Project description for Admission to the PhD Program

Search for di-Higgs in $\gamma\gamma$ + multi lepton final states with the ATLAS detector

Ruben Guevara

October 12, 2023

1 Main objective and summary of the project

After the Higgs boson, the last ingredient of the Standard Model (SM), was discovered in 2012 [1, 2], we still have questions of its nature. Is it the long-awaited SM particle, or is it the first of a series of new scalar particles that go Beyond the Standard Model (BSM)? With the advent of higher energies and higher collision rates, the Large Hadron Collider (LHC) continues the voyage towards new physics phenomena. The ambitious LHC physics programme may shed light on some of the greatest mysteries in physics today.

The focus of this work is to further study the Higgs boson, in particular the production of two Higgs bosons from the same collision, which is referred to as di-Higgs production. This work will focus on a subset of the many possible di-Higgs decay channels which are characterized by one Higgs boson decaying into a two photon ($\gamma\gamma$) final state, with the other Higgs boson decaying into a final state containing multiple leptons. To study the elusive di-Higgs channels we will make use of Machine Learning (ML), in particular supervised learning. Both a Neural Network (NN) and Boosted Decision Tree (BDT) will be studied using Monte Carlo (MC) simulated data samples featuring well defined predictions from the SM. The analysis will use proton-proton collisions at the LHC from Run 2 (2015-2018) at 13 TeV with an integrated luminosity of 139 fb^{-1}, as well as Run 3 (2022-2025) at 13.6 TeV, currently being recorded by the ATLAS detector and planned to have an integrated luminosity of 250 fb^{-1}. As the di-Higgs production has a very low cross-section, meaning that we expect few events to be recorded, we will also study novel ML methods to estimate the likelihood function directly after training, opening the door for a more general statistical test of the theories against the recorded data.

2 Project background and scientific basis

There have been several searches for the di-Higgs production throughout the years by both the ATLAS and CMS collaborations. The Higgs boson can directly decay to all massive particles in the SM, however the branching ratio of the decay is higher for heavier particles than others. As events with di-Higgs are already seldom seen, most analyses have been done on the decay of Higgs to the second heaviest quark predicted by the SM; the bottom/beauty quark, b. ATLAS has published papers for the searches for $HH \to b\bar{b}b\bar{b}$ in both a resonant [3] and non-resonant region [4], using data from Run 2 setting upper limits on the cross section to this decay channel. Resonant meaning that there is an intermediate state that gives a peak in the di-Higgs invariant mass, while non-resonant assumes no such structure. While still decaying to the beauty quark, ATLAS has also published results for the decay of the Higgs to the taus, τ. the heaviest lepton in the SM, in the decay $HH \to b\bar{b} + \tau^-\tau^+$ [5]; and to photons, γ, in the resonant region for the decay $HH \to b\bar{b} + \gamma\gamma$ [6]. In addition to the decay of the Higgs to the beauty quark, ATLAS has studied the di-Higgs production in association with a vector boson [7], meaning the W boson decaying to $W \to \ell\nu$, and Z boson decaying to $Z \to \ell\ell, \nu\nu$. There are also unpublished results from the ATLAS collaboration (available as conference notes), for example the search for non-resonant $HH \to b\bar{b} + \gamma\gamma$ [8], and the non-resonant $HH \to 2b + 2\ell + E_{\mathrm{T}}^{\mathrm{miss}}$. The last two notes also studied BSM models given the triple Higgs vertex coupling, called the Higgs self-coupling. The Higgs self-coupling is also related to the self-coupling strength with respect to the SM, denoted by κ_λ. The measurement of the

Higgs self coupling still has high uncertainties and can potentially open the door to new physics. This Higgs self-coupling factor has also been interpreted in the framework of Effective Field Theory (EFT) [9, 10, 11], specifically assumptions of EFT such as SMEFT [12] and HEFT [13]. However, as mentioned in the summary, the decay channel to be covered by this work will entail a final state consisting of multi leptons (ml) + $\gamma\gamma$, which is currently being studied by ATLAS. Previously the $HH \to WW^*\gamma\gamma$ channel has been studied by ATLAS [14, 15] (Run 1 and 2 respectively) and CMS [16].

