technische universität
dortmund

Advanced Laboratory Course

Particle Physics

Search for $t\bar{t}$ resonances with ATLAS data

Koen Denekamp & Riana Shaba

Advisor: Aaron Van der Graaf

May 2025

Contents

1 Introduction

1.1 The ATLAS detector

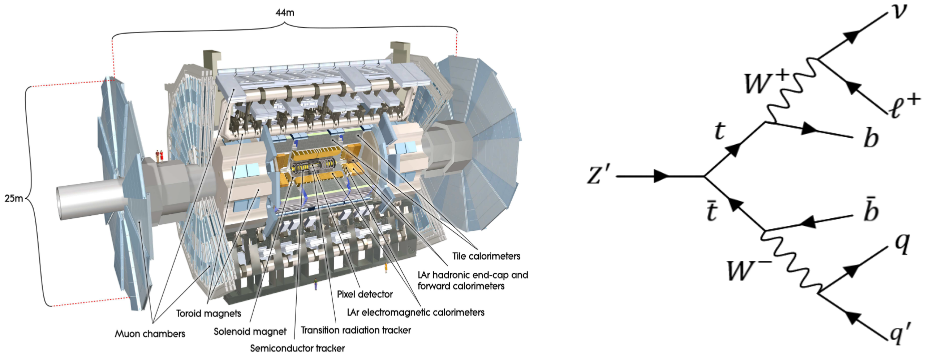
The ATLAS (A Toroidal LHC ApparatuS) experiment is a multipurpose particle detector installed in the experimental cavern Point 1 of LHC at CERN. Its purpose is to search and investigate particle events and prosperously new physics mainly from proton-proton collisions.

An overview of the ATLAS detector is given in Fig. 1a. The detector has four major subsystems. The first subsystem is the Inner Tracker (IT), which is the first area of the detector encountered by particles after the proton-proton collision. The second subsystem is the large magnet, that allows to distinguish positively, negatively and neutrally charged particles, as well to calculate their momentum. The third subsystem are the two calorimeters. One is specifically designed for electromagnetic particles, such as photons and electrons, while the other focuses on hadrons. These two detector systems measure the particle energies. Finally, there is the muon spectrometer that attempts to measure the muons, which pass through the subsystems as they have a very small interaction probability.

Together, these systems attempt to give a full overview of the properties of the particles that pass through it.

1.2 Physics

The goal in this analysis is to search for resonances resulting from a hypothetical massive boson $Z' \to t\bar{t}$ decay, where data is taken from the ATLAS detector. This would indicate Beyond Standard Model (BSM) physics, where the exact mass of the resonance is a free parameter. Taking all the different topologies into account, we are in particular interested on the lepton-jets channel, as these final states are predicted to have much lower Quantum ChromoDynamics (QCD) backgrounds than fully hadronic decays, while still having a significantly larger branching ratio than final states with two charged leptons. A Feynman diagram of the decay is found in Fig. 1b. The resulting quarks will hadronize, so this channel is thus expected to show up in the detector as one charged lepton, and four jets, two of which will be b-tagged jets that are explained in detail in Section 3.1. This will influence the cut criteria later on in this report.



(a) ATLAS detector. Taken from Ref. [1].

(b) Feynman diagram of $t\bar{t}$ production.

Figure 1: An overview of the ATLAS detector subsytem and a Feynman diagram of $t\bar{t}$ production through the BSM Z' boson and its decay in the lepton+jets channel.

2 Data

The data used in this analysis are of two kinds. There is simulated data that is further explained in Section 2.1. And secondly, a data file containing ATLAS data has been provided. A short description of this data is given in Section 2.2. Additionally, there were two example files provided, that follow the same procedure as the files explained in the next two sections, so they are not explained in much detail here.

2.1 Simulated data

The simulated data came as many different files, for different types of background processes and different types of signal.

