Charged Higgs boson study with $H^+ \to WH$ and $H \to \tau\tau$ Decay Mode in Proton-Proton Collision Simulation – Summer Student Report

Sarah Galea

CERN Summer Student 2024

Supervisor: André Sopczak

IEAP CTU in Prague

Abstract

This project focuses on advancing the search for the charged Higgs boson tbH^+ in the $\tau\tau$ production channel, focusing on events characterized by two same-sign electrically charged light leptons and one hadronically decaying tau (the $2\ell SS1\tau$ channel). Data processing for statistical and systematic analyses was performed for the v09 signal and background, with the results compared against the v08 analysis. Additionally, v09 includes systematic uncertainties on the tau lepton. The study was evaluated across seven mass points: 250 GeV, 300 GeV, 800 GeV (twice) 1500 GeV, 2000 GeV, and 3000 GeV for the charged Higgs boson.

1 Introduction

The discovery of the neutral Higgs boson in 2012 by the ATLAS and CMS experiments at CERN was a monumental achievement in particle physics, confirming the mechanism of electroweak symmetry breaking as predicted by the Standard Model (SM) [1]. However, the SM does not account for all existing phenomena and to address these gaps, several extensions of the SM have been proposed, including the Minimal Supersymmetric Standard Model (MSSM) and other two-Higgs-doublet models (2HDM), which predict the existence of additional Higgs bosons, including a charged Higgs boson ($H^{\pm}$). The detection of a charged Higgs boson would provide crucial insights into the nature of electroweak symmetry breaking and could potentially reveal new physics beyond the SM. However, identifying the charged Higgs boson in a proton-proton collider environment poses significant challenges due to the complex backgrounds and the need for precise measurements of its properties.

The Large Hadron Collider (LHC) at CERN accelerates proton beams to nearly the speed of light and collides them to explore the fundamental particles of the Universe. Each proton beam is composed of bunches, each containing up to 10^{11} protons, operating at a design luminosity of $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$. The significant radiation doses, high particle multiplicities, and elevated energy levels produced in these collisions necessitate the use of sophisticated detectors, such as ATLAS (A Toroidal LHC ApparatuS) shown in Figure 1, to observe and analyze the resulting events [2].