As mentioned in the summary above, we will also use novel techniques to estimate the likelihood function. While the likelihood function can be extracted by MC itself, this is time consuming and computationally expensive [17]. The motivation behind using ML to extract the likelihood function from the data is because it has shown the potential to increase the signal sensitivity, which is important for making statistical predictions. This work will expand upon the "likelihood-free" or "simulation-based inference" concept shown by Cramner et al. in 2016 in the paper "Approximating Likelihood Ratios with Calibrated Discriminative Classifiers" [18], where the key relies in doing a dimensionality mapping $\mathbb{R}^n \mapsto \mathbb{R}$ on the likelihood ratio, under the assumption that the corresponding transformation is itself monotonic with the likelihood ratio. Another novel ML algorithm we will study further is the use of Parametrized-NN (PNN) proposed by Baldi et al. [19]. The main idea behind PNNs is that, for new physics searches, it takes the (unknown) mass of the proposed new particles as an argument when training, meaning that one can train using the whole new physics dataset. Baldi et al. also propose that the PNN can by itself, interpolate between the mass points that are available in the dataset, meaning that we would not need to use other tools for the statistical interpolation between mass points. The idea is to combine both of these tools, as Cramner et al. already have done in the paper "The frontier of simulation-based inference" [20] and further developed to be applicable to the field of high energy physics [21, 22]. This method is also being studied by ATLAS in a Higgs analysis with promising results.

3 Research questions and scientific challenges

By using a new method, where we estimate the likelihood function directly from the ML output, we will go beyond the current state-of-the-art ML method used in high-energy physics where a binary classification is used to give a binned score of the events with different signal-to-noise ratios. By estimating the likelihood instead of doing the standard binary classification we hope to achieve a better use of the available information resulting in a more sensitive analysis. This is crucial for searches where the expected events are small, as the binning plays a central role in both estimating the uncertainties and signal thresholds.

The Cramner et al. article from 2016 [18] already showed promising results, with some areas for improvements. One of these areas to improve upon comes from the interpolation between the data points the ML algorithm was trained upon. Cramner et al. used a Bayesian optimization procedure. However, as time went on and we look at the Cramner et al. article from 2020 [20], we see that this area has already been improved upon this by including direct information from the simulator as well as including elements from PNNs. This project will expand further upon the idea of using a PNN to interpolate between points in training. This is different than what is currently being studied by ATLAS, as we will will potentially build a new framework that fully utilizes the power of the simulation-based inference technique.

In addition to the work entailed in this thesis, the candidate will partake in a Qualification Task (QT) in order to become an ATLAS author. Such a task entails taking into account the special skills and availability of the candidate and corresponds to a workload of about 80 full working days. Normally the project is completed within a year. The details on what the technical work on the QT will entail are yet to be decided by the activity leader in the collaboration and the project supervisors.

In this search and many others in high energy physics, a question that comes in testing statistical hypotheses, is the estimation of the likelihood criterion λ of Neyman & Pearson [23] in $-2\log\lambda$. This $-2\log\lambda$ is approximated by a χ^2 distribution with large samples [24]. However, in the cases with low statistics it is difficult to estimate what its distribution should be, especially for physics searches with rare processes that generally have low statistics. Box [25] showed that even for samples of moderate size the $-2\log\lambda$ can be approximated to have the same moments as a χ^2 by multiplying $-2\log\lambda$ by a scale factor. This scaling was first used by Bartlett in 1937 [26], where the object was to show that for

any likelihood function satisfying certain very general conditions an improved χ^2 test of this type is, in theory, possible. This was further studied by Lawley in 1956 [27] expanding the likelihood even further. In high energy physics, one tests with Monte Carlo techniques whether the approximation to a χ^2 is valid. If there is time, a study of moments in $-2\log\lambda$ to improve the asymptotic approximation will be carried out, resulting in an additional article.

4 Scientific method

This work, as many ML works in all fields, can be categorized into three parts; Data preparation, Network training, and Interpretation of results. Thus we can express the scientific method in these categories:

Data preparation

To further probe our knowledge of the Higgs boson, we will use real data collected in the ATLAS detector from proton-proton collisions in the Run 2 and Run 3 at the LHC at a center of mass energy of 13 TeV and 13.6 TeV, respectively, and with an integrated luminosity of 139 fb^{-1} and 250 fb^{-1}, respectively. In addition to using real data from the experiment, we will use Monte Carlo simulated samples that are based on the detector structure and accelerator conditions, as well as theoretical predictions of the signal and background calibrated to the luminosity and energy we are studying.

As mentioned we will do the analysis using ML, more specifically supervised learning, where we know whether a simulated event is signal (di-Higgs event) or background (the rest of the SM). The utilization of the ATLAS data provides a crucial foundation for training and evaluating the ML algorithms, ensuring their relevance and applicability to real-world physics phenomena.