In our search for a signal from the hypothetical particle Z' boson decaying into a $t\bar{t}$ pair, it is important to account for several known background processes, and in this report we consider five of these. Each of them contributes differently in terms of kinematics and final states, and they are all very important and need to be taken into account for our model because all these processes can be mistaken for the signal that we are interested to find.

The dominant background arises from the Standard Model $t\bar{t}$ pair production itself which produces the same decay products as the signal we are looking for. To pass the upcoming event selection, the decay must match the target channel, which in our study is the lepton + jets channel. The $t\bar{t}$ pair production has the exact topology as the signal, so it passes the selection directly. Single top quark production in the tW channel produces top quarks and b-jets which can be easily mistaken for resembling $t\bar{t}$ final states when both W bosons decay. Since single top events involve only one top quark, they naturally have fewer jets and b-jets. Therefore, passing selection often depends on additional jets from QCD background and on misidentification from b-tagging.

In addition to these, there are processes such as $W + jets$, $Z + jets$ and dibosons, which contribute especially in the leptonic or semi-leptonic decay channels, but don't produce the same final state as the signal. They can typically pass the selection due to detector shortcomings, such as jet misidentification as b-jets, and also because of the high number of jets from QCD background. In $W + jets$ events, the W boson decays into a lepton and a neutrino, as well as jets coming from the QCD background. Hence it can resemble semi-leptonic $t\bar{t}$ decays if the jets are mistaken for b-jets. Similarly, $Z + jets$ events contribute significantly in the dilepton or neutrino decay channels, as they can resemble the dileptonic $t\bar{t}$ decay mode or missing transverse momentum misidentification in the all-hadronic channel. Diboson processes like WW, WZ, ZZ are also background processes because of their production of leptons, neutrinos and jets. These backgrounds can be reduced by applying a better event selection or improving the detector performance.

These files are created using Monte Carlo (MC) simulations, the process of which is beyond the scope of this report. All the files are `ROOT` files, which contain *ntuples* in ROOT's datatype `TTree`, which in turn contain a branch for each variable given.

In total, there were 40 variables, and thus 40 branches given. The branch name, as well as a short description of these variables can be found in Table 1.

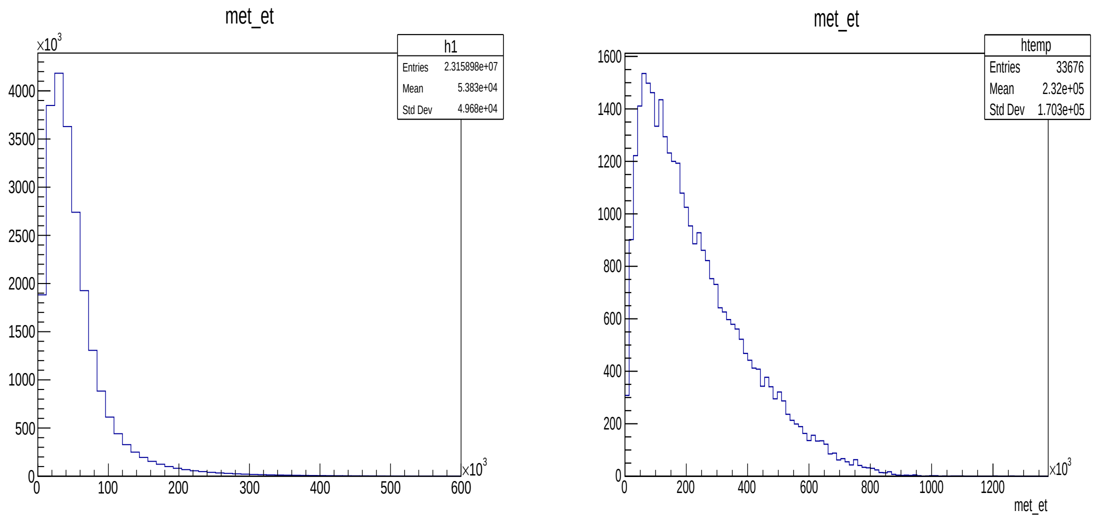
Then there are the simulated signal files. There are 12 of these files, under the name `Zprime<MASS>.root`, where `<MASS>` is replaced with different Z' masses in GeV. This is due to the fact that the Zprime mass is a free parameter in the theory and thus it is unknown. In this report, the masses of 400 GeV, and 500-3000 GeV, with steps of 250 GeV, are considered, and these are thus the data files that are used. These files have different numbers of entries, but as it is seen in Section 3.3, the higher the mass of the Z' boson, the less is the number of entries because less events are expected.