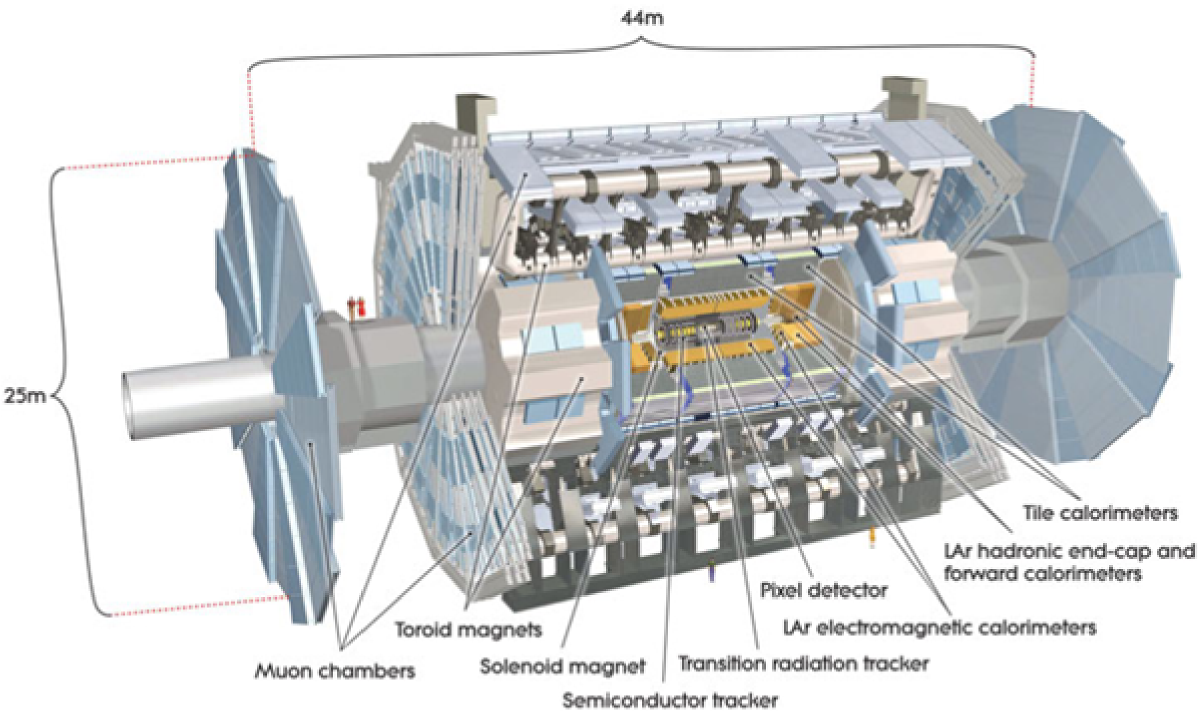

Figure 1: Cut-away view of the ATLAS detector [2].

1.1 Charged Higgs Boson Production

The charged Higgs boson ($H^{\pm}$) is a theoretical particle predicted by extensions of the SM, such as the MSSM and 2HDM [3]. Unlike the neutral Higgs boson discovered in 2012, the charged

Higgs boson carries an electric charge and represents a significant test of the SM's completeness and possible new physics beyond it. The production of charged Higgs bosons in a proton-proton collider can be categorized into two mass ranges, based on the mass of the Higgs relative to the top quark mass (m_t):

- **Low Mass Region ($m_{H^\pm} < m_t$):** Here, the production of $H^\pm$ predominantly occurs via the decay of a top quark ($t \to H^\pm b$) within top-antitop ($t\bar{t}$) pairs.
- **High Mass Region ($m_{H^\pm} > m_t$):** In this regime, charged Higgs bosons are mainly produced through the fusion of bottom and top quarks, resulting in the production processes $pp \to H^\pm tb$.

The subsequent decay modes of the charged Higgs boson depend on its mass and the specific model under consideration. In the context of this study, we focus on the decay $H^\pm \to WH$ followed by $H \to \tau\tau$, leading to the final state $tbH^+ \to tbWH \to tbW\tau\tau$. This decay chain is analyzed in the $2\ell SS1\tau$ channel, which involves events with two same-sign electrically charged leptons and one hadronically decaying tau lepton, as shown in Figure 2.

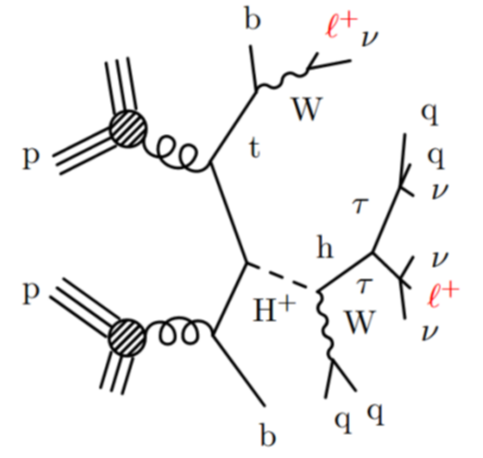

Figure 2: Feynman diagram showing the process of interest, tbH^+ [4].

1.2 Mass Points and Data Analysis

Given that the mass of the charged Higgs boson is not known, the analysis explores several mass points. Following Martin Rameš' thesis [5], the charged Higgs boson mass is sampled at multiple values: 250 GeV, 300 GeV, 800 GeV, 1500 GeV, 2000 GeV and 3000 GeV. For this project, both the older mass points (300 GeV, 800 GeV, 1500 GeV and 2000 GeV) and new mass points (250 GeV, 800 GeV, and 3000 GeV) are considered.

The analysis is conducted using the version 09 (v09) datasets Ntuples to evaluate the performance and consistency of the results. In particle physics, an Ntuple is a streamlined data format used to easily access specific parameters related to an event. It organizes data in a row-column structure, similar to a spreadsheet, where each row corresponds to a collision event and each column contains a variable linked to that event [6]. The new v09 Ntuple results were analyzed for each mass value and compared to the version 08 (v08) Ntuple results. This analysis aims to refine the search for the charged Higgs boson by improving the discrimination between signal and background, particularly in the context of the $2\ell SS1\tau$ final state.

1.3 Machine Learning Methods

Machine learning (ML) techniques have become increasingly valuable in high-energy physics, particularly for their ability to classify complex data. In this study, ML is employed to develop a classifier that assigns each event a probability of being associated with the process of interest. The implementation is achieved by the production of the friend Ntuples. By systematically excluding events with lower probabilities, the analysis aims to optimize the significance of signal events in the specified region, thereby improving the sensitivity of the search.

