoiding extensive scans in the space of hyperparameters is desirable. It is now common practice to perform either randomized grid searches or use dedicated algorithms for model configuration [14]. Surprisingly, a simple random search with hundreds of trials may perform as good as, or even better than, a manual search.

For large parameter spaces, however, it has been demonstrated that Bayesian optimization performs better than either manual or randomized searches [15]. The algorithm described in Ref. [15], implemented in the Python library `Hyperopt` [16], is based on the so-called sequential model-based optimization (SMBO) technique [45]. This class of algorithms suggests a new model (a new configuration of parameters) at each iteration in order to optimize the criterion of expected improvement (EI), which is the expectation that under a model M of a function f, $y = f(x)$ will exceed some threshold y^c

$$EI_{y^c}(x) = \int_{-\infty}^{+\infty} \max(y^c - y, 0) p_M(y|x) dy \tag{4.2}$$

in the search for the minimum of f.

The major obstacle in computing $EI(x)$ is estimating the conditional probability $p_M(y|x)$. `Hyperopt` overcomes this difficulty by means of the Bayes rule, $p_M(y|x) = \frac{p(x|y)p(y)}{p(x)}$,

where $p(x)$ is an assumed prior distribution of the parameters. By keeping a sorted list of observations of $y = f(x)$, it is possible to compute the quantiles $\gamma = p(y < y^c)$, while $p(x|y)$ is a nonparametric distribution estimated from previous observations along the run of the algorithm. The strategy to evaluate $p(x|y)$ in `Hyperopt` is known as a tree-structured Parzen estimator approach, TPE for short. In TPE, $p(x|y)$ equals $\ell(x)(g(x))$ if $y < y^c (y \geq y^c)$, thus providing an nonparametric estimate of $p(x|y)$ from previous runs of the algorithm. Further details of the algorithm can be found in Ref. [15] and references therein.

This way, it is possible to show that $EI_{y^c}(x)$ is such that

$$EI_{y^c}(x) \propto \left(\gamma + \frac{g(x)}{\ell(x)}(1-\gamma)\right)^{-1} \tag{4.3}$$

where, on each iteration, the algorithm returns the point on the parameters space x^c with greatest expectation improvement. The algorithm is efficient once $EI_{y^c}(x)$ grows as the ratio $g(x)/\ell(x)$ drops, that is, as $\ell(x)$ accumulates with the learning process and $g(x)$ represents more rare configurations.

The main result of this section is to use Bayesian optimization to look for better discriminating kinematic cuts. In this case, x is a point in a kinematic multivariable space designed to discriminate between signal and backgrounds and $f(x)$ is an approximated mean significance (AMS) function, a significance metric as defined in Eqs. (A1), (A2), (A3). In Sec. V C, we will investigate an augmented searching space comprising the thresholds of the kinematic variables for cuts and the hyperparameters which models a boosted decision trees algorithm, thus performing a joint cuts plus hyperparameters search.

B. Results using Bayesian optimization in `Hyperopt`

We use `Hyperopt` [16] for the search with the TPE strategy described above. The inputs of the program are a Python dictionary with the names and variation ranges of the variables, the prior random distributions assumed for those variables, the objective function to be minimized, and the number of experiments which the algorithm is allowed to perform in the search, that is, the number of trials. The algorithm can be easily parallelized as described in Ref. [15], but our searches were all obtained within a single thread of the computer, thus the running time of cut searches could be greatly reduced. In Appendix B, we display a simple code that can be adapted by the reader for immediate use in a cut-and-count analysis.

In Table III, we show the kinematic variables used for cut optimization and their ranges of variation. For all of them, we assume uniform priors. The corresponding number of points in such a grid would be staggering 1.86368×10^{14} possible cut strategies.

Compared to the variables of Table II, we also experimented with the invariant mass of the hardest b jet and photon, $M_{b_1\gamma_1}$. We required the same ΔR cut for b jets and photons pairs and for all $b\gamma$ combinations according to the first two rows of Table III. We also put the same cut on the transverse momentum of the hardest(second hardest) photon and bottom. Of course, we could have chosen different cuts for each particle p_T and ΔR pair. The rapidity cuts are kept constant throughout the experiments, $|\eta| < 2.4$ for all photons and jets.

TABLE III. The kinematic variables used for cuts and their allowed variation ranges in `Hyperopt`. The prior distributions for all these variables are set to uniform distributions over the ranges shown in the table within the steps shown as the last entry of each vector.

Kinematic variable	Variation range in `Hyperopt`
$\Delta R_{ii} <$	(1.4,4,0.05)
$\Delta R_{ij} >$	(0,2,0.05)
$p_T(1) >$	(30,100,1) GeV
$p_T(2) >$	(20,70,1) GeV
$p_{T_{ii}} >$	(0,200,5) GeV
$M_{b\bar{b}\gamma\gamma} >$	(0,400,5) GeV
$M_{b_1\gamma_1} >$	(0,200,5) GeV
$\lvert M_{\gamma\gamma} - m_h\rvert <$	(5,15,1) GeV
$\lvert M_{bb} - m_h\rvert <$	(10,30,1) GeV

In Table IV, we show the set of cuts that achieves the largest significance in a cut-and-count analysis found with the Bayesian search after 200 trials. The first row shows the optimized cuts and the significance, computed with $S/\sqrt{B}$, reached for the same backgrounds of Ref. [4]. In the second row, we show the results for the extended backgrounds including $c\bar{c}\gamma\gamma$, $b\bar{b}\gamma j$ and $c\bar{c}\gamma j$ events. The last row displays the results for the cuts of Azatov *et al.*, Ref. [4]; the upper subrow is the significance computed with the same backgrounds of that work, while the lower subrow contains the $S/\sqrt{B}$ with the extended set of backgrounds considered in our work.

First of all we note that the learning process selects somewhat different sets depending on the actual size of the backgrounds. The Best (1) strategy of the first row, with smaller backgrounds, relied mainly on the $M_{b\bar{b}\gamma\gamma}$ and $p_{T_{\gamma\gamma}}$ variables to eliminate backgrounds. The Best (2) set of the second row, for extended backgrounds, put a stronger cut on the ΔR_{ii} compared to the Best (1) set, while the other cuts remained more or less the same. This confirms that the ΔR_{ii} variables are indeed discriminative. Second, both strategies found better discrimination putting cuts on $M_{b\bar{b}\gamma\gamma}$ and $p_{T_{bb,\gamma\gamma}}$ which also confirms the usefulness of these variables. Third, we observe that the optimized sets relax the p_T cuts on the softer b's and photons whereas strengthening the cut on the hardest particles. As in previous studies, the window around the $b\bar{b}$ peak is wider than the $\gamma\gamma$ peak. Finally, ΔR_{ij} and $M_{b_1\gamma_1}$ were found to be less important in the discrimination as observed in Table IV.

TABLE IV. The rows show a set of cuts at the left column, and the number of signal and backgrounds after these cuts and the significance(signal-to-background ratio) in the subsequent columns. The first row shows the results reached for the same backgrounds as Ref. [4] after 200 Hyperopt trials. In the second row, we show the results for the extended backgrounds including $c\bar{c}\gamma\gamma$, $b\bar{b}\gamma j$ and $c\bar{c}\gamma j$ events again with 200 Bayesian searches. The last row displays the results for the cuts of Azatov *et al.*, Ref. [4]; the upper numbers in red in this row are computed with the same backgrounds of that work, while the lower ones in blue contains are computed with the extended set of backgrounds considered in our work.

Kinematic cuts	N_S	N_B	$S/\sqrt{B}(\sigma)$ (S/B)
Best (1): $p_T(1) > 90$ GeV, $p_T(2) > 21$ GeV $\Delta R_{ij} > 0.65$, $\Delta R_{ii} < 3.75$ $M_{b\bar{b}\gamma\gamma} > 385$ GeV, $p_{T_{ii}} > 100$ GeV, $M_{b_1\gamma_1} > 60$ GeV $\lvert M_{bb} - m_h\rvert < 24$ GeV, $\lvert M_{\gamma\gamma} - m_h\rvert < 7$ GeV	21.0	55.1	2.81(0.38)
Best (2): $p_T(1) > 86$ GeV, $p_T(2) > 22$ GeV $\Delta R_{ij} > 0.4$, $\Delta R_{ii} < 1.85$ $M_{b\bar{b}\gamma\gamma} > 390$ GeV, $p_{T_{ii}} > 100$ GeV, $M_{b_1\gamma_1} > 25$ GeV $\lvert M_{bb} - m_h\rvert < 24$ GeV, $\lvert M_{\gamma\gamma} - m_h\rvert < 8$ GeV	18.0	52.1	2.48(0.35)
Default: $p_T(1) > 30$ GeV, $p_T(2) > 50$ GeV $\Delta R_{ij} > 1.5$, $\Delta R_{ii} < 2$ $\lvert M_{bb} - m_h\rvert < 20$ GeV, $\lvert M_{\gamma\gamma} - m_h\rvert < 5$ GeV	12.8 12.8	37.1 48.7	2.1(0.34) 1.85(0.27)

We now investigate how often Hyperopt finds cuts with higher significances compared to the cuts of Ref. [4] and with the same background assumptions of that work. For this investigation we performed 500 trials and created histograms for the number of sets cuts in a given $S/\sqrt{B}$ interval as shown at the left plot of Fig. 2. The blue(red) [green] histogram displays the number of sets for a given AMS interval after 100(300)[500] trials.

Around 90% of all Bayesian optimization searches yielded a greater significance than the 2.1σ achieved by the cuts of Azatov *et al.*, represented by the dashed line at the left plot of Fig. 2. The 300 and 500 trials histograms also make evident the way the algorithm improves the objective function, $S/\sqrt{B}$ in this case. The bins of higher significances get more populated as we increase the number of trials indicating that the algorithm learns with past cut-and-count experiments in order to search for better ones as expected. This is no surprise, since the Bayesian optimization is actually a generative machine learning algorithm as described in the previous section.

In the inset frame of the left plot of Fig. 2, we show $S/\sqrt{B}$ as a function of the number of trials. We see that after 100–200 trials, the signal significance does not change much up to 500 trials. After 200 trials, the optimized cuts achieved a significance of 2.81σ against 2.1σ of the manual search of Ref. [4], a 34% improvement. With extended backgrounds,

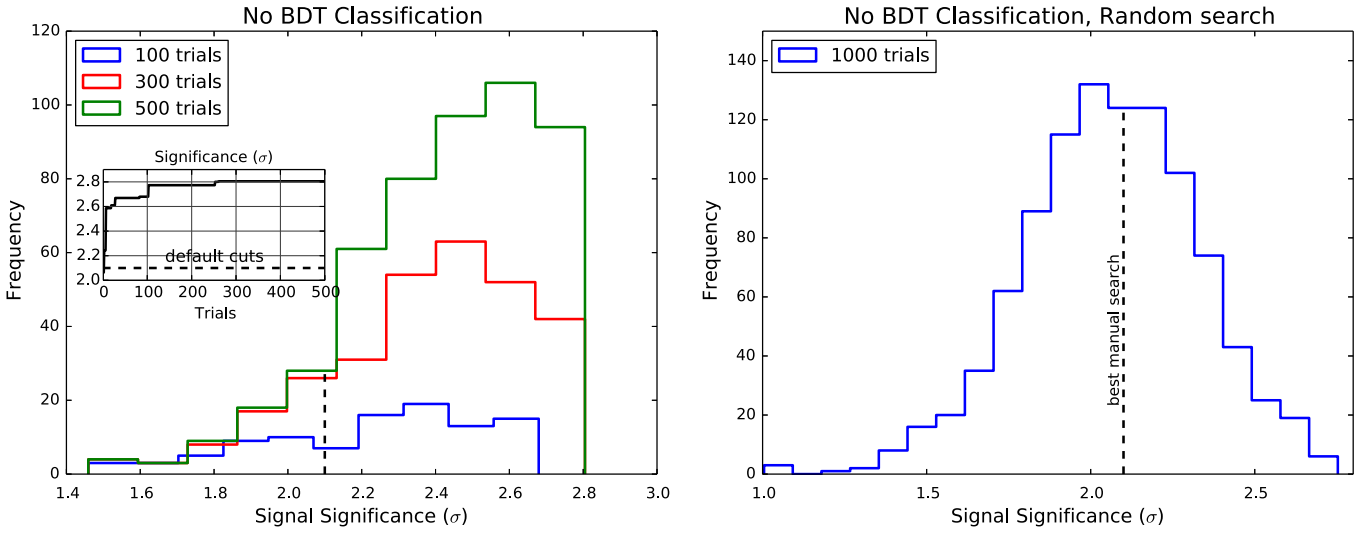

FIG. 2. The histograms of number of cut strategies producing a given significance interval in a cut-and-count analysis. At the left plot we show the optimized search with the TPE algorithm in Hyperopt. The inset frame in the left plot shows the significance as a function of the number of trials. At the right plot we display a nonoptimized random search after 1000 trials. No systematics are assumed, the backgrounds are those of Ref. [4] and the $S/\sqrt{B}$ used to compute the signal significances. The black dashed line represents the results obtained with the default cuts of Azatov *et al.*, Ref. [4] in all plots.

the Bayesian search reached 2.48σ against 1.85σ of the cuts of Ref. [4], again roughly the same improvement. A larger S/B was also achieved as shown in Table IV.

