Comparing Simulations for Dark Matter Production in Association with a Single Top Quark at the LHC

Oliver Powell
Advisor: Claudia Seitz
Deutsches Elektronen-Synchrotron DESY
Durham University
United Kingdom

September 4, 2019

Abstract

This report considers the isolation of one lepton decay events from the tW channel of the DMt signal for dark matter production at the LHC for specific samples derived by Monte Carlo techniques. It is of interest to produce sampling techniques that are less computationally expensive than those that involve the full detector simulation, yet contain as much of the same information as possible, for making meaningful predictions about how the real collisions for these events would behave. This could be used to help to validate theoretical models beyond the Standard Model assumed to allow the existence of dark matter and it's association to Standard Model particle physics, such as the $2HDM + a$ model, which is explained in this paper. Methods to discount the background Standard Model interactions are investigated, with selection criteria for the kinematic variables measured for the samples. Tools for comparing samples are introduced, including overlaid histograms and 'cutflow' tables, with an analysis of the differences. It is shown that the chosen Monte Carlo (TRUTH) samples are a promising starting point for replacement of the samples involving full detector simulations (RECO) samples, as the discrepancies in events surviving selection criteria within the different samples are mostly within 5%. Some further work does need to be done however on the finer details as there are some discrepancies, such as differences for the am_{T2} kinematic variable.

Contents

1 Introduction

Within the worldwide physics community, dark matter (DM) continues to be one of nature's leading unknowns driving a combined scientific effort to discover its origin and characteristics. Its existence has been inferred from several mainly gravitational effects: its effect on galaxy rotation curves [1], the cosmic microwave background (CMB) [2] and cause for instances of strong gravitational lensing [3] are some examples.

One of the current leading lines of research into DM is the search for a weakly interacting massive particle ($WIMP$) origin [4]. Particle colliders such as the LHC are used as a tool to collide standard model (SM) particles together, while the products of these collisions are analysed in the hope to find signatures for as-yet-undetected particles. If DM does have a particle nature, it is likely that it would pass through collidor detectors undetected due to its weak interactions. Therefore 'missing energy' is an important calculated variable to account for both standard model particles with the smallest cross-sections, like neutrinos and potential new particles incorporating DM. This is incorporated into the variable E_T^{miss} or 'MET', calculated by considering conservation of energy in the transverse plane to the beam and is further detailed in Section 3.

Run 2 of the LHC concluded in autumn 2018, with work now underway for the next running pariod of the ATLAS and CMS experiments, aiming to reach the intended centre of mass energy $\sqrt{s} = 14TeV$ for Run 3 in 2021. Therefore in context of the DM search, it's important to look for the particular interactions that could produce DM within the already-collected data and to develop a viable search strategy for the same interactions for data obtained in Run 3.

A theoretical model to allow the possible interactions producing DM from collisions undertaken at the LHC is described in section 2. A vital consideration of this analysis is the isolation of the DM-producing events of interest compared to the standard model (SM) background. To do this, several variables measured in the events need to be considered to decide how the unwanted background events can be removed, which is explored in section 3. It's important first to define the theoretical model to make this paper's purpose clear.

2 Theoretical Model

The two Higgs doublet model associated with a pseudoscalar mediator ($2HDM + a$) is incorporated as an extension of the SM. This means the mediator particle that couples to the DM obtains its couplings to the SM fermions from mixing with the second Higgs doublet. A detailed explanation of the 2HDM+a model and why it is used in context of the search for DM can be found in this paper [5]. There are several parameters in this model of interest for determination, notably the mixing angle of the two doublets, α, and the ratio of vacuum expectation values of the two Higgs doublets, $tan\beta$. The mass values of the particles involved in the model, the two charged Higgs $H\pm$, the two neutral Higgs H_0 and h (one of which is the scalar particle discovered in 2012 [6]) and the mediator a, are also important to determine.