The implemented approach follows the work of Martin Rameš, using a Multilayer Perceptron (MLP) ML model. MLP is a type of artificial neural network that consists of multiple layers of nodes, or neurons, with each layer fully connected to the next [5]. A schematic of the MLP ML model is illustrated in Figure 3. The MLP model is trained on labeled datasets to learn the distinguishing features between signal and background events. This novel ML models for signal detection and fake lepton determination using the template fit method. The effectiveness of this method has been developed by Martin Rameš, and serves as the foundation for the current analysis [5].

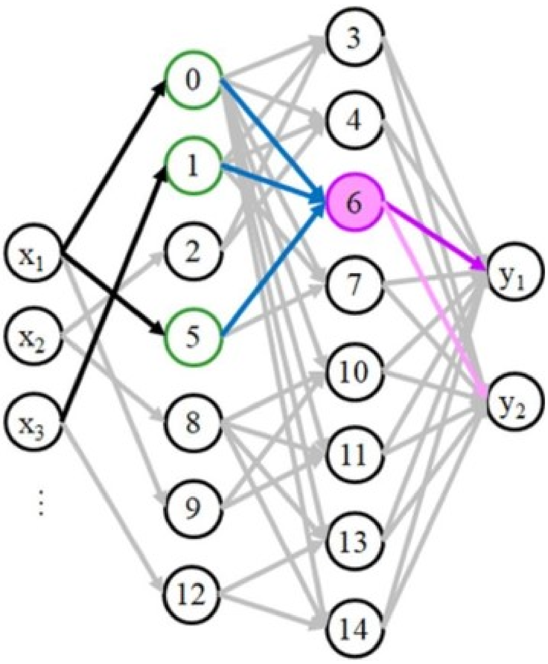

Figure 3: Schematic diagram of the MLP ML structure [7].

1.4 Version 09 Datasets

The primary goal of this project is to conduct a comprehensive analysis using the newly available v09 datasets, all within software release 21, and compare the outcomes with those from v08. This involves running the analysis on the v09 data, evaluating the results, and assessing how the updated dataset impacts the findings. The v09 datasets also have improved object definitions and requirements then the v08 dataset with better signal to background ratio. The luminosity of the v08 and v09 datasets is $140\,\mathrm{fb}^{-1}$.

The v09 datasets benefit from improved calibration procedures, ensuring more accurate measurements and corrections, leading to better data quality. Enhanced calibration reduces systematic uncertainties and provides a clearer view of the underlying physics processes, which is crucial for precise analysis and interpretation. The v09 datasets include a larger number of events compared to previous versions. The increased event statistics enhance the statistical power of the analysis, allowing for more robust and reliable conclusions. The v09 datasets come with better training and additional variables, inclusive of those needed for tau systematics. The inclusion of additional tau systematics allows for a more comprehensive treatment of uncertainties related to tau leptons.

2 Creation of the v09 Ntuples

The analysis presented in this report is based on the latest version of the Ntuples, referred to as v09. These Ntuples were generated specifically for the charged Higgs boson study, focusing on seven different mass points. The process of creating the v09 signal Ntuples involved running the GN2 Ntuple production code on datasets corresponding to the selected mass points of the charged Higgs boson The datasets used for this production are structured similarly to those used in the analysis of the top-associated Higgs production (ttH), with derivations labeled under the HIGG8D1 format. Each dataset is identified by a unique dataset identifier number (DSID), ensuring traceability and consistency across different stages of the analysis. The DSIDs for the signal datasets corresponding to the seven charged Higgs mass points are listed in Table 1.

Process Mass	**DSID**
New 250 GeV	512185
Old 300 GeV	510374_AF
Old 800 GeV	510375_AF
New 800 GeV	512187
Old 1500 GeV	510376_AF
Old 2000 GeV	510377_AF
New 3000 GeV	512186

Table 1: Mapping between DSID and processes.

2.1 Frameworks Used

2.1.1 Service for Web-based ANalysis (SWAN)

Service for Web-based ANalysis (SWAN) is a cloud-based platform provided by CERN that allows users to perform interactive data analysis. It leverages CERN's computing infrastructure to provide a seamless environment where users can access and process large datasets, such as those generated by the ATLAS experiment.