We also point out that the previous works of Refs. [46–48] approached in different ways the optimum locus of the variables space for better discernment between signal and backgrounds. Similarly to our findings, those works also highlight the relative importance of the $M_{b\bar{b}\gamma\gamma}$, $p_{T_{\gamma\gamma}}$, $p_{T_{bb}}$, and $\Delta_{\gamma\gamma,bb}$ variables.

C. Reliability of the Bayesian search

In order to probe the reliability of the Bayesian approach, we performed an exhaustive grid search in a reduced variables space. We choose the 4-dimensional $(\Delta R_{ii}, \Delta R_{ij}, M_{b\bar{b}\gamma\gamma}, p_T(1))$ space with ten evenly spaced values in each direction amounting to 10^4 different sets of cuts. We allowed `Hyperopt` to carry out up to 300 TPE trials. Both ΔR ranges were chosen to lie in (1,3,0.2), the $b\bar{b}\gamma\gamma$ invariant mass, (300,600,30) GeV, and the $p_T(1)$ variable, (20,70,5) GeV.

The maximum $S/\sqrt{B}$ found were

$$\begin{aligned}
&\textbf{Grid search}: 2.11\sigma, \quad \Delta R_{ii} < 1.6, \quad \Delta R_{ij} > 1.0, \\
&\qquad M_{b\bar{b}\gamma\gamma} > 390\,\text{GeV}, \quad p_T(1) > 25\,\text{GeV} \\
&\textbf{Optimized search}: 2.06\sigma, \quad \Delta R_{ii} < 1.6, \\
&\qquad \Delta R_{ij} > 1.8, \quad M_{b\bar{b}\gamma\gamma} > 390\,\text{GeV}, \\
&\qquad p_T(1) > 25\,\text{GeV} \qquad (4.4)
\end{aligned}$$

The only different cut was in the less discriminative variable ΔR_{ij}, for the all the other ones, Bayesian optimization was able to find the same cut thresholds of the Grid search. Of course, in a much larger searching space it is hard to tell how close to the best grid point the Bayesian optimization gets, but our results show that the cut strategies found with hundreds of trials improve significantly the statistical significances compared to the manual searches of Table II. We also point out that other open source algorithm optimization programs are available [49] for experimentation.

D. Random versus manual search

In phenomenological analyses, one frequently tunes the cut thresholds by visually estimating the regions of variable space which are more populated by signal or background events. Sometimes, after a first round of requirements, one looks for more discriminative variables to apply cuts on. The entire process, however, is not optimized. The similar results found by manual searches of this nature, for the cut thresholds displayed in Table II, suggest that the majority of cut strategies should indeed perform nearly identically by this method.

Another strategy to avoid large grid scans is simply performing a random search for cuts. As we discussed in the previous section, this approach presents good results in the search for ML hyperparameters according to Ref. [14]. In order to investigate how the manual strategies compare to a random search, we allowed for 1000 trials in `Hyperopt`, running in the random mode (see appendix B for more details), in the variables region of Table III. We then computed $S/\sqrt{B}$ for each set of cuts without systematics and with the same backgrounds of Ref. [4]. The search lasted around 20 minutes with a single thread.

In the right plot of Fig. 2, we show the histogram of the number of cut strategies for a given significance interval in this random search. The vertical dashed line is the significance of 2.1σ reached by the best manual search of Ref. [4]. The mean of the distribution is 2.06σ with 0.27 standard deviation. Around 45% of all cut strategies result in a signal significance larger than 2.1σ. In other words, a good manual search is likely to reproduce just the mean performance of a random search when we look for a promising region of the variables space for cut-and-count. We suspect that similar behavior can be observed in other phenomenological analysis based on cut-and-count.

As observed in Ref. [14], the Bayesian search performed slightly better than the random search in our case too. However, while a thousand experiments were necessary to reach an $\sim 2.7\sigma$ of significance in the random search, with just 200 trials is possible to reach around 2.8σ as we see in Fig. 2. Both searches, however, present an enhancement compared to the manual searches of Table II.

We now investigate how the Bayesian cut optimization works when systematic uncertainties are present.

E. Optimization with systematic uncertainties

As we observed in the previous section, the optimization procedure is able of not just increasing the signal significance but also the S/B ratio which is essential when we take systematic uncertainties into account in the statistical analysis. This observation leads us to investigate whether `Hyperopt` would also be able to find cuts with higher S/B in order to tame the systematics.

In Fig. 3, we show the signal significance in terms of the background rate systematic uncertainty ε_B from 0% to 20% after 100 trials. The red solid line represents the significance for the default cuts of Azatov *et al.* The points of the black dashed line are obtained by optimizing only the cuts of the 0% case and then using $S/\sqrt{B + (\varepsilon_B B)^2}$ to extrapolate the significance for other ε_B, keeping the same set of cuts found in the no systematics case. This is not the best that can be done, though, as the S/B ratio remains the same as in the no systematics scenario. The upper black solid line shows the results when we optimize the significance function for each systematics level. In this case, the Bayesian algorithm is able to find points with larger S/B ratio trying to overcome the systematics constraints. The inset plot shows that `Hyperopt` learned that S/B

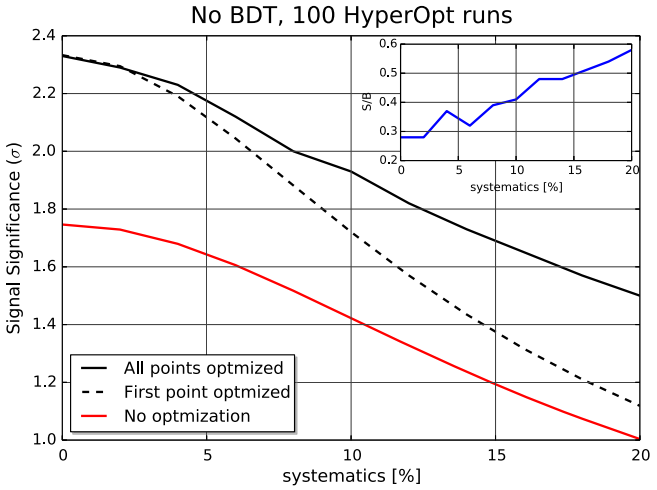

FIG. 3. The $S/\sqrt{B+(\epsilon_B B)^2}$ significance metric as function of ϵ_B, the systematic uncertainty in the total background rate. The red line represents the default cuts of Azatov *et al.*, Ref. [4], the black dashed assumes an optimized strategy just for the 0% systematics point, while for the solid upper line, the algorithm was solicited to learn the best cuts for each systematics level from 0% to 20%. In the inner plot, we show the S/B ratio for the point-to-point optimization case.

should double from the 0% to the 20% systematics case to reach larger significances.

As a consequence of larger S/B ratios, the difference between the point-to-point optimized significance and the 0% optimized curves gets larger as the systematics increase. It is also interesting to observe how the algorithm learns to increase the signal-to-background ratio and reach high significances as the systematics get more important. For that aim we show in Fig. 4 the cut thresholds of some key variables used in the analysis with systematics up to 30%.

Some clear tendencies are noticeable: the preferred variables to hardening cuts are ΔR_{ii}, the window around the bb and $\gamma\gamma$ mass peaks, especially this last one, and the transverse momentum of the softer particles shown in panels (c), (f) and (d) of Fig. 4, respectively. On the other hand, ΔR_{ij} and the transverse momentum of the harder particles become less relevant. We already knew that ΔR_{ij} is not so important for the discrimination as the other variables. The softening of the cut of $p_T(\text{hard})$, however, can be understood in view that we are not trying to optimize S/B but the significance metric, and the algorithm seems to find a way through the second hardest p_T cut instead. Despite being more erratic, a tendency to irrelevance is also observed in other discriminants like $M_{b\bar{b}\gamma\gamma}$, $M_{b_1\gamma_1}$ and $p_{T_{\gamma\gamma}}$, for example, as seen in panel (e) of Fig. 4. This can be explained in view of panels (a) and (b) which show the correlation between two of the most discriminative variables, $\Delta R_{\gamma\gamma}$ and $M_{b\bar{b}\gamma\gamma}$. A hard cut on $\Delta R_{\gamma\gamma}$ makes a cut on $M_{b\bar{b}\gamma\gamma}$ somewhat irrelevant and vice-versa. Of course, this does not mean that this is the only way to increase S/B, but it does suggest that not all the kinematic variables are relevant for that task at the same time.

We especially note that for the level of background rate systematics estimated by the ATLAS and CMS Collaborations, around 10% [5,7,43], the optimized cuts give a significance of 1.9σ against $\sim 1.4\sigma$ of the default cuts of Azatov *et al.*, all with the extended backgrounds.

V. SIGNAL VERSUS BACKGROUND DISCRIMINATION WITH BOOSTED DECISION TREES

The analysis presented in the previous section has been based solely on cut-and-count and can be employed in any phenomenological study where optimal cuts are necessary to clean up backgrounds and raise the signal significance. In Appendix B, we give more details about implementing this procedure in a simple and fast Python code.

In this section, we go beyond the cut-and-count analysis and focus exclusively on proposing tools to obtain even larger significances in the search for double Higgs production at the LHC, with and without systematics. Our goal now is to show that training a Boosted Decision Tree (BDT) algorithm to better classify signal and backgrounds events, in addition to the procedure of using optimal cuts to select the best volume of the features space for the BDT training, increases the signal significance dramatically.

We present our results in three stages. In Sec. V A, we first introduce the kinematic observables used in the BDT analysis, and provide a discussion of the interplay between BDT classifiers and cut selections, without addressing the question of cut optimization. In Sec. V B, we sequentially optimize the cuts on the kinematic observables using `Hyperopt`, and then optimize the BDT hyperparameters. Finally, in Sec. V C, we perform a joint optimization of the kinematic cuts and the BDT hyperparameters.

A. BDT analysis without cut optimization

The performance of any ML algorithm aimed to better classify signal and background events, or even an MVA analysis based on likelihood ratios, depends strongly on the portion of the feature space from which the events are selected, in other words, the number of signal and background are as follows,

$$N_{ev} = N_{ev}(\{x_k^c, k=1,\ldots,n_c\} \cup \{x_{\text{ML}}^c(\theta_{\text{ML}}, \{x_k^c, k=1,\ldots,n_c\})\}, \quad (5.1)$$

where θ_{ML} represents the hyperparameters of the ML algorithm and $N_{ev} = S(B)$ is the number of signal(total background) events. This is especially true in subtle searches for new physics, and is the reason we have investigated the Bayesian optimization method thoroughly in the previous section.