2.2 ATLAS data

The actual data were taken from the ATLAS detector in proton-proton collisions at a center-of-mass energy $\sqrt{s}$ = 13 TeV, corresponding to an integrated luminosity of 1 fb^{-1}.

`runNumber`	A number that identifies the ATLAS data-taking run
`eventNumber`	An event number and run number that combined identify each event uniquely
`channelNumber`	A number that uniquely identifies the simulated dataset in ATLAS
`mcWeight`	The weight of the simulated event
`SumWeights`	The sum of all weights generated by the Monte Carlo process
`XSection`	The total cross-section, taking selection efficiency and higher-order correction factor into account
`jet_E`	The energy of the jet
`jet_MV2c10`	The output of the MV2c10 algorithm, with the goal to b-tag jets
`jet_eta`	The pseudorapidity, often written as η, of the jet
`jet_n`	The number of pre-selected jets
`ket_phi`	The azimuthal angle, often written as ϕ, of the jet
`jet_pt`	The transverse momentum, often written as p_T, of the jet
`jet_trueflav`	The true flavour of the simulated jet
`jet_truthMatched`	Information on whether the jet is matched to a simulated jet
`lep_E`	The energy of the lepton
`lep_charge`	The electric charge of the lepton
`lep_eta`	The pseudorapidity of the lepton
`lep_etcone20`	The scalar sum of the track transverse energy, often written as E_T in a cone of $R = 0.2$ around the lepton
`lep_isTightID`	Information on the satisfaction or not of the tight ID reconstruction criteria
`lep_n`	The number of leptons in the event
`lep_phi`	The azimuthal angle of the lepton
`lep_pt`	The transverse momentum of the lepton
`lep_ptcone30`	The scalar sum of the track transverse momentum in a cone of $R = 0.3$ around the lepton
`lep_trackd0pvbiased`	d_0 of the track associated to the lapton at the point of closest approach
`lep_track0pvunbiased`	d_0 significance of the track associated with the lepton at the point of closest approach
`lep_trigMatched`	Information on whether the lepton triggered the event
`lep_truthMatched`	Information on whether the lepton is matched to a simulated lepton
`lep_type`	A number identifying the lepton type
`lep_z0`	The z-coordinate of the track associated to the lepton with respect to the primary vertex
`met_et`	The transverse energy of the missing momentum vector
`met_phi`	The azimuthal angle of the missing momentum vector
`scaleFactor_BTAG`	The scale-factor for the b-tagging algorithm at a 70% efficiency working point
`scaleFactor_ELE`	The scale-factor for the electron efficiency
`scaleFactor_LepTRIGGER`	The scale-factor for the lepton triggers
`scaleFactor_MUON`	The scale-factor for the muon efficiency
`scaleFactor_PILEUP`	The scale-factor for the pileup reweighting
`scaleFactor_PhotonTRIGGER`	An unused scale-factor for photon triggers
`scaleFactor_COMBINED`	The product of all scale-factors
`trigE`	Information on whether the event passes a single-electron trigger
`trigM`	Information on whether the event passes a single-muon trigger

Table 1: Overview of all branches in the `TTrees`



(a) Simulated diboson background

(b) Simulated Zprime2000 signal

Figure 2: Histograms showing the distribution for the missing transverse energy for two data files.

