Full systematics study for the search of a Higgs boson produced in association with a single top using the ATLAS detector - Summer Student Report

Jassimar Singh

CERN Summer Student 2023

Supervisor: André Sopczak

IEAP CTU in Prague

Abstract

The Higgs boson coupling with top quarks is of great interest in particle physics research. The particular case of a tH final state from proton-proton collisions is a very rare phenomenon where the number of signal events is much smaller than the number of background events. Previous work on this topic used machine learning algorithms for the tH analysis to better separate the signal events from the background events in order to maximise the sensitivity for the signal process. This report extends the previous work using a limited number of background processes and partial systematic uncertainties to a full background and full systematic uncertainty analysis.

1 Introduction

Owing to the top quark being the heaviest elementary particle in the Standard Model of particle physics, it is expected to have a strong interaction with the Higgs field. Thus, one of the large achievements of Higgs boson research was investigating the top quark pair coupling with the Higgs: the ttH process. This process is very rare and only accounts for about 1% of Higgs boson productions. Almost a factor 10 more rare is the single top quark and Higgs boson production - the tH process. Figure 1 shows the corresponding Feynman diagram. This project focuses on the $H \to bb$ channel.

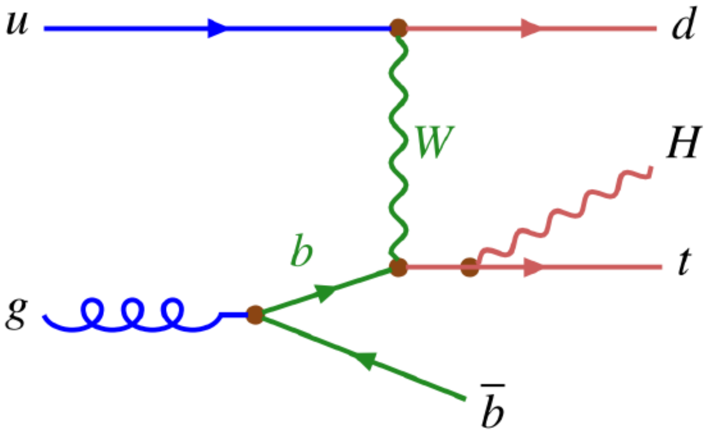

Figure 1: Feynman diagram of the tH process [1].

This report extends the work by a previous masters student - Martin Vatrt [1] on the data analysis of this process.

2 Reproduction of previous work

The first aspect of this project was to reproduce the final plots and results in Martin Vatrt's work. He had conducted a machine learning analysis on the $tH(H \to bb)$ process to obtain a high sensitivity of the tH target process. After determining the optimal machine learning models, he used the TRExFitter framework [2] to perform a statistical analysis. The first aim at this stage is to use this framework to reproduce his statistical analysis.

Martin's analysis had two main aspects: purely statistical effects, and statistical and some systematic uncertainty effects. The systematic ntuples were not ready in time for the completion of his thesis so while the purely statistical analysis used the full background (section 3.2 [1]), the inclusion of systematic uncertainties only utilised a few main background sources - ttb, ttc and ttL where the tt systematics were implemented.

Using the relevant TRExFitter config files and after modifying the ntuple paths, the same analysis was reproduced. The analysis generates a variety of plots that are all presented in the internal note [3]. This report will only highlight the key outcomes such as Figure 2 which shows the signal and background yields after implementation of the neural network model.