Ideally, the least biasing portion of any variables space is the one with minimal cuts, possibly requiring just acceptance and trigger cuts. However, in processes with low

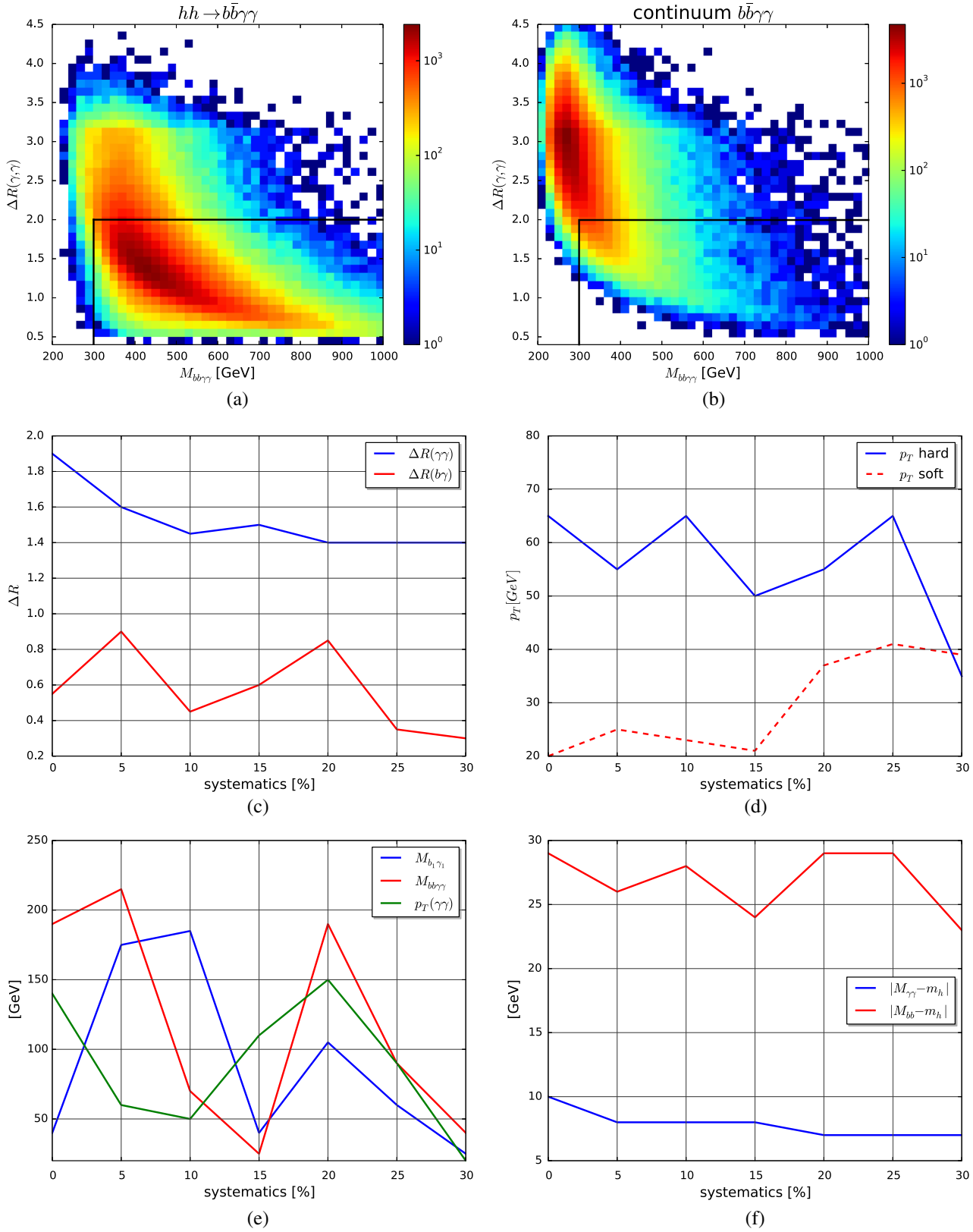

FIG. 4. The learning process evolution of the cut variables in the search for maximum significance in the presence of increasing systematic uncertainties are displayed in panels (c-f). Panels (a) and (b), show the correlation between two of the most discriminative kinematic variables, $\Delta R_{\gamma\gamma}$ and $M_{b\bar{b}\gamma\gamma}$ for the signal and the dominant $b\bar{b}\gamma\gamma$ background, respectively. In panel (c), we show the distance of pairs of particles in the (η, ϕ) plane. Panel (d) displays the transverse momentum of the hardest and second hardest b's and photons. In panels (e) and (f), various invariant mass combinations used in the discrimination plus the transverse momentum of the pair of photons.

signals and large backgrounds like $pp \to b\bar{b}\gamma\gamma$, if one employs just acceptance cuts, detection efficiencies and even takes b-tagging into account, one is still presented with backgrounds that are many orders of magnitude larger than the signal. This would require a ML classifier with an extremely exquisite signal acceptance versus background rejection performance, which cannot be reached in practice. On the other hand, applying harder cuts may not necessarily degrade the ML performance to the point of making them useless for further discrimination.

Therefore, a trade-off between cuts and ML performance should be expected in a phenomenological analysis. We now proceed to study this interplay.

We use the `XGBoost` [17] implementation of BDTs for Python for its very good discrimination performance, speed and capacity of parallelization. The events features used to train the BDT are as follows:

(1) transverse momentum of the two hardest b jets and photons: $p_T(b_1, b_2)$ and $p_T(\gamma_1, \gamma_2)$
(2) transverse momentum of $b\bar{b}$ and $\gamma\gamma$ pairs: $p_T(bb)$ and $p_T(\gamma\gamma)$
(3) invariant mass of all four combinations of a b jet and a photon: $M_{b_i\gamma_j}, (i, j) = 1, 2$
(4) invariant mass of the two b jets and two photons of the event: $M_{b\bar{b}\gamma\gamma}$
(5) distance between pairs of bottoms and photons: $\Delta R(bb)$, $\Delta R(\gamma\gamma)$ and all the four combinations of a b jet and a photon $\Delta R(b_i\gamma_j), (i, j) = 1, 2$
(6) the Barr variable [50,51] between all the six combinations of two particles in the event defined as $\cos\theta^*_{ij} = \tanh(\frac{\Delta\eta_{ij}}{2})$ where $\Delta\eta_{ij}$ is the rapidity separation of the i and j particles

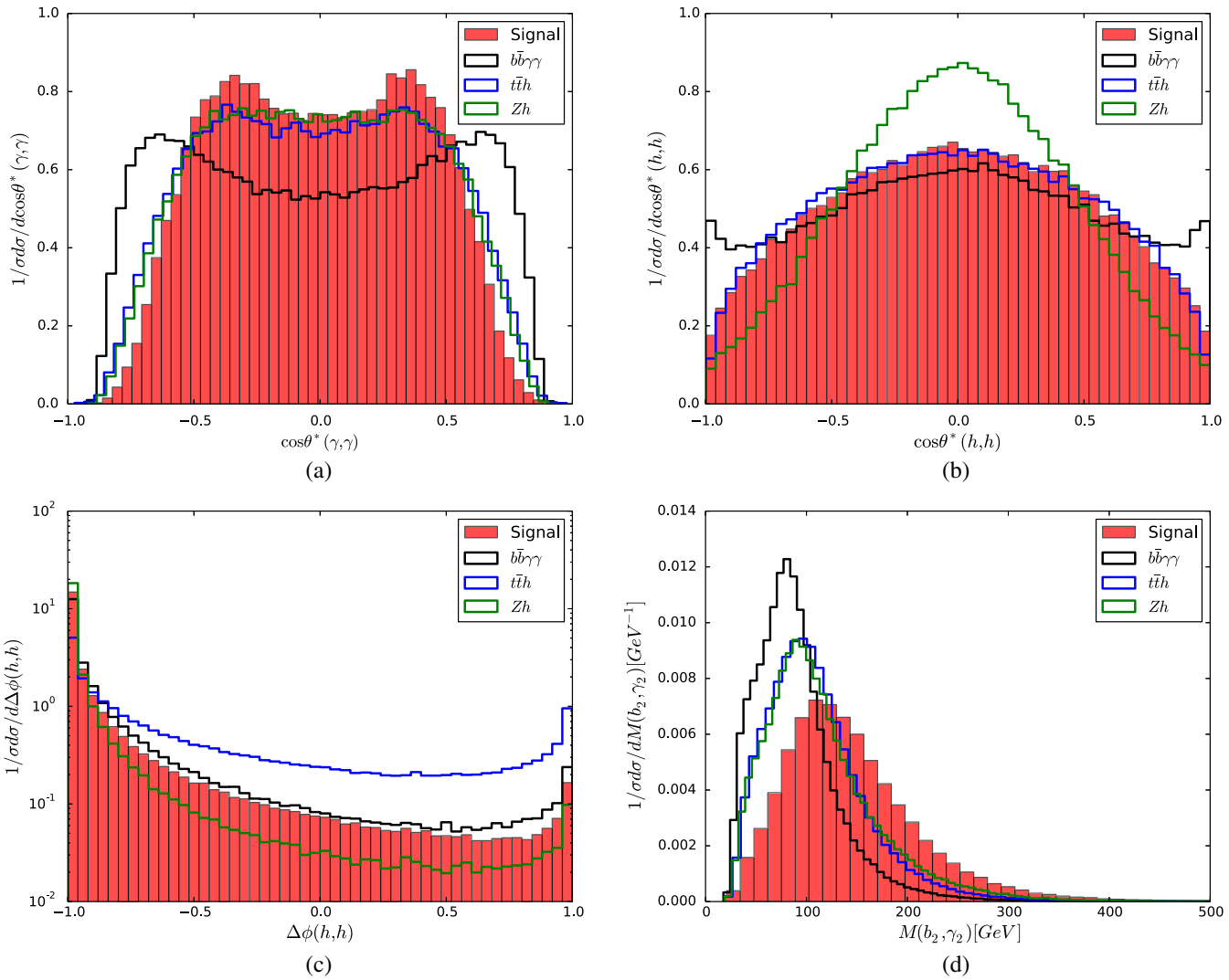

FIG. 5. Four out of the 27 kinematic distributions of signal (shaded red), and the backgrounds $b\bar{b}\gamma\gamma$ (black), $t\bar{t}h$ (blue) and Zh (green), used for the BDT discrimination. In (a), we show the Barr variable of the two photons (see the text for its description). Plots (b) and (c) display the Barr variable and the difference of the azimuthal angle of the reconstructed Higgs pair, respectively. In panel (d), the invariant mass of the second hardest b jet and photon.

(7) the Barr variable between the two reconstructed Higgs bosons, $\cos\theta^*_{hh}$
(8) azimuthal angle difference between the two reconstructed Higgs bosons, $\Delta\phi(h,h)$
(9) missing energy of the event
(10) the number of charged leptons with $p_T > 20$ GeV and $|\eta| < 2.5$

These are 27 features in total. We do not use all of them for kinematic cuts; just those shown in the first and second rows of Table III. The missing energy and the number of charged leptons are used to better distinguish the multijet backgrounds and semileptonic $t\bar{t}h$ backgrounds. In Fig. 5, we show some other good features besides the ones shown in Fig. 1. We simulated ∼240000 signal and ∼640000 background events to train, test and cross-validate the BDTs. After optimized cuts we observed that the number of Monte Carlo samples of signal and background events get much more balanced.

We preprocess the features prior to the BDT training which improves their performances. First, to the distributions with skewness larger than 1.0 we add a small value of 10^{-8}, the logarithm is taken and then they are normalized as in Ref. [52]. All the features are rescaled to smaller and standardized ranges better suited for the training process.