Network training

For the network training, the current standard approach is to create a signal region where the signal to background ratio is as high as possible to then calculate the statistical significance of the signal to either set limits on the theory or claim a discovery. However, as mentioned on the summary of the project, the novel method that will be used in this work is to estimate the likelihood with ML. To do this the method from Cramner et al. [18] will be explored and similar methods to get the most out of the dataset in a hopefully unbinned manner.

However, the approach of how the ML algorithms will be trained depends on the final size of the simulated training data set. As mentioned by Cramner et al. [18], if the data size is not big enough, the output from a BDT will not be a good likelihood estimation, due the nature of decision trees when splitting data. Another direct consequence from the nature of decision trees, limits its ability to include a new parameter (mass points or pdf. parameters) as an input for training in the same manner as the PNN described by Baldi et al. [19]. Thus if the training data sets size, consisting of simulations, is good enough we will study how BDTs can estimate the likelihood function for the di-Higgs search.

Regardless of the data size, we will study NNs, as we could use shallow NNs (with little data) or deep NNs (with more data). Due to the freedom and mathematical nature of the Fast Forward algorithm, the implementation of a new parameter can efficiently be included to generalize the training.

Interpretation of results

After evaluating the likelihood-ratio from the network training we will compare the results for the signal strength with current constraints on the value to see if there is any significant improvement to the sensitivity. If there is time we will also set constraints on the triple Higgs vertex value given by the data, including a possible preparation of the results such that they can in the future be used in an EFT interpretation.

5 Expected impact

The expected impact of the di-Higgs search depends highly on what the results yield. If we find a statistical significance that is compatible with the SM predictions, then we are closer to estimating the

important triple Higgs vertex parameter that is predicted by Quantum Field Theory. Estimating this parameter is of great importance due the fact that it can both further confirm that the SM is the leading theory of the building blocks of the universe and allow us to include more constraints on future BSM models which try to explain the phenomena that the SM lacks. The other, perhaps more exciting outcome, is that if we find a statistically significant difference between the SM predictions and the data, that would open the door to a new realm of physics we can explore with further research in the future.

In addition to this, a major goal of this project is to fine tune and combine already existing state-of-the-art methods to do hypothesis testing, then the methodology of this project can be utilized on all fields of science which wish to test any hypothesis that can be simulated. In addition, as the principle behind the ML algorithms that are used in advanced high energy physics can be boiled down to a simple binary classification, the methods used to prepare data and train networks can be used for works in academia as well as in the industry outside of academia.

6 Ethics

There are no direct ethical challenges with this work. The only thing that may become an ethical problem is the misuse of the algorithm, but to be able to change the structure of the binary classifier that gives us a likelihood function estimation, to something malicious, would require the same amount of work as creating a malicious ML algorithm from scratch, making this highly unlikely.

7 Project timeline

Autumn 2023 (First day 14.08.2023)

Combination of theoretical curriculum, in-depth statistics, and getting to know the Higgs community. The courses that will be taken are: STK9011 - Statistical Inference Theory, giving the foundation of the statistical theory for the whole project, and FYS9580 – Introduction to Nuclear Reactor Physics, as this is a special interest for me considering the societal relevance.

Spring 2024

Courses, ATLAS qualification task, teaching and ATLAS induction. Setting up the analysis framework and begin looking at Run 3 data. The courses are: MNSES9100 - Science, Ethics and Society, and MNPED9000 - Teaching in STEM.

Autumn 2024

Finish up the ATLAS qualification task, teaching and research. Made progress in optimizing and replacing BDT with NN on current analysis, and starting to implement the Simulation-based inference technique. Start work on asymptotic paper.

Spring 2025

Final approval of the qualification task, teaching and research. Get a traditional fit up and running with a fully re-optimized analysis using Run 2 + partial Run 3 data. Further work on asymptotic paper.

Autumn 2025

Teaching and research. End of Run 3 data taking. Include systematic uncertainties, as well as background modeling, on the analysis. Publish asymptotic paper. Start writing a potential paper on simulation-based inference.

Spring 2026

Teaching and research. Include full Run 3 data into analysis. Measuring the feasibility of the simulation-based inference method applied to the analysis. Further work on simulation-based inference paper.

Autumn 2026

Teaching, writing and research. Preparation of the final results, including the latest ATLAS Combined-Performance group recommendations. Publish paper on simulation-based inference.