By fixing all but two of these variables and exploring the parameter space of the two remaining such as $tan\beta$ and $m(H^-)$, we can simulate collision data that the LHC could produce at its highest intended centre of mass energy of $14TeV$. This study has been undertaken and published in paper [7] as Figure 8. We can exclude certain regions at the 95% confidence level in the effort for determination of these values, as an advisory study at where best to look in the future analysis of collision data.

It is of interest to produce graphs similar to Figure 8 in [7] instead with use of data simulated with full detector modelling, referred to through the rest of this report as reconstructed (RECO) samples. This generating software is very computationally expensive, making it very difficult to adequately explore the full parameter space for exclusion regions. An alternative strategy is to fill gaps in the parameter space with data simulated by other methods that are faster to generate but hopefully retain the information held by the RECO samples. In this report, an example of a less computationally advanced sample technique relying on Monte Carlo generation (referred to through the rest of this report as TRUTH samples) will be compared to the RECO samples for various $H^{+/-}$ and a mass points to deem it's feasibility for use in the way described above and in other search methods.

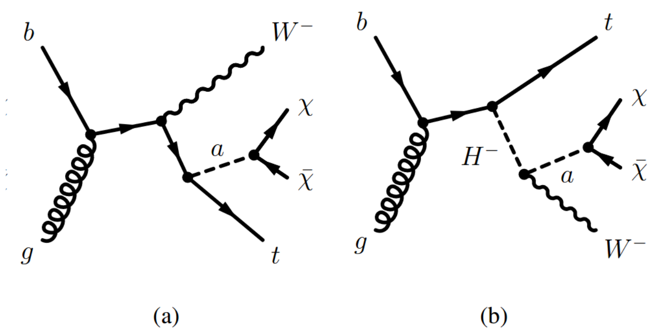


Figure 1: Representative Feynman diagrams for tW channel production of DM (χ) in association with a single top quark and a W boson. These diagrams were taken from paper [7].

The study in [5] finds the mono-Higgs and mono-Z signatures cover a large section of the parameter space of the 2HDM+a model, along with the signature for DM produced in association with a top anti-top pair ($DMt\bar{t}$). This signal in particular is of interest as the $t\bar{t}$ pair kinematics give information about the CP properties of the mediator a. This paper will focus on a variant of this signature producing heavy quarks in the final state, namely DM in association with a single top quark, DMt. This process has been shown in [8] to increase the coverage of current analyses targeting the $DMt\bar{t}$ processes. We will only consider the processes displayed in Figure 1, where both tW channel interactions interfere with each other destructively, ensuring unitarity (that the probability of all outcomes sums to one.) The tW channel is one of the 3 main contributions at leading order in the QCD calculations, the other two not being considered in this study being t-channel and s-channel production.

Within this simulated new physics, a method to isolate the events of interest coming from the tW channel DMt production is required due to the large SM background, as mentioned in the introduction, Section 1. This paper only considers the one lepton decay events, where the neutrino also produced is undetected by the collider. This is achieved by introducing requirements to the sample of events on selected variables that have been measured for each event, for the event to be 'counted' and not removed from consideration. The variables that were selected for this purpose are explained in Section 3. It should be expected if the RECO and TRUTH samples behave in a similar way that a similar amount of each sample is removed after each successive selection criterion towards isolating the tW channel events looked for in Table 2.

3 Kinematic Variables

Firstly the number of b jets above 50GeV is cut for events containing only one b-tagged jet. A complex filter is ran on all events in the sample to work out from the jets present whether they can be traced back to an initial bottom quark, in which case the jet is 'b-tagged'. We select for events that contain one b-tagged jet. By referring to Figure 1, once the top quarks decays after it's short lifetime into a W boson and a b quark, the W is unlikely to decay into a further bottom quark (with an associated top) due to the large top mass.