3 Procedure

In this section, both the procedure, as well as the thought behind some choices that were taken will be considered. The code for this project was written in the `C++` language and `ROOT`, which is a data analysis framework designed for high energy physics, written as well in `C++`.

3.1 Event selection

The first step of the analysis is to remove most of the background in a process known as event selection. In this report we use a simple way to select our events.

We apply cuts to both the true data and the MC data. This is most easily understood by using an example using a histogram. In Fig. 2, we show two of these histograms. The plot in Fig. 2a shows the `MET_et` distribution for the diboson background, while the plot in Fig. 2b shows the `MET_et` distribution for the Zprime signal with a Z' mass of 2000 GeV.

We then want to remove as much of the diboson data, while keeping as much of the Zprime2000 data. Then, for example, if we remove every event that has a `MET_et` value smaller than $100 \cdot 10^3$ MeV, we can greatly reduce the diboson background, while still keeping a decent amount of signal. We can apply similar evaluations on other variables, and combine this with physical considerations.

We decided to make cuts based on four feature criteria. The first is the number of leptons (lep_n). As we want to consider the lepton-jets channel, as explained in Sec. 1.2, we require that there is one lepton. Additionally, we expect four quarks to hadronize and form jets, but we cannot exclude that more jets are created in this procedure. Hence, we require that there are at least four jets in the event. At least two of those four quarks are expected to be b quarks, but we cannot exclude more, so we require at least two $b-$tagged jets.

A jet is said to be b-tagged if the jet_MV2c10 value is larger than a certain value. What that value is depends on the desired efficiency Working Point (WP). In this case, we opted to use a WP of 70%, as this gives a good balance of efficiency, while still retaining a good number of events. This means that the jet_MV2c10 threshold we use to b-tag is 0.83. Hence, we keep events where at least two of the jets have a jet_MV2c10 value larger than 0.83.

Lastly, we require the cut on the missing transverse momentum. This completes our event selection. We consider such a cut, as in the model, the Z' boson will be heavy. This means that the decay products for this particle will most likely be very energetic. The most prominent background, which is the ttbar.root background, will have a similar number of leptons, jets, and $b-$jets, but the expected energy is lower. Hence, the expected missing transverse momentum is expected to be higher in our signal than this background.

This event selection was applied to all the data, signal and background data files.

For the Wjets.root, Zjets.root, diboson.root background files, the btag_count cut is the most impactful. This only kept 0.77%, 1.7%, and 0.69% of these backgrounds respectively. However, since b-jets are already expected in the singleTop.root (keeping 14%) and the ttbar.root (keeping 32.9%) background, here the met_et cut is the most effective. This cut only keeps 9.8% of the singleTop.root background, and only 17% of the ttbar.root background. The lep_n cut is the least impactful in all background files, only removing 50% of background in the best case (which was in Zjets.root.

Afterwards, the most prominent background was the ttbar.root background, which is to be predicted, as it is expected to be the most similar to the signal. Regarding the signal, the btag_count criterion had the biggest impact in reducing the amount of events. For example, in Zprime1000.root this criterion only kept approximately 38.4% of events.

Feature	Description
met_et	E_T^{miss}: The magnitude of the missing transverse momentum.
delta_phi	$\Delta\phi(E_T^{\text{missing}},$ lepton): The difference in the azimuthal angle of the missing transverse energy and the azimuthal angle of the lepton.
invMasse3jets	The invariant mass of the system formed by the three jets with the largest p_T.
invMasse4jetslv	The invariant mass of the system formed by the four jets, the lepton and the neutrino. After selection, this corresponds with the entire invariant mass of the event.
pseudo_rapidity	The pseudorapidity of the system formed by the four jets, lepton and neutrino.

Table 2: Table outlining the high level observable under consideration.