SWAN integrates with the CERN EOS storage system, enabling users to access data directly from the CERN file system. It supports multiple programming languages including Python, which was used in this analysis, and provides an environment for executing Jupyter notebooks. This flexibility makes SWAN an ideal platform for developing and running analyses remotely, without the need for local installations of complex software stacks.

In this project, SWAN was used to develop, test, and execute Python scripts that processed the Monte Carlo (MC) simulated datasets. Its integration with the EOS system allowed for efficient handling of large data files, facilitating the analysis process.

2.1.2 Python 3.0

Python is a versatile and widely-used programming language in the field of data analysis and scientific computing. It provides powerful libraries for numerical computation, data manipulation, and statistical analysis, making it an essential tool in high-energy physics research.

In this study, Python 3.0 scripts were written to handle the preprocessing of Ntuples, applying preselection criteria, and preparing data for further analysis. Python was also used to implement and train MLP ML models for classifying events in the analysis.

Python's extensive ecosystem of libraries and its compatibility with CERN's computing environment made it a crucial component of the analysis workflow.

2.1.3 ROOT

ROOT is a data analysis framework developed by CERN, specifically designed for handling large datasets in high-energy physics. It provides a comprehensive set of tools for data processing, statistical analysis, and visualization.

In this project, ROOT's TTree and TNtuple classes were used to efficiently store and manage large datasets, such as those generated in MC simulations. ROOT provided a wide range of statistical tools that were employed to perform fits, hypothesis testing, and other analyses necessary for the search for the charged Higgs boson. The framework's powerful graphing capabilities were utilized to generate histograms, scatter plots, and other visual representations of the data, aiding in the interpretation of results.ROOT's integration with the C++ programming language and its ability to handle large-scale data processing tasks made it an indispensable tool in this study.

2.1.4 TRExFitter

TRExFitter is a powerful tool used in the statistical analysis of high-energy physics data [8]. It is built on top of ROOT and RooStats and it allows for a wide range of actions, including fitting distributions, calculating significance and exclusion limits, and plotting various outputs by pseudomerging them. In this study, several TRExFitter options were used to perform different tasks in the analysis of the charged Higgs boson search by using a tagged version.

- **`n` Option: Reading Input Ntuples**
  The `n` option in TRExFitter is specifically used to read input Ntuples. This is crucial when dealing with raw data stored in the form of Ntuples, as it allows TRExFitter to process these inputs directly for further analysis. This option ensures that the statistical analysis begins with the appropriate data format, enabling accurate fitting and plotting in subsequent steps.

- **`wd` Option: Creating the Workspace**
  The `wd` option is utilized to create the RooStats XML files and workspace necessary for the fitting process. This step is vital as it prepares the environment in which the statistical fits will be conducted, organizing the data, models, and configurations into a structured workspace.

- **`fl` Option: Fitting the Workspace**
  The `fl` option triggers the actual fitting of the workspace. This action is central to the analysis as it involves optimizing the model parameters to best describe the observed data. By fitting the workspace, TRExFitter extracts key information such as signal strengths, uncertainties, and correlations.

- **`r` Option: Drawing Ranking Plot**
  The `r` option in TRExFitter is used to draw ranking plots. A ranking plot provides a visual representation of the importance of different parameters or components in the fitting process. In this analysis, the ranking plot helps evaluate which parameters or systematic uncertainties have the most significant impact on the fit results.

- **`pas` Option: Plotting Post-Fit Results**
  The `pas` option is used to plot the post-fit results. This option visualizes the fitted model compared to the data, offering insight into the performance of the fit. Post-fit plots are critical for assessing model-data agreement and presenting the final results of the analysis.

3 Analysis

3.1 Preselection

Initially, the preselection criteria is applied to an MC-simulated dataset. These criteria are designed to enhance the proportion of tbH^+ events within the signal region while retaining as many events as possible to minimize statistical uncertainties. The pt cut on the lepton in this analysis is of 10 GeV. The discriminating function for preselected events for the 1500 GeV mass is shown in Figure 4 for v08 (left) and v09 (right), with a cumulative histogram for the various background contributions and the signal.

3.2 Small Ntuples Production

The small Ntuples are compact datasets containing a subset of the information from the full v09 Ntuples, tailored for specific analyses. A Python virtual environment containing all required libraries was used for the production of the small Ntuples from the pre-existing necessary data. In cases where the virtual environment did not function properly, the environment was completely removed and re-created from scratch, ensuring a clean setup.