The behavior of the statistical significance in terms of the output scores may sometimes oscillate very badly if the number of test samples is small as a consequence of not too smooth signal and background scores distributions [53]. We checked that the AMS function, in terms of the score cut threshold for one of the five evaluations of the BDT in

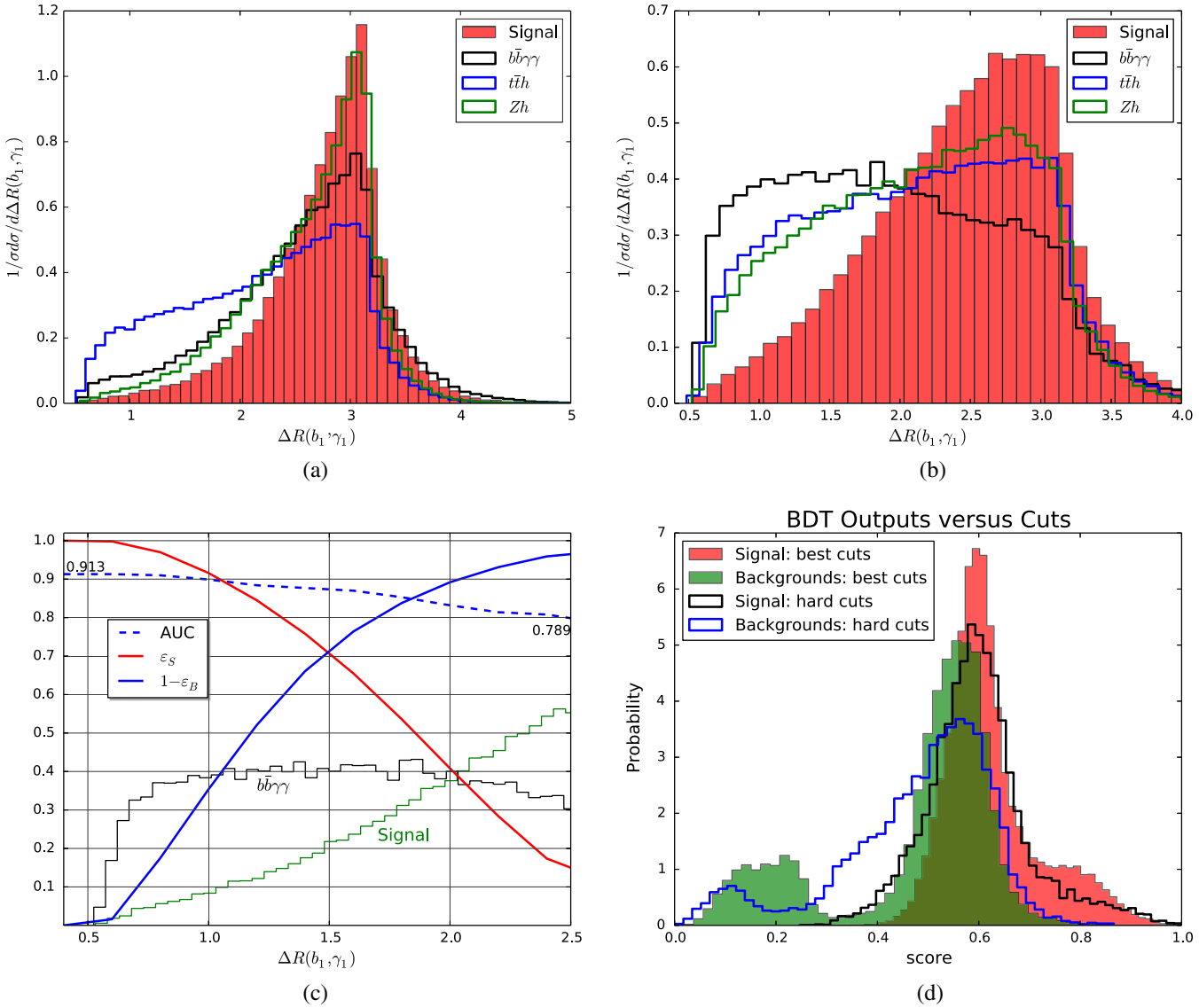

FIG. 6. Panels (a) and (b) show the $\Delta R_{b_1\gamma_1}$ distributions of signal and backgrounds requiring the acceptance (default) cuts of Eq. (2.2) (Azatov *et al.*, Ref. [4], last row of Table II). In panel (c), we present the results of the effects of cutting on $\Delta R_{b_1\gamma_1}$ for the BDT performance, see the text for further details. The output scores of the BDT are shown in panel (d) for signal and backgrounds for the optimized set of cuts and a hard set of cuts.

the fivefold cross validation for an optimized set of cuts, is very smooth and well behaved. The maximum AMS, in this case, occurs for scores cut around 0.5. For all the cut strategies with BDTs, the threshold score is chosen in order to achieve the maximum significance.

We next go on to an investigation of how the cuts affect the discrimination power of the BDT. We fixed the set of cuts as the default cuts of Azatov *et al.* shown in Table II, except the ΔR_{ij} variables. Panels (a) and (b) of Fig. 6 show the $\Delta R_{b_1\gamma_1}$ distribution with just the acceptance cuts of Eq. (2.2) and after imposing the default cuts of Azatov *et al.*, respectively. Interestingly, the cuts seem to make the distributions more distinctive in this case; contrary to intuition, therefore, the cuts may help the ML classification in some cases.

Panel (c) of Fig. 6 shows the normalized $\Delta R_{b_1\gamma_1}$ histograms for the signal and the $b\bar{b}\gamma\gamma$ continuum background, the signal efficiency(background rejection) is the red(blue) line, and the area under the Receiver-Operator curve (ROC), AUC, is the dashed line. The bigger the AUC, the better the performance of a cut-and-count analysis based on that distribution. To eliminate backgrounds we should demand that an event has a large $\Delta R_{b_1\gamma_1}$. The effect of hardening this cut is that the total background rejection increases and the signal efficiency decreases as expected. For example, requiring $\Delta R_{b_1\gamma_1} > 1.5$, as in the default cuts of Azatov *et al.* almost exactly rejects 70% of backgrounds at the same time that it retains 70% of signal. However, as the cuts get harder the AUC drops from 0.913 to 0.789 as we see from the dashed line. It is common that tiny increments in AUC represent a significant increase in the significance, thus the magnitude of difference in AUC in this case represents a large decrease in the ML performance.

The BDT scores distributions for signal and backgrounds are shown in panel (d) of F 6. We chose to place harder cuts to make the signal and backgrounds scores distributions more similar. In fact, the hollow histograms of events with hard cuts overlap more noticeably than the scores of the best set of cuts found with the Bayesian optimization. We note, especially, that the green shaded histogram of backgrounds with best cuts presents a more pronounced hill on the left compared to the hollow blue histogram showing some degradation. This isolated left hill is populated mainly by the reducible backgrounds with charged leptons and missing energy. The signal histograms also show marked differences, especially the right hill of best cuts which disappears from the hard cuts of the red hollow histogram.

We next turn to a discussion of the results for the BDT analysis with optimized cuts.

B. Sequential search for optimal cuts and BDT hyperparameters

In this section, we study how best to perform an optimization of the cut analysis and the selection of BDT hyperparameters, in a sequential manner.

The necessity of tuning BDT hyperparameters before optimizing the cuts arises from the need to avoid overfitting and underfitting. This used to be a costly part of a ML analysis. Beside keeping the complexity of the algorithm under control to achieve a good generalization performance, an efficient way to avoid overfitting is to use a large number of training samples whenever possible. For our ML analysis we simulated ∼880000 events as discussed in the previous section. Depending on the cuts, however, the total number of events usually drops to around 100000–300000 events which also turned out to be a sufficient number of samples to keep overfitting under control.

Our first approach was to apply the default cuts of Azatov *et al.* and run 500 `Hyperopt` trials in the space of the chosen hyperparameters of `XGBoost` in the search for the highest AUC over 1/3 of the total samples, the other 2/3 were used for a fivefold cross validation by randomly splitting the remaining samples in the 2:1 proportion for training and testing the BDT, respectively. The hyperparameters chosen were the `number of boosted trees`, from 100 to 500, the `learning rate` from 0.001 to 0.5, the `maximum depth of the trees`, from 2 to 15 final leaves, and the minimum sum of instance weight needed in a child to continue the splitting process of the tress, `min_child_weight`, from 1 to 6. Once we found the best hyperparameters, we then checked the learning curves of the algorithm, as the classification error and the log-loss, to confirm that it generalizes well from the training to the testing samples.

From this initial tuning, we fixed

$$\begin{aligned} \texttt{number of boosted trees} &= 200, \\ \texttt{learning rate} &= 0.1 \\ \texttt{maximum depth of the trees} &= 6, \\ \texttt{min child weight} &= 1. \end{aligned} \tag{5.2}$$

Hyperparameters like the `number of boosted trees`, `maximum depth of the trees` and the `min_child_weight` are directly related to the complexity of the algorithm by controlling the number, size and configuration of the trees. The `learning rate`, also known as `shrinkage` in this context, is a parameter that controls the weight new trees have to further model the data. A large value permits a larger effect from new added trees and might lead to more severe overfitting. There are other parameters which can be eventually used to prevent overfitting and loss of generalization power as explained in Refs. [17,54], but we found that tuning these parameters was sufficient to achieve a good performance.

In principle, it would also be possible to tune the BDT for each set of cuts. But that would be computationally expensive. As we show going forward, keeping these parameters fixed already leads to very good results in terms of signal significance.

In Fig. 7, we repeat the analysis presented in Fig. 2, but now after performing the BDT classification. The black

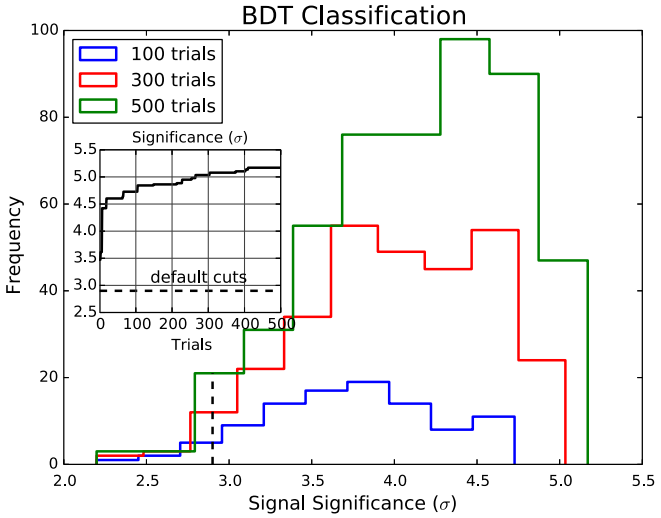

FIG. 7. The histogram of number of cut strategies producing a given significance interval in a BDT-aided classification analysis. The inset plot shows the significance as a function of the number of Hyperopt trials. No systematics are assumed, the backgrounds are those of Ref. [4] and the $S/\sqrt{B}$ used to compute the signal significances. The black dashed line represents the results obtained with the default cuts of Azatov *et al.*, Ref. [4].

dashed line is the maximum signal significance encountered by cutting on the BDT output scores distributions of signal and backgrounds with no systematics using the $S/\sqrt{B}$ metric for the default cuts of Azatov *et al.* In this case, we take only the backgrounds of the Ref. [4] for the comparison.