3.2 Determination of final discriminant

So far, we have only considered low level observables to discriminate signal from background. While these have had decent success, we believe that a better discrimination of signal and background is possible. In order to find this discriminant, we take into account five derived quantities, also referred to as high level observables. For this we also revisit the magnitude of the missing transverse energy E_T^{miss}. An overview of these is given in Table. 2.

For each of these features, histograms were plotted for each data and background file, and after a visual analysis inspection, we chose to use `invMasse4jetslv` as our final discriminant. To illustrate this choice, we have shown `invMasse4jetslv` and `pseudo_rapidity` for the `ttbar.root` background and the `Zprime1000` signal files, in Fig. 3 and 4 respectively.

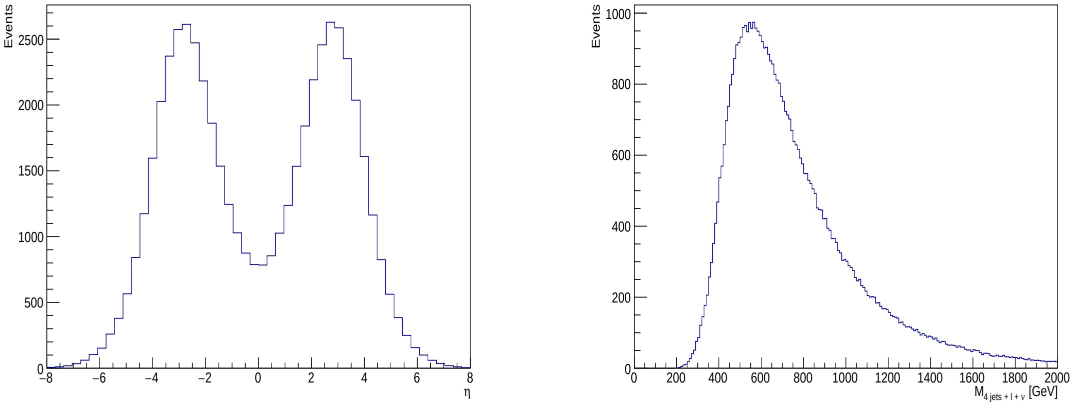
Here we can see that while the `pseudo_rapidity` plots look very similar, there is a clear difference in the distribution for the `inv4Massejetslv` plots.

3.3 Agreement of simulation and data

In this section we discuss the accuracy of the simulated background processes, as we compare the number of events predicted by the Monte Carlo simulations to the number of events observed in the real data after applying the full event selection. This comparison is necessary for ensuring that the simulation correctly models both the detector response and the underlying physics process we are interested in.

After the MC simulation were normalized and weighted properly, the entries of the selected events are already weighted, so they give the expected event counts. These numbers are given in Table 3.

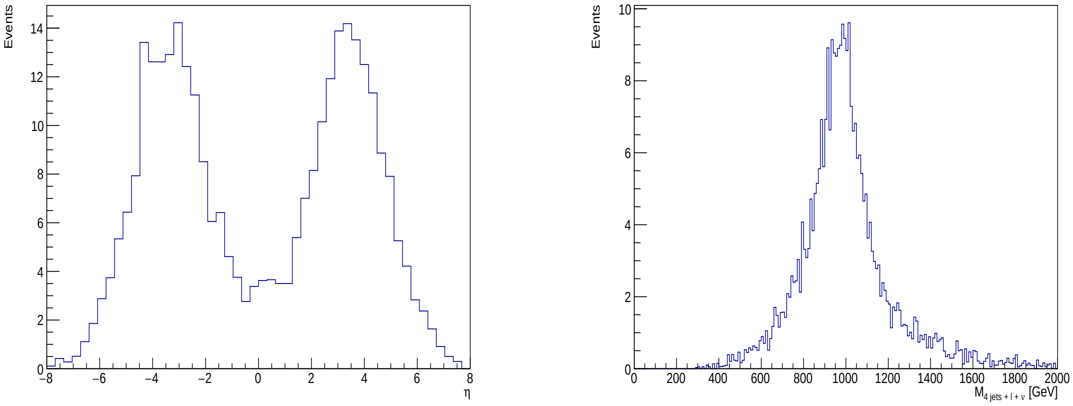
Comparing the sum of all the numbers of background events in simulation with the number of observed events in the data, we see a good agreement between the simulation and real ATLAS data.