Spring 2027

Teaching, Presenting final or near-final results at bi-annual Nordic Conference on Particle Physics (January 2027) and finishing up thesis. Possibly publishing di-Higgs multi-lepton paper with full Run 2 + full Run 3 data.

8 Statement from principal supervisor

The primary supervisor (Prof. Read) has worked on Higgs boson decays to the diphoton final state with the ATLAS experiment for more than 10 years. The secondary supervisor (Postdoc Shope) is an expert on Higgs boson decays to the W^+W^- final state with the ATLAS experiment and played, for example, a leading role in a very recent publication on Run 2 results. The rare process that Guevara will search for in Run 2 and Run 3 data and prepare prospects for High Luminosity LHC running in 2029-2037 (ca.) takes advantage of the expertise of both supervisors.

The ATLAS experiment is one of the flagship activities in the Norwegian Centre for CERN-related Research (NorCC) and in the Section for High Energy Physics at the Department of Physics, University of Oslo. The confirmation (or perhaps, surprisingly, the non-observation) of di-Higgs production, is one of the key targets for the High Luminosity LHC at CERN, which is a major part of the research program at CERN in the coming 15-20 years. Although we (Guevara et al) have not targeted the most sensitive final states, the sensitivity of the combined final results will depend on squeezing out every bit of sensitivity in all final states. The Higgs boson was discovered in 2012 by combing search results of several final states, and the story is likely to be repeated for the search for di-Higgs boson production.

9 Cooperation with external parties

As attaining an ATLAS authorship is required to publish papers with the collaboration, the QT will have an external supervisor for the technical work needed to gain the authorship. As the candidate will be part of a big research collaboration, collaboration with smaller and focused subgroups is expected. This collaboration entails attending ZOOM meetings or occasionally meeting the group in person at CERN.

References

1. ATLAS-Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. Physics Letters B 2012 Sep; 716:1–29. DOI: 10.1016/j.physletb.2012.08.020
2. CMS-Collaboration. Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC. Phys. Lett. B 2012; 716:30–61. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex]
3. ATLAS-Collaboration. Search for resonant pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state using pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. Phys. Rev. D 2022 May; 105(9):092002. DOI: 10.1103/PhysRevD.105.092002. Available from: https://link.aps.org/doi/10.1103/PhysRevD.105.092002
4. ATLAS-Collaboration. Search for nonresonant pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. Phys. Rev. D 2023 Sep; 108(5):052003. DOI: 10.1103/PhysRevD.108.052003. Available from: https://link.aps.org/doi/10.1103/PhysRevD.108.052003
5. ATLAS-Collaboration. Search for resonant and non-resonant Higgs boson pair production in the $b\bar{b}\tau^+\tau^-$ decay channel using 13 TeV pp collision data from the ATLAS detector. JHEP 2023; 07:040. DOI: 10.1007/JHEP07(2023)040. arXiv: 2209.10910 [hep-ex]