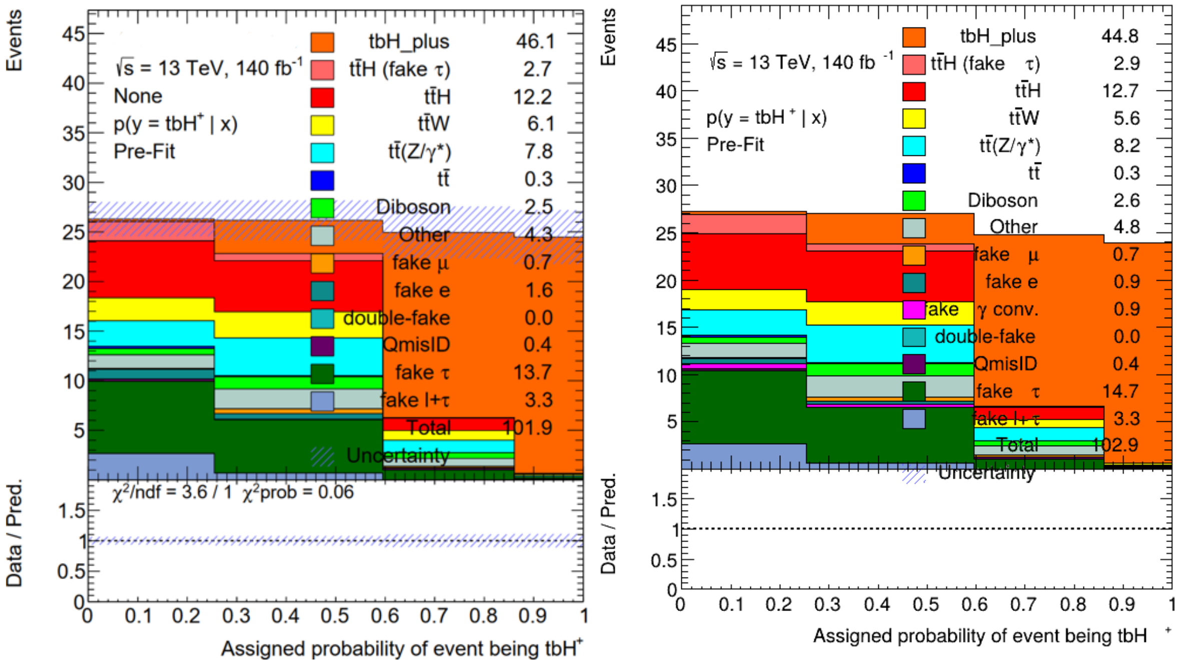

Figure 4: Contributions to the preselected MLP discriminating function for preselected events for the 1500 GeV v08 (left) and v09 (right) signal, showing close similarity. The assumeed signal cross-section value is 0.1 pb.

The script `produce_small.py` was then executed to generate the small Ntuples. After production, the file size and contents of the newly produced v09 small Ntuples were carefully compared to the v08 small Ntuples to verify consistency and correctness.

3.3 Friend Ntuples Production

The friend Ntuples are auxiliary datasets containing additional variables or derived quantities that complement the main Ntuples which are critical for the application of advanced analysis techniques, including ML.

In this phase, a MLP ML model was employed to enhance the classification of events by executing the script `generate_friend_ntuples.py`. The resulting friend Ntuples, containing the outputs of the MLP classifier, were subsequently used in the final analysis to improve the sensitivity to the charged Higgs boson signal.

4 Results

As an example, the presented preselection plot uses the old 1500 GeV signal mass point. Preselection plots, ranking plots, pulls, correlation factors, and normalization factors for all mass points are provided in the appendix of reference [9].

4.1 Preselection Plots

The preselection plots provide an initial visualization of the data before applying specific event selection criteria. These plots present the results obtained after training the models and evaluating their performance. In this analysis a cross-section of 0.1 pb is implemented. While the preselection results for other mass points are available in reference [9], Figure 4 shown a comparison of the v08 and v09 1500 GeV mass point. Preselection refers to the numbers in the plot and the hatched area in the plot represents the statistical uncertainties. These results provide valuable insights into the initial data distribution and the effectiveness of the machine learning models at this stage. Only small differences are determined between v08 and v09 distributions.