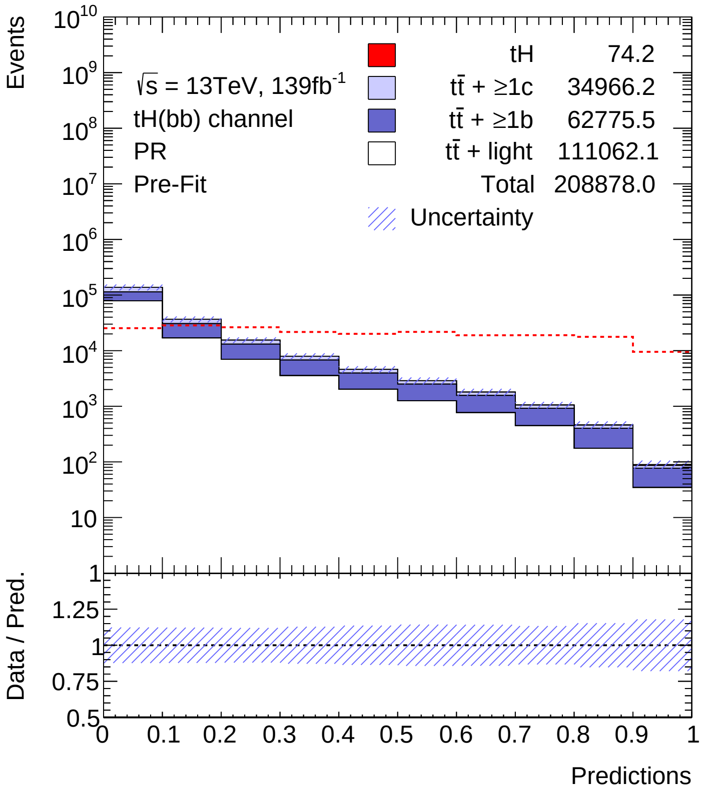

Figure 2: Reproduction of signal region prediction plot in Martin Vatrt's thesis [1].

Comparison of Figure 2 with Figure 12.1 in Martin Vatrt's thesis [1] shows that the plots are the same. In fact, every reproduced plot is almost the exact same as the corresponding plot in Martin Vatrt's thesis. There were a few subtle differences caused by the updated TRExFitter framework since Martin completed his thesis but they do not affect the analysis outcomes.

2.1 Reproduction of expected sensitivity

In the conclusion of Martin's thesis, the expected limit at 95% CL (median) of the signal strength μ was 3.89 (statistical uncertainties) and the limit with inclusion of systematic and statistical uncertainties was 6.35. The former will be discussed later in the full background analysis meanwhile, the latter was able to be replicated exactly as shown in Figure 3.

```
Correct bands
+2sigma:  12.028
+1sigma:  8.86676
-1sigma:  4.5727
-2sigma:  3.4061
Injected: 0
Median:   6.34607
Observed: -1
```

Figure 3: Expected signal strength limit (median) with the inclusion of statistical and systematic uncertainties (old ntuples), exactly the same values as in [1].

3 New Ntuples

Although the neural network ntuples with full systematics were not produced by the time Martin had finished his thesis, they were ready by the time this project had begun. The aim was to use all the backgrounds and systematics to update Martin's analysis. One noteworthy remark is that recently, the ATLAS collaboration had determined that the LHC Run-2 luminosity was actually 140 fb^{-1} [4] instead of 139 fb^{-1} which increases the Monte Carlo weights (in comparison to Martin's weights) used in the TRExFitter plots, affecting quantities like the yields.

3.1 Reproduction under same conditions

The first step was to use the new ntuples to do the same analysis as Martin and compare the differences. New config files were used that take into account the new file paths to create the relevant plots (given in the internal note [3]).

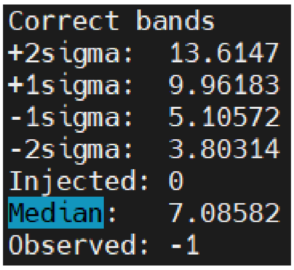

Figure 4: Expected signal strength limit (median) with the inclusion of statistical and some systematic uncertainties using the new ntuples with the same partial background as in Martin Vatrt's thesis [1].