As in the case of the cut-and-count analysis with no BDT classification, we used Hyperopt to search for the best cuts in 500 experiments with the same kinematic variables and ranges of Table III. Again, around 90% of all cut strategies produced a signal significance larger than the default cuts of Azatov *et al.* With 200 experiments, the maximum AMS found with the optimized search was 4.9σ and an AUC of 0.904, whereas for the default cuts the maximum AMS significance is 2.9σ with an AUC of 0.869. The best set of cuts found with 200 experiments was

$$\begin{aligned} &p_T(1) > 52\ \text{GeV}, \quad p_T(2) > 22\ \text{GeV} \\ &\Delta R_{ij} > 0.1, \quad \Delta R_{ii} < 2.75 \\ &M_{b\bar{b}\gamma\gamma} > 340\ \text{GeV}, \quad p_{T_{ii}} > 145\ \text{GeV}, \\ &M_{b_1\gamma_1} > 45\ \text{GeV} \quad |M_{bb} - m_h| < 26\ \text{GeV}, \\ &|M_{\gamma\gamma} - m_h| < 10\ \text{GeV} \end{aligned} \tag{5.3}$$

Including the extended backgrounds, again after 200 trials, Hyperopt found a significance of 4.5σ, AUC of 0.910, and for the default cuts of Azatov *et al.*, an AMS of 2.6σ and AUC of 0.869. In this case, the Bayesian optimization algorithm found another way into the variables space

TABLE V. Comparison of the performance of the BDT implementation in XGBoost trained with samples selected with the cuts various previous works in the literature and with the optimized set of cuts. The second column contains the maximum AMS obtained by cutting on the BDT outputs after a fivefold cross validation. The last column displays the AUC metric of the BDT for each set of cuts.

Reference	Max AMS(σ) with BDT	AUC
(A) [1]	2.36	0.884
(B) [2]	1.96	0.885
(C) [3]	2.43	0.885
(D) [4]	2.65	0.870
ATLAS [5]	2.67	0.883
Our work (with Hyperopt)	**3.88**	**0.901**

$$\begin{aligned} &p_T(1) > 52\,\text{GeV}, \quad p_T(2) > 20\,\text{GeV} \\ &\Delta R_{ij} > 0.65, \quad \Delta R_{ii} < 3.85 \\ &M_{b\bar{b}\gamma\gamma} > 90\,\text{GeV}, \quad p_{T_{ii}} > 160\,\text{GeV}, \quad M_{b_1\gamma_1} > 125\,\text{GeV} \\ &|M_{bb} - m_h| < 24\,\text{GeV}, \quad |M_{\gamma\gamma} - m_h| < 12\,\text{GeV} \end{aligned} \tag{5.4}$$

There is an enormous gain in the significance after using BDT to help classifying signal and background events. However, the $S/\sqrt{B}$ metric overestimates the significance when the number of signal events is not much smaller than the number of background events. In Table V, we show the maximum signal significance by cutting on the BDT scores with extended backgrounds and using the more conservative and best suited significance AMS of Eq. (A3). We display in this table the results for all the cut strategies of Table II plus the best cut strategy found with the Bayesian method.

First, whatever the cut strategy, the BDT classification significantly enhances the signal significance compared to the simple cut-and-count analysis. The larger AMS, however, is once again the one obtained by selecting the cut strategy with the optimized search, reaching $\sim 3.9\sigma$ with 3 ab^{-1} of integrated luminosity. It is interesting to note that the selection cuts found for AMS are different from those of Eq. (5.4) for $S/\sqrt{B}$

$$\begin{aligned} &p_T(1) > 92\,\text{GeV}, \quad p_T(2) > 20\,\text{GeV} \\ &\Delta R_{ij} > 0.2, \quad \Delta R_{ii} < 2.6 \\ &M_{b\bar{b}\gamma\gamma} > 10\,\text{GeV}, \quad p_{T_{ii}} > 125\,\text{GeV}, \quad M_{b_1\gamma_1} > 70\,\text{GeV} \\ &|M_{bb} - m_h| < 30\,\text{GeV}, \quad |M_{\gamma\gamma} - m_h| < 9\,\text{GeV} \end{aligned} \tag{5.5}$$

From the previous results and those of Eqs. (5.3), (5.4), (5.5), we observe that the Bayesian optimization algorithm learns basically two types of selection criteria to increase the significance: either relaxing the ΔR and hardening

some of the invariant masses and transverse momenta variables, or placing more stringent ΔR and relaxing invariant mass and transverse momentum cuts. This is perfectly understandable from the physics point of view: events with high p_T particles and large invariant masses are more likely to contain collimated photons and b jets, thus cutting both on ΔR and invariant masses, for example, would be redundant as also can be seen in panels (a) and (b) of Fig. 4. The job of the optimization algorithm is more a fine tuning of the cuts throughout the variables space.

Another feature of the best cut criteria found so far by the Bayesian approach is the b-tagging dependence with the transverse momentum as parametrized in the `Delphes` detector simulator. Once the b-tagging increases with the bottom quark transverse momentum, selection criteria with at least one high-p_T is likely to provide a better discrimination against important non-b-jet backgrounds as $jj\gamma\gamma$, $c\bar{c}\gamma\gamma$ and $c\bar{c}\gamma j$.

C. Joint search for best cuts and BDT hyperparameters

In the previous section, we carried out a sequential search for cuts and BDT hyperparameters, first adjusting the BDT to perform well on the baseline selection criteria, then, with the hyperparameters fixed, continuing to the search of best cuts.

In this section, we will investigate whether a joint search for all the parameters of the phenomenological analysis can also yield good results. The relevant parameters that need to be adjusted together are both the cut thresholds and the BDT hyperparameters. This represents a more thorough approach to the problem of getting the best performance possible using a ML algorithm.

For this global search we used `Hyperopt` with the parameters space of the Table VI. All the prior distributions were assumed to be uniform in the range indicated in the right column. As in the previous analysis, the objective function to be minimized was $-AMS$. All the BDT results were obtained from a fivefold cross validation by randomly splitting training and testing samples at the proportion of $2/3$ and $1/3$ of the total sample, respectively. As the parameter space is larger now, we allowed for 300 trials.

As in the previous sections, we plot, in Fig. 8, histograms of the number of cut strategies for a given significance interval for the joint search. The black dashed line now represents the maximum significance of the sequential search of the previous section. In contrast to the other cases, the global search found just a few better cut strategies, but the important fact is that it actually found a better strategy than the sequential search, showing that a joint search is not only possible but also beneficial to the AMS maximization.

The maximum AMS is 4.0σ for an AUC of 0.904, against 3.9σ and AUC of 0.901 of the sequential parameters search. The parameters of the joint search are the following

TABLE VI. The kinematic variables used for cuts and BDT hyperparameters and their allowed variation ranges in `Hyperopt` for the joint optimization. The prior distributions for all these variables are set to uniform distributions over the ranges shown in the table within the steps shown as the last entry of each vector. In a grid search, the number of evaluation points would be approximately 8.9×10^{20}.

Kinematic variable/BDT hyperparameter	Variation range in `Hyperopt`
$\Delta R_{ii} <$	(1.4,4,0.05)
$\Delta R_{ij} >$	(0,2,0.05)
$p_T(1) >$	(30,100,1) GeV
$p_T(2) >$	(20,70,1) GeV
$p_{T_{ii}} >$	(0,200,5) GeV
$M_{b\bar{b}\gamma\gamma} >$	(0,400,5) GeV
$M_{b_1\gamma_1} >$	(0,200,5) GeV
$\lvert M_{\gamma\gamma} - m_h\rvert <$	(5,15,1) GeV
$\lvert M_{bb} - m_h\rvert <$	(10,30,1) GeV
`number of trees`	(150,250,1)
`learning rate`	(0.001,0.5,0.001)
`maxixum tree depth`	(2,20,1)
`min_child_weight`	(1,6,1)

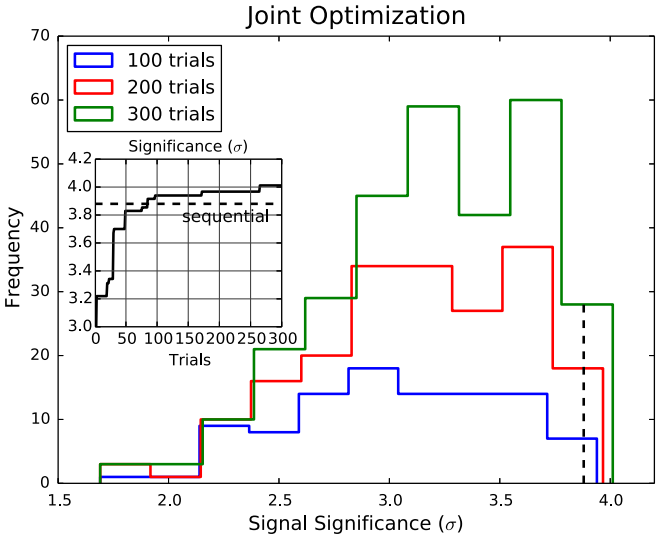

FIG. 8. The histogram of number of cut strategies producing a given significance interval with BDT adjusted in a joint optimization of cuts and hyperparameters. The inset plot shows the significance as a function of the number of `Hyperopt` trials. No systematics are assumed, the backgrounds are those of Ref. [4] and the $S/\sqrt{B}$ used to compute the signal significances. The black dashed line represents the results obtained with the default cuts of Azatov *et al.*, Ref. [4].

$$
\begin{aligned}
&p_T(1) > 72\ \text{GeV}, \quad p_T(2) > 20\ \text{GeV}\\
&\Delta R_{ij} > 0.15, \quad \Delta R_{ii} < 3.6\\
&M_{b\bar{b}\gamma\gamma} > 370\ \text{GeV}, \quad p_{T_{ii}} > 145\ \text{GeV}, \quad M_{b_1\gamma_1} > 100\ \text{GeV}\\
&|M_{bb} - m_h| < 27\ \text{GeV}, \quad |M_{\gamma\gamma} - m_h| < 11\ \text{GeV}\\
&\texttt{number of trees} = 157\\
&\texttt{learning rate} = 0.101\\
&\texttt{maximum tree depth} = 14\\
&\texttt{min child weight} = 5
\end{aligned}
\tag{5.6}
$$

The joint search was able to find a more regularized set of BDT hyperparameters to avoid overfitting. This is why the `number of trees` is smaller and `min_child_weight` bigger than those of the sequential search. This is a very welcome result—with harder cuts than those of the default cuts of Azatov *et al.*, `Hyperopt` learned how to control the loss in AMS that would be caused by a more dangerous overfitted BDT due a smaller size sample for training and testing. Moreover, it was able to tune the parameters to perform slightly better than the sequential search. Finally, we note that the optimized cuts are of the type that feature harder invariant masses and transverse momenta and relaxed ΔR cuts.

As a final investigation, we present in the next section a multivariate statistical analysis based on the BDT output scores and the inclusion of systematic uncertainties for our final more realistic prospects of discovering the double Higgs production at the LHC.

VI. FINAL RESULTS: FURTHER DISCRIMINATION WITH MULTIVARIATE ANALYSIS OF BDT OUTPUTS

In the previous sections, we employed a ML algorithm to boost our classification accuracy of signal and background events, relying exclusively on cut-and-count analysis and posterior calculation of the significance with an approximated median significance formula.

In this section, we will attempt to improve the signal significance by focusing on the statistical side of the analysis encouraged by the results of Ref. [46], around 4σ for 3 ab^{-1} with MVA based on kinematic variables but with no systematics included. This will be done by estimating the log-likelihood ratio statistics from the output scores of the BDT algorithm provided by `XGBoost`. This is a well known and established procedure used by the LHC Collaborations for a long time, but only recently more rigorously justified [18].