(a) `pseudo_rapidity`

(b) `inv4Massejetslv`

Figure 3: Histograms showing the distribution of two high level observables for the `ttbar.root` background data.



(a) `pseudo_rapidity`

(b) `inv4Massejetslv`

Figure 4: Histograms showing the distribution of two high level observables for the `Zprime1000.root` signal data.

Processes	Nr. of events
diboson_selected.root	108.547
singleTop_selected.root	3270.19
ttbar_selected.root	53458.5
Wjets_selected.root	1639.32
Zjets_selected.root	97.3748
Sum of background processes	58573.9318
data_selected.root	60128

Table 3: Number of events for each background process.

Z' events for each mass	Nr. of events
Zprime400_selected.root	216.865
Zprime500_selected.root	486.592
Zprime750_selected.root	581.605
Zprime1000_selected.root	312.021
Zprime1250_selected.root	145.618
Zprime1500_selected.root	65.3515
Zprime1750_selected.root	28.4533
Zprime2000_selected.root	13.0835
Zprime2250_selected.root	5.98606
Zprime2500_selected.root	3.07151
Zprime2750_selected.root	1.52924
Zprime3000_selected.root	0.779865

Table 4: Number of events for each Z' mass simulation.

However, using only comparison of the number of data events and simulation is not sufficient to quantify the agreement between them. There are several points that are very important and need to be considered when making these comparisons. The shape of the distributions, statistical tests and uncertainty (statistical and systematic) evaluations need to be taken into account, instead of only counts.

From Table 4 showing the number of events for each simulated mass of the Z' boson, we see that as the mass of the hypothetical particle increases, the number of events decreases. We also notice that for the Z' mass of 750GeV we have a larger mass, which is where he have a peak. Overall this behavior is exactly what we expect because for much heavier masses, the production of Z' requires much more energy, and at higher masses the cross section decreases significantly. As the mass keeps increasing, there is also less phase space for the Z' to decay into final states that can be detected considering the detector's efficiency and geometrical acceptance, so also the event selection efficiency experiences a drop when even more higher-momentum particles are produced.

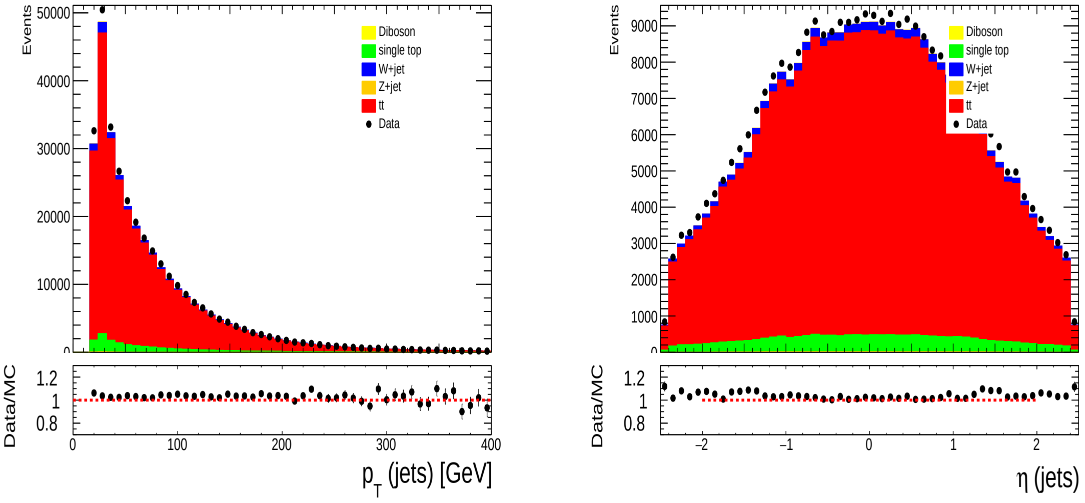
Since simply comparing the counts is not sufficient, we have created stacked plots of the backgrounds, on top of which we have created a scatterplot with uncertainties of the ATLAS data. We have also calculated the ratio $\frac{\text{Data}}{\text{MC}}$, where MC is the amount of simulated background.
If we observe a $\frac{\text{Data}}{\text{MC}}$ ratio that differs from 1 beyond uncertainties, this can be an indication of processes that were not included in the background simulations, and thus of BSM physics.
These plots were first created for the following parameters:

- Lepton p_T (`lep_pt`), η (`lep_eta`), ϕ (`lep_phi`), E (`lep_E`),
- Jet p_T (`jet_pt`), η (`jet_eta`), ϕ (`jet_phi`), E (`jet_E`) of all jets,
- Jet p_T, η, ϕ, E, but only of the jet with the largest p_T per event,
- Number of jets (`jet_n`),
- Number of b-tagged jets (`btag_count`),
- Magnitude of the missing transverse momentum (`met_et`).

We then visually analyzed these plot to see if there were any irregularities that might already point out either new physics, or a mistake in the analysis. We did not find any such artifacts. For illustration, we have added two of these plots in Fig. 5.
We then created the same plot for the high level final discriminant we chose in the previous subsection. This plot can be found in Fig. 6.
Visually analyzing this plot, as well as the ratios, we notice no significant difference between Data and Simulation. There is also no trend. The only obvious difference between bins is that in the bins where there is more data, around $m_{4\text{j},\ell,\nu} = 600$ GeV, the uncertainties are smaller.

3.4 Statistical analysis

Although a visual analysis of a graph is a good way to get a quick overview, to get a definitive answer on whether we have observed new physics or not, we must perform



(a) `jet_pt`

(b) `jet_eta`

Figure 5: Histograms showing the stacked background distribution and comparison to data of two low level observables, as well as the Data/MC ratio for each bin.

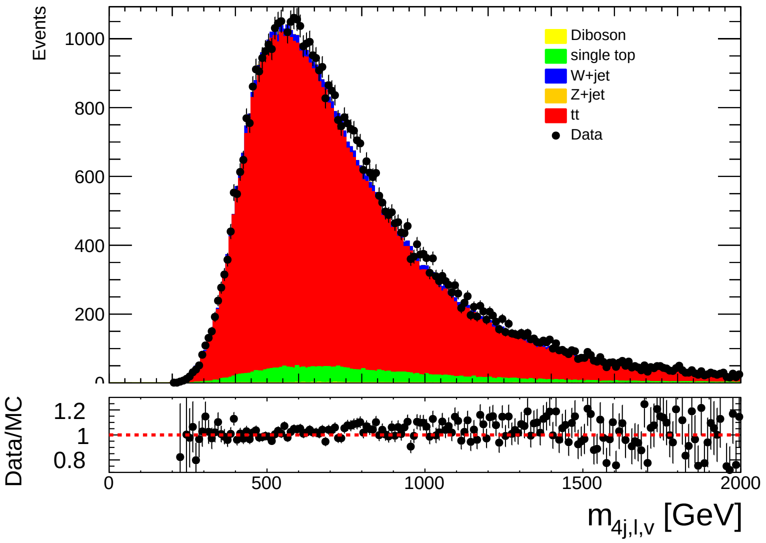


Figure 6: A stacked histogram showing the stacked background distribution and comparison to data of the discriminant `invMasse4jetslv`, as well as the Data/MC ratio for each bin.