With this new set of plots, many differences with respect to the previous plots are noticeable, however, the outputs of the analysis were generally similar and consistent. Although most plots have experienced some level of change, the difference in individual values themselves is small. For example, the new signal region prediction plot has increased yields by about 1% caused by the previously discussed luminosity update increasing the MC weight value. Additionally, in general, the systematic effects with higher uncertainty in the old ntuples maintain their high uncertainty in the new ntuples' analysis. On the other hand, the largest difference observed between the old and new ntuples' results is between Figure 3 and Figure 4. The expected signal strength limit has increased from 6.346 to 7.086. Given that the background influence is almost identical, this indicates a larger systematic contribution.

The natural next step for the the analysis is to add the full background.

3.2 Full background inclusion

After the partial background analysis was completed and showed consistency with Martin's previous work, the next step was to do the full background analysis. Once again, the provided config files contained the ntuple paths and sample MC weights which were copied into the original config file from Martin. The full background includes: $ttb, ttc, ttL, ttH, ttV, t+Z$, Wt-channel, $t-$channel, $s-$channel, non-top, rare top, non-prompt (fakes). The recorded data was also included, applying a blinding condition. For the statistics and systematic analysis, the fakes (non-prompt events) are excluded[1].

3.2.1 Statistical uncertainty only

The yields are shown in Figure 5. The expected signal strength limit (median) is shown in Figure 6.

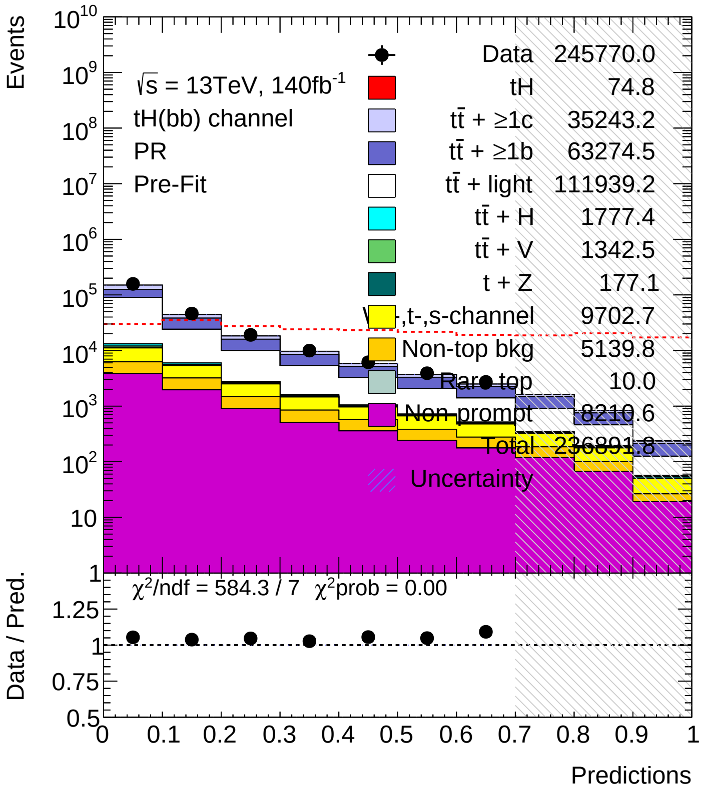

Figure 5: Signal region prediction using the new ntuples with the inclusion of full background. Data events with $s/b > 0.3\%$ are blinded.

```
Correct bands
+2sigma:   7.5784
+1sigma:   5.61296
-1sigma:   2.89763
-2sigma:   2.15838
Injected:  0
Median:    4.02139
Observed: -1
```

Figure 6: Expected signal strength limit (median) with only the inclusion of statistical uncertainties using the new ntuples with the full background.

[1]Initially, the data and fakes were excluded because the corresponding ntuples had not undergone the level 2 sklimming process network but these were updated later in time for the full systematic analysis, much after these initial plots had already been generated.

The signal region prediction plot has changed compared to previous iterations in this report. Note the logarithmic scale, most background has a very small contribution. The expected signal strength limit (median) 4.02 in Figure 6 slightly exceeds the signal strength limit (median) of 3.89 given in the conclusion of Martin's thesis [1]. An increase is expected due to the additional background reactions.