We calculate the log-likelihood ratio of the binned BDT output scores for signal s_i and backgrounds b_i, $i = 1, \ldots, N_{\text{bins}}$, after cuts, shown in the panel (d) of Fig. 6, according to [55]

$$
\Lambda = \sum_{i=1}^{N_{\text{bins}}} \left[-s_i + d_i \ln\left(1 + \frac{s_i}{b_i}\right)\right] \tag{6.1}
$$

We assume that the simulated data d_i follows either the null hypothesis with no signal, $d_i \sim \text{Pois}(x_{\text{BDT}_i}|B)$, to compute Λ_B, or the alternative hypothesis where $d_i \sim \text{Pois}(x_{\text{BDT}_i}|S+B)$ to compute Λ_{S+B}. The estimation of the nonparametric statistical distributions of Λ_B and Λ_{S+B}, $P(\Lambda|B)$ and $P(\Lambda|S+B)$, respectively, is done with a large number of pseudoexperiments with new statistically varied BDT output distributions assuming that the number of events in each bin is drawn from a Poisson distribution, $\text{Pois}(x|\mu)$, of mean μ. From these distributions the p-value of the background hypothesis is calculated

$$
p_B = \int_{\Lambda_{S+B}}^{+\infty} P(\Lambda|B) d\Lambda, \tag{6.2}
$$

and the statistical significance is computed as $\Phi^{-1}(1-p_B)$, where Φ is the cumulative distribution function of the standard Gaussian with zero mean and unit variance.

According to the Neyman-Pearson lemma [56], the likelihood ratio is the most powerful test statistic to discriminate a signal hypothesis for a fixed significance level of the background hypothesis (a fixed background efficiency) in the absence of systematic uncertainties.

In this work, in order to estimate $P(\Lambda|B)$ and $P(\Lambda|S+B)$, we performed 40000 pseudoexperiments from the binned BDT output scores. As in the previous sections, we used `Hyperopt` to search for the cut strategy with the biggest significance after training the BDTs and computing the AMS as described above. The BDT hyperparameters were fixed as in Eq. (5.2), so Bayesian search was applied in the sequential way.

The histogram of cut strategies as a function of the significance of Fig. 9, as in the other cases, shows that more than 90% of all cut selections found by `Hyperopt` lead to a better MVA performance than that of the default cuts of Azatov *et al.*, which are definitely not suited to MVA. Also, similarly to other cases studied previously, the maximum significance is found rather early in the searching, with 100 experiments, as shown in the inset plot of Fig. 9. A very high AMS is already obtained at that stage, and it is the best strategy up to almost the 500th experiment which improves it very slightly.

Our final analysis and results take into account systematic uncertainties of 10% and 20% and are shown in the Table VII. The systematic uncertainties are incorporated in MVA in mixed frequentist-Bayesian method, by marginalizing over the background rate in Eq. (6.2) assuming that the systematic errors are Gaussian. All the backgrounds are taken into account, including $c\bar{c}\gamma\gamma$, $b\bar{b}\gamma j$ and $c\bar{c}\gamma j$, the significance was calculated with the AMS formula (A3), and the integrated luminosity corresponds to 3 ab^{-1}.

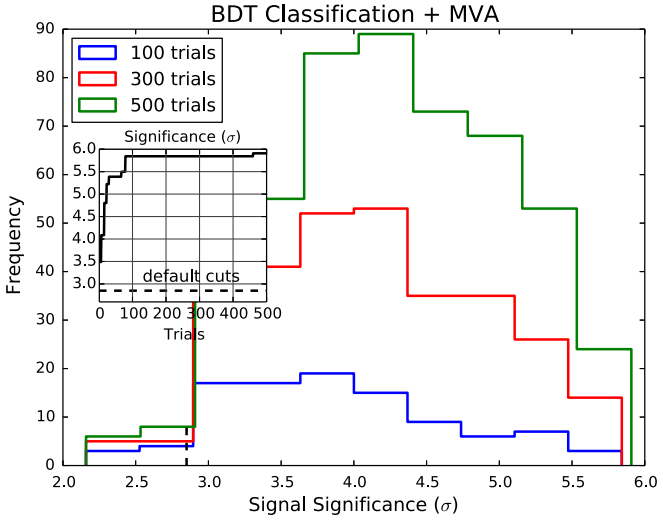

FIG. 9. The histogram of number of cut strategies producing a given significance interval in MVA. The inset plot shows the significance as a function of the number of `Hyperopt` trials. No systematics are assumed, the backgrounds are those of Ref. [4] and the $S/\sqrt{B}$ used to compute the signal significances. The black dashed line represents the results obtained with the default cuts of Azatov *et al.*, Ref. [4]. The optimization was of the sequential type.

With a very low level of systematics, the techniques proposed here with the selection criteria optimization may be able to confirm the production of a pair of SM Higgs bosons with 5σ. Within a more realistic projection of the level of systematics, around 10%, the optimization of cuts to train boosted decision trees combined with a multivariate analysis delivers a respectable significance of 4.6σ. This is the largest significance achieved so far in the $b\bar{b}\gamma\gamma$ channel with realistic assumptions concerning backgrounds and systematic uncertainties at the 14 TeV LHC. Even assuming large systematics of 20%, our analysis predicts a 3.6σ significance, which represents at least a strong evidence in favor of double SM Higgs production.

TABLE VII. Signal significances for cut-and-count, BDT and MVA are shown in the second, third and fourth column, respectively, for 0%, 10% and 20% systematics. We took all backgrounds into account for the computation of the AMS with optimized cuts and an integrated luminosity of 3 ab^{-1} at the 14 TeV LHC. The bold-face numbers represent the significances expected with the level of systematics anticipated by the experimental collaborations in Refs. [5,7,43]. The numbers inside brackets are the significances computed with the default cuts of Azatov *et al.*, Ref. [4], which we took as baseline results.

Systematics (%)	Cut-and-count	BDT	MVA
0	2.34[1.76]	3.88	5.05
10	**1.93**[1.43]	**3.57**	**4.64**
20	1.51[1.0]	3.10	3.60

Relying just on BDT classification with optimized cuts, for systematics below 20%, a robust evidence for double Higgs production is possible according to Table VII.

Compared to the default cuts of Ref. [4], which we took in this work as our baseline results, the cuts found from the Bayesian optimization are able to enhance the significance by 30%–50% with little computational efforts and speed. The results for the default cuts of Azatov *et al.* are shown between brackets in the second column of Table VII for comparison.

Finally, we elect from all the results presented, those of the second row of Table VII as the most representative of our findings, again stressing that these results take into account realistic backgrounds, the level of systematic uncertainties expected for this channel, and also better suited significance metrics for the number of signal and background events expected at the LHC with these selection criteria.

VII. CONCLUSIONS AND PROSPECTS

In this paper, we explored double Higgs production via gluon fusion at the LHC. Our analysis builds significantly on previous studies in that we used tools from the ML literature to discriminate signal and background events. We also incorporated background contributions coming from light flavor jets or c jets being misidentified as b jets and electrons or jets being misidentified as photons.

First we used Bayesian optimization, implemented in `Hyperopt`, to select cuts on kinematic variables, obtaining a 30%–50% increase in the significance metric $S/\sqrt{B}$ compared to current results in the literature. Then, we used BDTs implemented in `XGBoost` to further discriminate signal and background events. At this stage, we showed that a joint optimization of kinematic cuts and BDT hyperparameters results in an appreciable improvement in performance. Finally, we turned to the statistical side of the analysis by estimating the log-likelihood ratio statistics from the output scores of the BDT algorithm provided by `XGBoost`. The final results of our paper are presented in Table VII. We find that assuming a very low level of systematics, the techniques proposed here will be able to confirm the production of a pair of SM Higgs bosons at 5σ level. Assuming a more realistic projection of the level of systematics, around 10%, the optimization of cuts to train BDTs combined with a multivariate analysis delivers a respectable significance of 4.6σ. This is the largest significance achieved so far in the $b\bar{b}\gamma\gamma$ channel with realistic assumptions concerning backgrounds and systematic uncertainties at the 14 TeV LHC. Even assuming large systematics of 20%, our analysis predicts a 3.6σ significance, which represents at least strong evidence in favor of double SM Higgs production.

We pause for a moment to recapitulate the reasons behind the larger significances obtained in this paper, compared to previous studies. What makes the significances larger is

precisely the better discrimination between the signal and background classes achieved by the machine learning algorithms as they find more profound correlations among the kinematic features and those classes. These correlations cannot be fully explored in simple/manual rectangular cut-and-count analyses. There is a tradeoff between the efficiency of the cuts and the ML performance which is usually neglected in phenomenological works where these tools are employed. The reasoning is simple: cutting harder cleans up more backgrounds but weakens the correlations between the kinematic variables and the event classes, thereby decreasing the ML performance. On the other hand, relaxing the cuts makes the correlations stronger helping to boost ML but the discrimination power gained might not be enough to get a good significance with a large number of surviving background events. Finding the optimal performance from this competition is the core of the method present in the paper.

We now turn to some future prospects. One immediate future goal is to study the prospects of measuring deviations of λ_3 from the SM prediction at the high-luminosity LHC using our work on double Higgs processes in the $b\bar{b}\gamma\gamma$ channel [57]. In this context, it would also be interesting to pursue the ensuing implications for the electroweak phase transition within an effective potential framework. Another set of goals is to extend our work to other final states like $b\bar{b}\tau^+\tau^-$, $b\bar{b}W^+W^-$, and $b\bar{b}b\bar{b}$, as well as other production channels.

There are also several directions one can pursue that are not necessarily related to studies of the Higgs sector. The Bayesian optimization approach to the cut selection presented in this work can be used in other phenomenological studies. For example, it would be very interesting to use our methods to reevaluate the discovery prospects for compressed supersymmetric searches or dark matter [58–60]. The Bayesian optimization can also be used to design a cut selection that helps to overcome the effect of various types of systematic uncertainties which affect the shape of the distributions and the normalization of the cross sections.

The measurement of particles masses, couplings and quantum numbers like spin and CP also depend strongly on the kinematic selection criteria. This is another target for optimization using `Hyperopt`. As we showed in this work, a multivariate analysis used for hypothesis tests can be greatly enhanced with a careful set of cuts aimed to keep strong correlations but eliminating as much backgrounds as possible. Some other discrimination techniques which suffer with hard cuts, such as the calculation of asymmetries [50,51,61–63], are also worth investigating using our methods.

While we performed a discovery analysis in this work, an optimized set of cuts, with or without further classification with the help of ML tools, can also be employed to obtain stringent limits in exclusion studies.

It is certain that cut optimization will be able to improve the performance of other classifiers such as neural networks and naive Bayes-inspired algorithms which are commonly explored in phenomenological studies, although it is difficult to estimate the extent. One might anticipate that the cut selection which optimizes a given classifier should not correspond to the selection that improves another. The joint optimization presented here is also a potential target of further investigation as the cut optimization can be performed at the same time of hyperparameters tuning of classifiers as decision trees and neural networks. These are directions that can also be pursued in the future.

ACKNOWLEDGMENTS

A. A. would like to thank Fundação de Amparo à Pesquisa de Estado de São Paulo (FAPESP), Process No. 2013/22079-8, and Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, Grant No. 307098/2014-1. K. S. would like to thank Minho Son for a very illuminating discussion and the IBS Center for Theoretical Physics of the Universe, Daejeon, South Korea, for hospitality when this work was in progress. He would also like to thank Paul Padley, Ankit Patel, and Jamal Rorie for helpful discussions. T. G. is supported by the U.S. Department of Energy Grant No. de-sc 0016013.

APPENDIX A: STATISTICAL SIGNIFICANCE METRICS

A comparative study of various statistical significance metrics can be found in [64]. In that work, the problem of incorporating systematic uncertainties in the background normalization for a Poisson process is addressed and it is found that three most widely used significance metrics perform similarly in many situations concerning the relative number of signal and background events and the level of systematics.

The three significance methods are

(1) The naive and most simple way to incorporate systematic uncertainties in the calculation of the significances for S signal events and B background events in a Poisson process for a given integrated luminosity

$$\frac{S}{\sqrt{B+(\varepsilon_B B)^2}}. \tag{A1}$$

In all cases, we assume that the systematic uncertainty in the total background normalization is proportional to the number of background events, $\varepsilon_B B$. This is simple and fast, but it somewhat overestimates the discovery reach with or without systematics.