Type of uncertainty	χ^2 value	p-value
No uncertainty	244.738	0.01693
14% uncertainty	214.682	0.226619

Table 5: A table showing the results of the statistical tests.

a statistical analysis. To do this, we calculate the χ^2 value between the data and background using all the bins in Fig. 6.
We do this once for the data as we were given it, and once for the data when we introduce a flat uncertainty of 14% (which is the sum of a proposed 4% uncertainty from the luminosity and a 10% estimated systematic uncertainty). The results of these two statistical tests, both with 200 degrees of freedom, are given in Table 5.
As we require a p-value of $2.7 \cdot 10^{-3}$ or less, corresponding to 3σ, to claim evidence of a signal, and an even smaller p-value of $5.7 \cdot 10^{-7}$, corresponding to 5σ, to claim observation of a signal, we can claim neither of these for neither of the uncertainties.
The next step now would be to calculate the 95% CL exclusion limits on the Z' production cross section. While this calculation is beyond the scope of this laboratory report, the main idea behind this analysis is to find the χ^2 value where the probability of the comparison of background and signal with true data is smaller than 0.05. We were given a plot showing the results of these calculations. This is shown in Fig. 7. From this plot we can deduce that we can exclude with 95% confidence, that the Z' boson has a mass below approximately 1750 GeV, as this is where the limits are stricter than the theoretical expectations predict.

4 Conclusions

In this report, we used a combination of simulated background and signal data to create a procedure to follow to look for Beyond Standard Model physics in the $Z' \rightarrow t\bar{t}$ decay channel and to optimized the background-to-signal ratio. This procedure was then applied on real data, and a statistical analysis was performed to quantify the likelihood of Beyond Standard Model physics appearing in our data.

So in the end we could not find any evidence of a possible new intermediate boson Z' predicted by Beyond Standard Model Physics, in a mass range from 400Gev to 3000GeV.

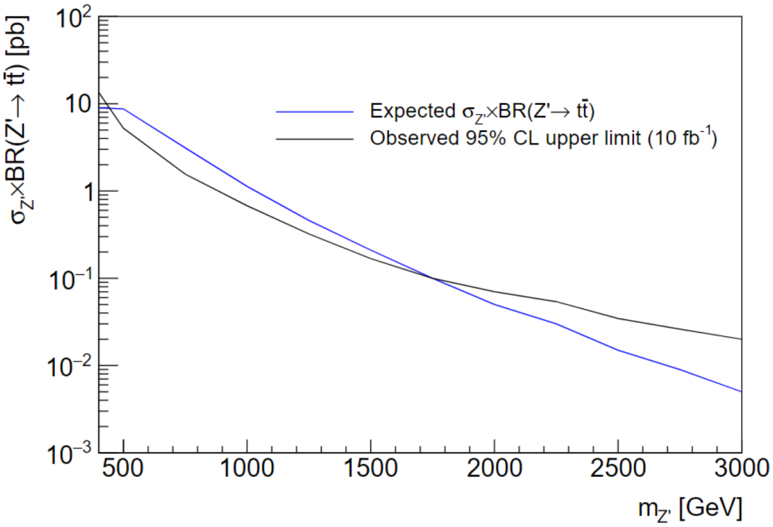


Figure 7: Theoretical limits and 95% CL exclusion limits on the production cross section for different Z' masses. Taken from Ref. [2]

References

[1] ATLAS Collaboration. “The ATLAS experiment at the CERN Large Hadron Collider: a description of the detector configuration for Run 3”. In: *JINST* 19.05 (2024). 233 pages in total, author list starting page 214, 116 figures, 15 tables, published in JINST. All figures including auxiliary figures are available at http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/GENR-2019-02/, P05063. DOI: 10.1088/1748-0221/19/05/P05063. arXiv: 2305.16623. URL: https://cds.cern.ch/record/2859916.

[2] Benedikt Gocke et al. “Search for $t\bar{t}$ resonances with ATLAS data.” In: (2025).