6. ATLAS-Collaboration. Search for Higgs boson pair production in the two bottom quarks plus two photons final state in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. Phys. Rev. D 2022 Sep; 106(5):052001. DOI: 10.1103/PhysRevD.106.052001. Available from: https://link.aps.org/doi/10.1103/PhysRevD.106.052001
7. ATLAS-Collaboration. Search for Higgs boson pair production in association with a vector boson in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. Eur. Phys. J. C 2023; 83:519. DOI: 10.1140/epjc/s10052-023-11559-y. arXiv: 2210.05415 [hep-ex]
8. ATLAS-Collaboration. Studies of new Higgs boson interactions through nonresonant HH production in the $b\bar{b}\gamma\gamma$ final state in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. Tech. rep. Geneva: CERN, 2023. Available from: https://cds.cern.ch/record/2870214
9. Burger AM. Probing the nature of electroweak symmetry breaking with Higgs boson pairs in ATLAS. 2023. Available from: https://cds.cern.ch/record/2871536
10. ATLAS-Collaboration. Constraints on the Higgs boson self-coupling from single- and double-Higgs production with the ATLAS detector using pp collisions at s=13 TeV. Physics Letters B 2023; 843:137745. DOI: https://doi.org/10.1016/j.physletb.2023.137745. Available from: https://www.sciencedirect.com/science/article/pii/S0370269323000795
11. Haisch U and Koole G. Off-shell Higgs production at the LHC as a probe of the trilinear Higgs coupling. JHEP 2022; 02:030. DOI: 10.1007/JHEP02(2022)030. arXiv: 2111.12589 [hep-ph]
12. Ellis J. SMEFT Constraints on New Physics Beyond the Standard Model. 2021. arXiv: 2105.14942 [hep-ph]
13. Dong ZY, Ma T, Shu J, and Zhou ZZ. The new formulation of higgs effective field Theory. Journal of High Energy Physics 2023 Sep; 2023. DOI: 10.1007/jhep09(2023)101. Available from: https://doi.org/10.1007%2Fjhep09%282023%29101
14. ATLAS-Collaboration. Searches for Higgs boson pair production in the $hh \to bb\tau\tau, \gamma\gamma WW^*, \gamma\gamma bb, bbbb$ channels with the ATLAS detector. Phys. Rev. D 2015; 92:092004. DOI: 10.1103/PhysRevD.92.092004. arXiv: 1509.04670 [hep-ex]
15. ATLAS-Collaboration. Search for Higgs boson pair production in the $\gamma\gamma WW^*$ channel using pp collision data recorded at $\sqrt{s} = 13$ TeV with the ATLAS detector. Eur. Phys. J. C 2018; 78:1007. DOI: 10.1140/epjc/s10052-018-6457-x. arXiv: 1807.08567 [hep-ex]
16. CMS-Collaboartion. Search for nonresonant Higgs boson pair production in the WW$\gamma\gamma$ channel in pp collisions at $\sqrt{s} = 13$ TeV. Tech. rep. Geneva: CERN, 2022. Available from: https://cds.cern.ch/record/2840773
17. Neal Radford M. Computing Likelihood Functions for High-Energy Physics Experiments when Distributions are Defined by Simulators with Nuisance Parameters. 2008. DOI: 10.5170/CERN-2008-001.111. Available from: https://cds.cern.ch/record/1099977
18. Cranmer K, Pavez J, and Louppe G. Approximating Likelihood Ratios with Calibrated Discriminative Classifiers. 2016. arXiv: 1506.02169 [stat.AP]
19. Baldi P, Cranmer K, Faucett T, Sadowski P, and Whiteson D. Parameterized neural networks for high-energy physics. The European Physical Journal C 2016 Apr; 76. DOI: 10.1140/epjc/s10052-016-4099-4. Available from: https://doi.org/10.1140%2Fepjc%2Fs10052-016-4099-4
20. Cranmer K, Brehmer J, and Louppe G. The frontier of simulation-based inference. Proceedings of the National Academy of Sciences 2020 May; 117:30055–62. DOI: 10.1073/pnas.1912789117. Available from: https://doi.org/10.1073%2Fpnas.1912789117
21. Brehmer J, Kling F, Espejo I, and Cranmer K. MadMiner: Machine learning-based inference for particle physics. Comput. Softw. Big Sci. 2020; 4:3. DOI: 10.1007/s41781-020-0035-2. arXiv: 1907.10621 [hep-ph]
22. Brehmer J and Cranmer K. Simulation-based inference methods for particle physics. 2020 Oct. arXiv: 2010.06439 [hep-ph]
23. NEYMAN J and PEARSON ES. ON THE USE AND INTERPRETATION OF CERTAIN TEST CRITERIA FOR PURPOSES OF STATISTICAL INFERENCE. Biometrika 1928 Dec; 20A:263–94. DOI: 10.1093/biomet/20A.3-4.263. eprint: https://academic.oup.com/biomet/article-pdf/20A/3-4/263/1037410/20A-3-4-263.pdf. Available from: https://doi.org/10.1093/biomet/20A.3-4.263

24. Wilks SS. The Large Distribution of the Likelihood Ratio for Testing Composite Hypotheses. Annals of Mathematical Statistics 1938; 9:60–2. DOI: 10.1214/aoms/1177732360. Available from: http://dx.doi.org/10.1214/aoms/1177732360

25. Box GEP. A general distribution theory for a class of likelihood criteria. Biometrika 1949; 36:317

26. Bartlett MS. Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences. 1937; 160(901):268. Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1937.0109

27. LAWLEY DN. A GENERAL METHOD FOR APPROXIMATING TO THE DISTRIBUTION OF LIKELIHOOD RATIO CRITERIA. Biometrika 1956 Dec; 43:295–303. DOI: 10.1093/biomet/43.3-4.295. eprint: https://academic.oup.com/biomet/article-pdf/43/3-4/295/987568/43-3-4-295.pdf. Available from: https://doi.org/10.1093/biomet/43.3-4.295