(2) The Bayesian-frequentist hybrid recipe to the estimation of the systematics impact on the significance. Assuming that systematic errors are normally distributed we marginalize over the systematic errors to obtain the p-value

$$p_B = \sum_{k=S+B}^{+\infty} \int_{-\infty}^{+\infty} \frac{e^{-B(1+z\varepsilon_B)}}{k!} [B(1+z\varepsilon_B)]^k \times \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} dz \tag{A2}$$

and the significance is computed as $Z = \Phi^{-1}(1-p_B)$, where $\Phi(z)$ is the cumulative distribution function of the standard normal distribution.

This is method of incorporating systematics into the significance was employed in Sec. VI the MVA analysis and it is computationally more demanding.

(3) The profile likelihood method originally proposed in [65] in astrophysical searches with subsidiary measurements of the background adapted to a high energy experiment where the systematics is a fraction of background events, $\varepsilon_B B$

$$\text{AMS} = \begin{cases} \sqrt{2}\left\{(S+B)\ln\left[\left(1+\frac{1}{B\varepsilon_B^2}\right)\frac{S+B}{S+B+1/\varepsilon_B^2}\right] + \frac{1}{\varepsilon_B^2}\ln\left[\frac{B+1/\varepsilon_B^2}{S+B+1/\varepsilon_B^2}\right]\right\}^{\frac{1}{2}}, & \varepsilon > 0 \\ \sqrt{2}\left[-S+(S+B)\ln\left(1+\frac{S}{B}\right)\right]^{\frac{1}{2}}, & \varepsilon = 0 \end{cases} \tag{A3}$$

Among the three metrics this is the most conservative and reliable, and it is as simple and fast to compute as the naive metrics of Eq. (A1). Moreover, its performance is very close to the consistent frequentist approach for tests of the ratio of Poisson means implemented in ROOT [66], for example.

A comparison of these three methods are also investigated in Ref. [67] in the context of the search for dark matter production in the mono-Z channel confirming all the features anticipated in Ref. [64]. In the case of double Higgs production, we also checked that the naive formula of Eq. (A1) always provide larger significances with or without systematics compared to the other metrics for the same number of signal and background events.

APPENDIX B: PYTHON CODE OF THE OPTIMIZATION METHOD USING `Hyperopt`

We show a snippet of the code used to optimize the cut strategies right below.

Listing 1: Python snippet of the optimization code.

```
# loading packages
import numpy as np
from hyperopt import hp, fmin, tpe, STATUS_OK, Trials
from functools import partial
# loading data
data = np.genfromtxt('data/data.csv', delimiter=',')
n_data, ncol = data.shape
print data.shape
ncol=ncol-1
# raw data
X_raw = data[:,1:ncol] #vector of features, 27 for HH production
y_raw = data[:,0]      #labels
weights=data[:,ncol]   #events weights
print ('finish loading '+str(n_data)+' samples from csv file')
# evaluation parameters
nevals=200
# building the selector function
def selector(y):
    aux_min=int(min(y))
    aux_max=int(max(y))
    sel=[[] for i in range(int(aux_max))]
    for i in range(int(aux_max)):
```

```
        sel[i] = np.array([y[k] == float(i+1) for k in range(len(y))])
    return sel
# AMS metrics
def ams(s,b,sys):
    breg=0.0
    #return s/np.sqrt(b+breg+(sys*(b+breg))**2)
    #return np.sqrt(2.0)*np.sqrt( (s+b+breg)*np.log(1.0+s/(b+breg))-s)
    if b==0. and sys!=0.:
        aux=np.sqrt(2*s*(1.+1./sys))
    else:
        b=b+breg
        if sys==0.:
            aux=np.sqrt(2.0)*np.sqrt(-s+(s+b)*np.log(1.+s/b))
        else:
            aux=np.sqrt(2.0)*np.sqrt((s+b)*np.log((1.+1./(b*sys**2))*(s+b)/(s+b+1./sys**2))+(1./sys)**2*np.log((1.+b*sys**2)*(1/sys**2)/(s+b+1./sys**2)))
    return aux
# computing the number of signal, backgrounds events for a given selection
def Nevents(w,sel):
    nev=len(y)
    # number of events of each class
    nevS  = np.sum(np.array([w[sel[0]]])) #signal
    nevB1 = np.sum(np.array([w[sel[1]]])) #background 1
    nevB2 = np.sum(np.array([w[sel[2]]])) #backgriund 2
    nevB3 = np.sum(np.array([w[sel[3]]])) #background 3
    nevB  = nevB1+nevB2+nevB3
    events= np.array([nevS,nevB1,nevB2,nevB3])
    return events
##############################
# Passcuts Boolean function #
##############################
# variables contained in the vector of features
vars={'pT1':1, 'pT2':2, 'Mii':3, 'Mij':4, 'Rij':5}
# defining cut variables
mh=125.0
vd1=data[:,vars['pT1']]
vd2=data[:,vars['pT2']]
vd3=data[:,vars['Mii']]
vd4=data[:,vars['Mij']]
vd5=data[:,vars['Rij']]
vd6=abs(vd4-vd3+mh*np.ones(n_data))
# cuts function
def passcuts(cut,a):
```

```
    if a[0]>=cut[0] and a[1]>=cut[1] and abs(a[2]-mh)<=cut[2] \
       and a[3]>=cut[3] and a[4]<=cut[4] and a[5]>=cut[5]:
        aux=True
    else:
        aux=False
    return aux
################################
# CUT-AND-COUNT: TPE/HyperOpt #
################################
best_cc=[[] for i in range(22)]
best_cut=[[] for i in range(22)]
def objective(cuts):
    cut=np.array([cuts['pT1_cut'],cuts['pT2_cut'],cuts['Wii_cut'], \
                  cuts['Mij_cut'],cuts['Rij_cut'],cuts['Mxx_cut']])
    data_cut=np.array([data[i] for i in range(n_data) if \
                       passcuts(cut,[vd1[i],vd2[i],vd3[i],\
                                     vd4[i],vd5[i],vd6[i]])])
    if len(data_cut)!=0:
        y_cut = data_cut[:,0]
        n_cut = len(y_cut)
        w_cut = data_cut[:,ncol]
        sel_cut = selector(y_cut)
    # number of events of each class
        nevS, nevB1, nevB2, nevB3 = Nevents(w_cut,sel_cut)
        nevB  = nevB1+nevB2+nevB3
        loss=-ams(nevS,nevB,sys)
        print -loss
    else:
        print 'no events passed cuts'
        loss=0.
    return{'loss':loss, 'status': STATUS_OK}
# Cuts dictionary
cuts={
        'pT1_cut': hp.quniform("pT1_cut", 30., 100., 1.),    #70
        'pT2_cut': hp.quniform("pT2_cut", 20., 60., 1.),     #30
        'Wii_cut': hp.quniform("Wii_cut", 5., 15., 1.),      #10
        'Mij_cut': hp.quniform("Mij_cut", 100., 300., 10.), #20
        'Rij_cut': hp.quniform("Rij_cut", 0.4, 1.4, 0.1),    #10
        'Mxx_cut': hp.quniform("Mxx_cut", 100., 200., 10.)   #20
    }
print '-----HyperOpt SEARCH: '+str(nevals)+' experiments ------'
for j in range(0,25,5):
    sys=j*0.01
    print 'systematics = '+str(j)+'%'
    trials = Trials()
    best = fmin(fn=objective,
```

```
111                 space=cuts,
112                 algo=partial(tpe.suggest, n_startup_jobs=10)#rand. for
    random search
113                 max_evals=nevals,
114                 trials=trials)
115     print 'best:'
116     print best
117     best_cut[j]=best
118     # best ams calculation
119     cut=np.array([best['pT1_cut'],best['pT2_cut'],best['Wii_cut'], \
120                   best['Mij_cut'],best['Rij_cut'],best['Mxx_cut']])
121     data_best=np.array([data[i] for i in range(n_data) if \
122                         passcuts(cut,[vd1[i],vd2[i],vd3[i], \
123                                       vd4[i],vd5[i],vd6[i]])])
124     y_best = data_best[:,0]
125     n_best=len(y_best)
126     w_best = data_best[:,ncol]
127     # number of events of each class
128     nevS, nevB1, nevB2, nevB3 = Nevents(w_best,y_best)
129     nevB  = nevB1+nevB2+nevB3
130     best_cc[j]=ams(nevS,nevB,sys)
131     print 'sys, AMS, S/B =', sys, ams(nevS,nevB,sys), nevS/nevB
```

This code illustrates the basic steps to optimize a cut strategy with a single signal class and three different background classes as an example. It cannot be immediately used, but should be adapted to the reader analysis.

First, we load the basic Python packages `NumPy` and `Hyperopt` and also load the data from lines 1 to 14. If the data size is too big, it might be necessary to load it in batches. In line 16, we set the number of `Hyperopt` trials.

Signal and background samples need to be identified in several steps of the computation, we then create an event selector with the event labels as input in line 18. Significance metrics discussed in the previous appendix can be chosen in the definition of the `ams` function at line 26, $s(b)$ is the number of signal(backgrounds) events and `sys` the systematics level in the background rate ε_B.

In line 40, we define a function that returns the number of signal and background events given a selector vector.

From lines 50 to 70, we build a Boolean function which returns `True` if an event pass the cuts, otherwise it returns `False`. This function is inspired in the Fortran routine `PASSCUTS` found in the `MadAnalysis` package for `MadGraph` [19]. This function needs to be adjusted by the user according to his/her selection criteria. In this example, we put cuts on all the features of the event but this is not mandatory, of course. Instead, we construct in line 62 another cut variable which does not compound the features vector.

Now comes the part of the code where we actually perform the optimization. At line 76, we define our `objective` function which is going to be minimized by `Hyperopt`, its input is the cut dictionary placed at line 77. In this case, we are interested in maximize the `ams` function, that is, minimize `-ams`. Note that prior to the computation of `ams` we select those events which pass the cuts designed in `passcuts`. The labels and weights of these selected events are denoted by `y_cut` and `w_cut`, respectively. With `y_cut` we set the events selector at line 86 and then call the `Nevents` function to calculate the number of signal and background events after cuts, these numbers feed the `-ams` function at line 90.

In line 97, we set the a Python dictionary for the cut thresholds to be chosen by `Hyperopt` with the corresponding priors, in this case, all the priors were chosen to be uniform distributions. In Ref. [15], the user can find all the options to set the functioning of the program.

At line 106, we start a loop in the systematics level `sys` from 0% to 20%, from 5% to 5%. The TPE search is called in 110 in order to find the best cuts (with a warm-up phase of 10 trials) which return the larger AMS within `nevals` trials. From line 118 until line 131, we calculate

and print the results of the optimization for a given systematics. If one wants to perform a random search instead of using TPE, line 112 should be modified to `algo=rand.suggest`.

Note that the quantile γ discussed in Sec. IV A is, in principle, an adjustable parameter, but as far as we know there is no option to change it in `Hyperopt`. In Ref. [14], however, the authors keep this parameter at 0.15 for their studies.

[1] U. Baur, T. Plehn, and D. L. Rainwater, Probing the Higgs selfcoupling at hadron colliders using rare decays, Phys. Rev. D **69**, 053004 (2004).

[2] J. Baglio, A. Djouadi, R. Gröber, M. M. Mühlleitner, J. Quevillon, and M. Spira, The measurement of the Higgs self-coupling at the LHC: theoretical status, J. High Energy Phys. 04 (2013) 151.

[3] P. Huang, A. Joglekar, B. Li, and C. E. M. Wagner, Probing the electroweak phase transition at the LHC, Phys. Rev. D **93**, 055049 (2016).