3.2.2 Statistical and systematic uncertainties

The tt systematics used in Martin's analysis were implemented into the analysis (data and fakes excluded). The detailed results are shown in the internal note [3] and the expected signal strength limit (median) is shown in Figure 7.

```
Correct bands
+2sigma:  14.4275
+1sigma:  10.5517
-1sigma:  5.40565
-2sigma:  4.02655
Injected: 0
Median:   7.50206
Observed: -1
```

Figure 7: Expected signal strength limit (median) with the inclusion of statistical and partial systematic uncertainties, no fakes, using the new ntuples with the full background.

There are small difference in the plots compared to the partial background analysis, however, there is a very high level of consistency, similar to previous discussion between old and new ntuples. The individual numbers themselves do not vary significantly and the important systematic uncertainties still stand out and are the same. The signal strength in Figure 7 has increased by about 0.42 which is significant and shows the effect of the added background.

4 Full systematic analysis

Initially, the full systematic analysis was performed without the inclusion of the fake lepton events since the recorded data files did not contain the NN variable being used for the analysis. Then, Martin Vatrt added the NN variable to the files and this allowed the analysis to use the fake (non-prompt) events as well.

4.1 Results

The inclusion of all systematics is a challenge. It requires a lot of memory and CPU time, and thus just running the TRExFitter code in one job would take too long. The appropriate approach is to split up the TRExFitter code into many jobs and then perform those jobs in parallel using HTCondor. The previously written code was updated to intake the files of interest and perform the analysis. Once the root files are generated, they were combined into one 'signal region' root file which is fed into TRExFitter for the analysis. The plots from the analysis are shown in the internal note [3] and the signal region prediction and the expected signal strength limit (median) are given in this report and shown in Figure 8 and Figure 9, respectively.