4.2 Limits

This subsection focuses on the expected upper limits at 95% CL set on the signal cross-section as a function of mass, using the asymptotic limit type. Using the statistical tools employed in this analysis, the expected limits are calculated. Table 2 presents the expected asymptotic upper limits considering the statistics only of all seven mass points. The values presented here serve as a baseline for understanding the limitations of the current analysis in detecting new signals.

	v08 Limit [pb]	**v09 Limit [pb]**
New 250 GeV	0.0987	0.0939
Old 300 GeV	0.0776	0.0825
Old 800 GeV	0.0168	0.0183
New 800 GeV	0.0168	0.0161
Old 1500 GeV	**0.0110**	**0.0115**
Old 2000 GeV	0.0209	0.0214
New 3000 GeV	0.0364	0.0356

Table 2: Expected asymptotic upper limits of statistics only (StatOnly: TRUE).

Table 3 lists the expected asymptotic upper limits considering the statistics and systematics, of all seven mass points. This table emphasizes the impact of systematic uncertainties alongside statistical considerations. Notably, the limits obtained with the v09 Ntuples are only slightly weaker, as determined by the 1500 GeV mass point. By including both statistical and systematic uncertainties, this table provides more accurate values of the expected upper limits.

	v08 Limit [pb]	**v09 Limit [pb]**
New 250 GeV	0.1128	0.4020
Old 300 GeV	0.0921	0.2675
Old 800 GeV	0.0180	0.0309
New 800 GeV	0.0180	0.0299
Old 1500 GeV	**0.0116**	**0.0119**
Old 2000 GeV	0.0231	0.0279
New 3000 GeV	0.0381	0.0712

Table 3: Expected asymptotic upper limits of statistics and systematics.

5 Conclusions

In this project, progress was made in the challenging search for the charged Higgs boson particle by the inclusion of the contributions of the new systematic in v09 Ntuples datasets. The incorporation of additional systematic in v09 provided a better understanding of uncertainties associated with tau leptons, enhancing the accuracy and reliability of the results. Notably, the limits, using the recent version with v09 are only slightly weaker than those obtained in the previous version.

6 Suggestions for Future Work

Future work shall focus on exploring different binning strategies and compare the asymptotic limits produced with toy limits. Furthermore, the need for new MC files is critical to update the limit plots across the full mass range of the charged Higgs boson. These updates are essential for refining the analysis and providing more comprehensive results for the entire mass spectrum.

7 Acknowledgements

I would like to express my sincere gratitude to my supervisor André Sopczak for his consistent guidance and mentorship during the summer student project. I would also like to thank the Conseil Européen pour la Recherche Nucléaire and the CERN & Society Foundation for their support in making this summer student project possible. A special acknowledgment goes to Janick Böhm for generously sharing relevant scripts and invaluable knowledge, as well as for his unwavering patience and guidance.

I would also like to warmly acknowledge the discussions, advice, and assistance provided by the tbH^+ working group, whose support has been instrumental in the success of this project.

References

1. *The Higgs Boson* https://home.cern/science/physics/higgs-boson.
2. ATLAS Collaboration. *ATLAS : a 25-year insider story of the LHC experiment* (Singapore : World Scientific, 2019. - 359 p, 2019).
3. Stål, O. *Prospects for Charged Higgs Discovery at Colliders* tech. rep. (Stockholm University, 2017).
4. Pospísil, J. *Application of Machine Learning for the Charged Higgs Boson Search Using ATLAS Data* MA thesis (Czech Technical University, Prague, 2022). https://cds.cern.ch/record/2812372?ln=en.
5. Rameš, M. *Search for tbH+ with Performance Optimisation for signal and Background Separation Using Machine Learning with ATLAS Data* MA thesis (Czech Technical University, Prague, 2023). https://cds.cern.ch/record/2862250.
6. *Ntuple— ATLAS Open Data* https://opendata.atlas.cern/docs/documentation/data_format/ntuple.
7. Dertat, A. *Applied Deep Learning - Part 1: Artificial Neural Networks* https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6.
8. *TRExFitter Documentation* https://trexfitter-docs.web.cern.ch/trexfitter-docs/.
9. Galea, S. & Sopczak, A. *Charged Higgs boson study with $H^+ \to WH$ and $H \to \tau\tau$ Decay Mode in Proton-Proton Collision Simulation – Summer Student Project* tech. rep. (CERN, Geneva, 2024). https://cds.cern.ch/record/2917268.