[4] A. Azatov, R. Contino, G. Panico, and M. Son, Effective field theory analysis of double Higgs boson production via gluon fusion, Phys. Rev. D **92**, 035001 (2015).

[5] The ATLAS Collaboration, Report No. ATL-PHYS-PUB-2014-019 (2014).

[6] V. Barger, L. L. Everett, C. B. Jackson, and G. Shaughnessy, Higgs-Pair Production and Measurement of the Triscalar Coupling at LHC(8,14), Phys. Lett. B **728**, 433 (2014).

[7] The ATLAS Collaboration, Report No. ATL-PHYS-PUB-2017-001 (2017).

[8] U. Baur, T. Plehn, and D. L. Rainwater, Examining the Higgs boson potential at lepton and hadron colliders: A Comparative analysis, Phys. Rev. D **68**, 033001 (2003).

[9] M. J. Dolan, C. Englert, and M. Spannowsky, Higgs self-coupling measurements at the LHC, J. High Energy Phys. 10 (2012) 112.

[10] A. Papaefstathiou, L. L. Yang, and J. Zurita, Higgs boson pair production at the LHC in the $b\bar{b}W^+W^-$ channel, Phys. Rev. D **87**, 011301 (2013).

[11] D. E. Ferreira de Lima, A. Papaefstathiou, and M. Spannowsky, Standard model Higgs boson pair production in the $(b\bar{b})(b\bar{b})$ final state, J. High Energy Phys. 08 (2014) 030.

[12] D. Wardrope, E. Jansen, N. Konstantinidis, B. Cooper, R. Falla, and N. Norjoharuddeen, Non-resonant Higgs-pair production in the $b\bar{b}$ $b\bar{b}$ final state at the LHC, Eur. Phys. J. C **75**, 219 (2015).

[13] J. K. Behr, D. Bortoletto, J. A. Frost, N. P. Hartland, C. Issever, and J. Rojo, Boosting Higgs pair production in the $b\bar{b}b\bar{b}$ final state with multivariate techniques, Eur. Phys. J. C **76**, 386 (2016).

[14] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, Algorithms for hyper-parameter optimization, *Advances in Neural Information Processing Systems*, edited by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (Curran Associates, Inc., New York, 2011), Vol. 24, pp. 2546–2554.

[15] J. S. Bergstra, D. Yamins, and D. Cox, Hyperopt: A Python library for optimizing the hyperparameters of machine learning algorithms, in *Proceedings of the 12th PYTHON in Science Conference, SciPy* (2013).

[16] `Hyperopt` software package, https://github.com/jaberg/hyperopt.

[17] T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785 (2016), https://github.com/dmlc/xgboost.

[18] K. Cranmer, J. Pavez, G. Louppe, and W. K. Brooks, Experiments using machine learning to approximate likelihood ratios for mixture models, J. Phys. Conf. Ser. **762**, 012034 (2016).

[19] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, J. High Energy Phys. 07 (2014) 079.

[20] H. L. Lai, J. Huston, S. Kuhlmann, J. Morfin, F. Olness, J. F. Owens, J. Pumplin, and W. K. Tung, Global QCD analysis of parton structure of the nucleon: CTEQ5 parton distributions, Eur. Phys. J. C **12**, 375 (2000).

[21] P. M. Nadolsky, H. L. Lai, Q. H. Cao, J. Huston, J. Pumplin, D. Stump, W. K. Tung, and C.-P. Yuan, Implications of CTEQ global analysis for collider observables, Phys. Rev. D **78**, 013004 (2008).

[22] E. W. N. Glover and J. J. van der Bij, Higgs boson pair production via gluon fusion, Nucl. Phys. **B309**, 282 (1988).

[23] D. A. Dicus, C. Kao, and S. S. D. Willenbrock, Higgs boson pair production from gluon fusion, Phys. Lett. B **203**, 457 (1988).

[24] T. Plehn, M. Spira, and P. M. Zerwas, Pair production of neutral Higgs particles in gluon-gluon collisions, Nucl. Phys. **B479**, 46 (1996); Erratum, Nucl. Phys. **B531**, 655(E) (1998).

[25] S. Dawson, S. Dittmaier, and M. Spira, Neutral Higgs boson pair production at hadron colliders: QCD corrections, Phys. Rev. D **58**, 115012 (1998).

[26] B. A. Kniehl and M. Spira, Low-energy theorems in Higgs physics, Z. Phys. C **69**, 77 (1995).

[27] D. de Florian and J. Mazzitelli, Two-loop virtual corrections to Higgs pair production, Phys. Lett. B **724**, 306 (2013).

[28] D. de Florian and J. Mazzitelli, Higgs Boson Pair Production at Next-to-Next-to-Leading Order in QCD, Phys. Rev. Lett. **111**, 201801 (2013).

[29] J. Grigo, K. Melnikov, and M. Steinhauser, Virtual corrections to Higgs boson pair production in the large top quark mass limit, Nucl. Phys. **B888**, 17 (2014).

[30] D. de Florian and J. Mazzitelli, Higgs pair production at next-to-next-to-leading logarithmic accuracy at the LHC, J. High Energy Phys. 09 (**2015**) 053.

[31] D. Y. Shao, C. S. Li, H. T. Li, and J. Wang, Threshold resummation effects in Higgs boson pair production at the LHC, J. High Energy Phys. 07 (2013) 169.
[32] S. Borowka, N. Greiner, G. Heinrich, S. P. Jones, M. Kerner, J. Schlenk, U. Schubert, and T. Zirke, Higgs Boson Pair Production in Gluon Fusion at Next-to-Leading Order with Full Top-Quark Mass Dependence, Phys. Rev. Lett. **117**, 012001 (2016); Erratum, Phys. Rev. Lett. **117**, 079901 (2016).
[33] S. Borowka, N. Greiner, G. Heinrich, S. P. Jones, M. Kerner, J. Schlenk, and T. Zirke, Full top quark mass dependence in Higgs boson pair production at NLO, J. High Energy Phys. 10 (2016) 107.
[34] D. de Florian, M. Grazzini, C. Hanga, S. Kallweit, J. M. Lindert, P. Maierhfer, J. Mazzitelli, and D. Rathlev, Differential Higgs Boson pair production at next-to-next-to-leading order in QCD, J. High Energy Phys. 09 (2016) 151.
[35] G. Heinrich, S. P. Jones, M. Kerner, G. Luisoni, and E. Vryonidou, NLO predictions for Higgs boson pair production with full top quark mass dependence matched to parton showers, arXiv:1703.09252.
[36] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, An Introduction to PYTHIA 8.2, Comput. Phys. Commun. **191**, 159 (2015).
[37] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi, DELPHES 3: a modular framework for fast simulation of a generic collider experiment, J. High Energy Phys. 02 (2014) 057.
[38] S. Dittmaier (LHC Higgs Cross Section Working Group Collaboration), Report No. CERN-2011-002, (2011).
[39] S. Dawson *et al.*, Higgs Working Group Report of the Snowmass 2013 Community Planning Study, arXiv: 1310.8361.
[40] M. L. Mangano, M. Moretti, F. Piccinini, and M. Treccani, Matching matrix elements and shower evolution for top-quark production in hadronic collisions, J. High Energy Phys. 01 (2007) 013.
[41] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, J. High Energy Phys. 07 (2014) 079.
[42] G. Aad *et al.* (ATLAS Collaboration), Search For Higgs Boson Pair Production in the $\gamma\gamma b\bar{b}$ Final State using pp Collision Data at $\sqrt{s} = 8$ TeV from the ATLAS Detector, Phys. Rev. Lett. **114**, 081802 (2015).
[43] The CMS Collaboration, Report No. CMS PAS HIG-16-032 (2016).
[44] T. Aaltonen *et al.* (CDF Collaboration), Observation of Single Top Quark Production and Measurement of—Vtb—with CDF, Phys. Rev. D **82**, 112005 (2010).
[45] F. Hutter, H. Hoos, and K. Leyton-Brown, Sequential model-based optimization for general algorithm configuration, in LION-5 (2011).
[46] F. Kling, T. Plehn, and P. Schichtel, Maximizing the significance in Higgs boson pair analyses, Phys. Rev. D **95**, 035026 (2017).
[47] Q. H. Cao, G. Li, B. Yan, D. M. Zhang, and H. Zhang, Double Higgs production at the 14 TeV LHC and the 100 TeV pp-collider, arXiv:1611.09336.
[48] Q. H. Cao, B. Yan, D. M. Zhang, and H. Zhang, Resolving the degeneracy in single Higgs production with Higgs pair production, Phys. Lett. B **752**, 285 (2016).
[49] For example, the `Spearmint` Python package, J. Snoek, H. Larochelle, and R. P. Adams, Practical Bayesian optimization of machine learning algorithms, Advances in Neural Information Processing Systems (2012), https://github.com/HIPS/Spearmint.
[50] A. J. Barr, Measuring slepton spin at the LHC, J. High Energy Phys. 02 (2006) 042.
[51] A. Alves and O. Eboli, Unravelling the sbottom spin at the CERN LHC, Phys. Rev. D **75**, 115013 (2007).
[52] P. Baldi, P. Sadowski, and D. Whiteson, Enhanced Higgs Boson to $\tau^+\tau^-$ Search with Deep Learning, Phys. Rev. Lett. **114**, 111801 (2015).
[53] C. Adam-Bourdarios *et al.*, The Higgs boson machine learning challenge, in *JMLR Workshop and Conference Proceedings* (2015), Vol. 42, p. 19. Ill-behaved AMS was an important issue discussed along this Kaggle contest.
[54] J. Brownlee, XGBoost With Python, Gradient Boosted Trees With XGBoost and scikit-learn, ebook, Machine Learning Mastery Series (2017).
[55] G. Cowan, Statistics for searches at the LHC, arXiv: 1307.2487.
[56] A. Stuart, J. K. Ord, and S. Arnold, *Kendall's Advanced Theory of Statistics*, Vol. 2A, 6th Ed. (Oxford University Press, Oxford, United Kingdom, 1999).
[57] A. Alves, T. Ghosh, and K. Sinha (to be published).
[58] B. Dutta, W. Flanagan, A. Gurrola, W. Johns, T. Kamon, P. Sheldon, K. Sinha, K. Wang, and S. Wu, Probing compressed top squark scenarios at the LHC at 14 TeV, Phys. Rev. D **90**, 095022 (2014).
[59] A. G. Delannoy *et al.*, Probing Dark Matter at the LHC Using Vector Boson Fusion Processes, Phys. Rev. Lett. **111**, 061801 (2013).
[60] B. Dutta, T. Ghosh, A. Gurrola, W. Johns, T. Kamon, P. Sheldon, K. Sinha, K. Wang, and Sean Wu, Probing compressed sleptons at the LHC using vector boson fusion processes, Phys. Rev. D **91**, 055025 (2015).
[61] A. Alves, O. Eboli, and T. Plehn, It's a gluino, Phys. Rev. D **74**, 095010 (2006).
[62] J. M. Smillie and B. R. Webber, Distinguishing spins in supersymmetric and universal extra dimension models at the large hadron collider, J. High Energy Phys. 10 (2005) 069.
[63] C. Athanasiou, C. G. Lester, J. M. Smillie, and B. R. Webber, Distinguishing spins in decay chains at the Large Hadron Collider, J. High Energy Phys. 08 (2006) 055.
[64] R. D. Cousins, J. T. Linneman, and J. Tucker, Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process, Nucl. Instrum. Methods Phys. Res., Sect. A **595**, 480 (2008).
[65] T.-P. Li and Y.-q. Ma, Analysis methods for results in gamma-ray astronomy, Astrophys. J. **272**, 317 (1983).
[66] `ROOT`, An object-oriented data analysis framework, http://root.cern.ch.
[67] A. Alves and K. Sinha, Searches for dark matter at the LHC: A Multivariate analysis in the mono-Z channel, Phys. Rev. D **92**, 115013 (2015).