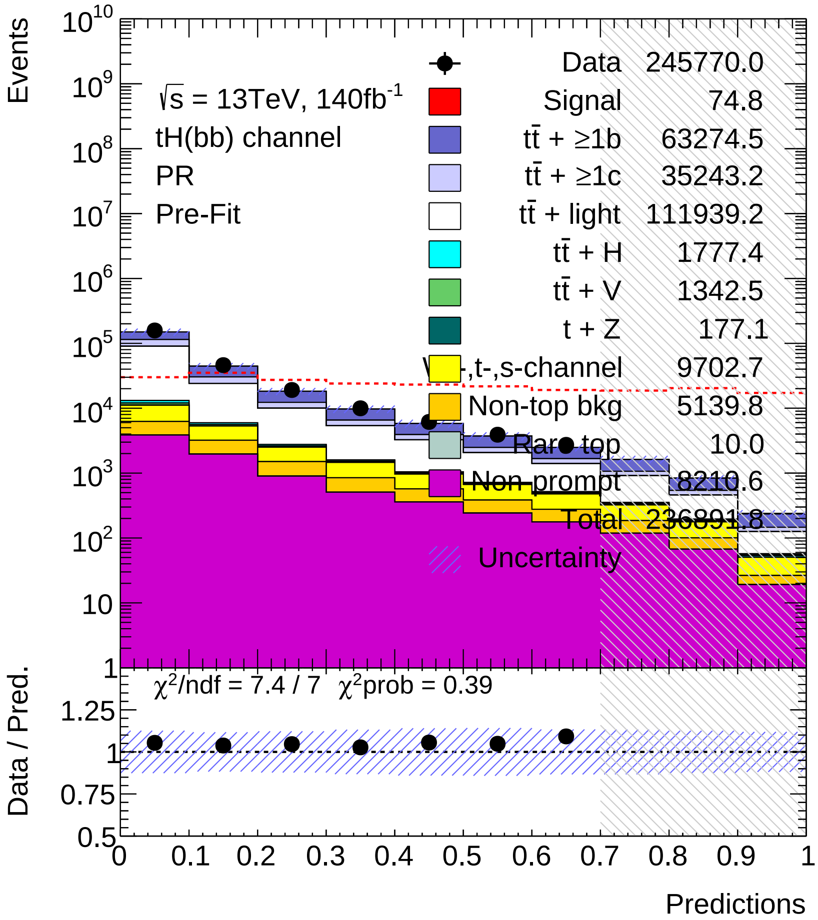

Figure 8: Signal region prediction using full systematic uncertainties. The recorded data is blinded (grey hatched region) if the signal over background ratio in a bin is larger than 0.33%. The uncertainties are indicated in the lower Data/Prediction ratio including statistics and systematics (blue hatched region).

```
Correct bands
+2sigma:   16.2668
+1sigma:   11.5992
-1sigma:   5.87448
-2sigma:   4.37577
Injected: 0
Median:    8.15271
Observed: -1
```

Figure 9: Expected signal strength median using the full systematic run.

4.2 Discussion

There are larger differences in systematic uncertainties and correlations in comparison to the previous analyses. The effect of the inclusion of the full set of systematic uncertainties is very noticeable because several of the new systematic uncertainties show meaningful contribution to the results. Note that while the important systematic uncertainties in the partial systematic analysis continue to play a big role, there are additional new systematic uncertainties from the extended analysis with full systematic uncertainties among the most important ones. This shows that while Martin's previous work had taken into account the majority of important systematic uncertainties, there are several others that contribute significantly.

The expected signal strength limit in Figure 9 demonstrates a decrease in sensitivity. With the inclusion of the full background and full systematic uncertainties, the expected limit increases to 8.15. A summary of the expected signal strength limits is given in Table 1.

Analysis	Ntuple	Background	Systematics	Expected limit
Martin Vatrt	old	full	none	3.89
This study	new	full	none	4.02
Martin Vatrt	old	partial	partial	6.35
This study	old	partial	partial	6.35
This study	new	partial	partial	7.09
This study	new	full	partial	7.50
This study	new	full	full	8.15

Table 1: Summary of expected signal strength limits (medians) for a previous analysis, using new ntuples, a complete set of background processes, and a complete set of systematic uncertainties, including the uncertainties from fake lepton determination with a Template Fit method.

5 Conclusion

This summer student project extended a previous tH analysis. While in the previous tH analysis for the systematic uncertainty study only the most important background and systematic uncertainties were considered, this analysis includes the full set of background processes and all available systematic uncertainties. First, the expected limit on the signal strength at 95% CL was determined with statistical uncertainties only, leading to $\mu = 4.02$. In the previous analysis, including only the most important backgrounds and the most significant systematic uncertainties, the limit was $\mu = 6.35$. This analysis using all background and available systematic uncertainties increased the expected limit to $\mu = 8.15$. The increase resulted mostly from the additional systematic uncertainties.

Acknowledgements

I would like to thank Andre's group and the tHq analysis group for the many discussions. They were of great help to me. Particular thanks go to Martin Vatrt, Mohamed Aly, Nazim Huseynov for their incredible patience with my barrage of questions and the fantastic, clear help in response. And of course, thanks to André Sopczak, my supervisor, for his continuous supply of help throughout the two month period.

References

1. Vatrt, M. *Development, Application and Representation of Algorithms for Discoveries with the ATLAS Forward Proton (AFP) Detector at CERN* Presented 16 May 2019 (2019). `https://cds.cern.ch/record/2690359`.
2. *TRExFitter Documentation* `https://trexfitter-docs.web.cern.ch/trexfitter-docs/`.
3. Singh, J. & Sopczak, A. *Full systematics study for the search of a Higgs boson produced in association with a single top using the ATLAS detector – Summer Student Project* ATL-COM-PHYS-2023-903. Geneva, 2023. `https://cds.cern.ch/record/2875007`.
4. ATLAS Collaboration. *Luminosity determination in pp collisions at* $\sqrt{s} = 13$ *TeV using the ATLAS detector at the LHC* 2022. arXiv: `2212.09379 [hep-ex]`.