Second we use the variable for the missing energy in transverse plane to the beam (MET). Using basic mechanics knowledge it should be clear that in the transverse direction all particle energy/momenta should total zero if the only acceleration is parallel to the beam, so it's possible to calculate the 'missing' energy in this plane taking all detected particle momenta into account. The missing energy parallel to the beam cannot however be calculated, since there is no method of accurately knowing the total energy in this direction. MET is made of a combination of neutrino and DM energy, neither of which are detected due to their small cross-sections. We cut on this variable above a level deemed confident as to include both DM and neutrino energy, rather than just the neutrino/chance undetected particle energy of the SM background interactions.

A slightly more involved variable considered is $\Delta\phi_{min}(E_T^{miss}$, 4 leading jets). This is defined as the minimum angle between the MET direction and one of the four leading p_T jets. Cutting on this specifically defined variable has success in taming the secondary backgrounds, like the single or double production of vector bosons, and gets rid of events containing fake MET from calorimeter effects. [9]

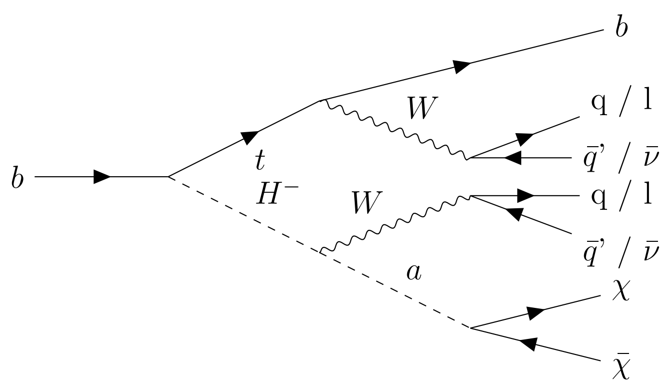


Figure 2: An expansion of the decay products of event (b) from Figure 1. The W bosons decay into either two jets or a lepton and neutrino. χ represents the DM particles.

The number of leptons measured is selected to be one, since these were chosen for analysis over the two lepton events. Furthermore, the number of jets measured in the decay products is selected to be larger than one. From the tW channel decay products shown in Figures 1 and 2, it's clear that to measure the one lepton events there must be more than one jet as decay products.

The transverse mass m_T is a variable built with components of the momenta in the transverse plane of the beam for the measured lepton and the invisible particles (DM and neutrinos.) The m_T is selected so only events around the W mass or higher are kept for analysis. This is designed to reduce SM background from single/double top quark production, as in the SM the m_T comes from the lepton and neutrinos as decay products from W bosons which constrain the overall possible mass/energy. Cutting above this would mostly isolate events with m_T including reasonable contribution from the DM particles.

Lastly, we utilise variables for the combined invariant mass of the leading p_T b-tagged jet and lepton $m(b_1, l)$ and of the leading p_T b-tagged jet leading p_T light quark (non b-tagged) jet $m(b_1, j_{1l})$. Selecting to only keep events in the top mass energy region or higher for $m(b_1, l)$ and the top mass energy region or lower on $m(b_1, j_{1l})$, events with one lepton decay coming from only the charged Higgs rather than the top can be isolated. This can be seen from Figure 2 where the branches for the top quark and Higgs must have asymmetric decay products for the events to produce single-lepton decays, the other branch producing jets. The top branch decay products are constrained by the top mass, while the Higgs branch has no such constraint due to the boson's variable mass. By cutting above the top mass for $m(b_1, j_{1l})$, it ensures the 1 lepton decay comes from the unconstrained Higgs branch, and cutting below the top mass for $m(b_1, j_{1l})$ has the same effect.

4 Method

Sample A	Sample B	Sample C
m(H-)=600GeV	m(H-)=800GeV	m(H-)=800GeV
m(a)=250GeV	m(a)=250GeV	m(a)=350GeV

Table 1: The charged Higgs masses $m(H^-)$ and mediator masses $m(a)$ used in each sample.

Firstly, three TRUTH derivative samples of events for different charged Higgs and mediator masses, $m(H^-)$ and $m(a)$, were converted into a flat ROOT tree in a workable format with the SimpleAnalysis C++ tool [1]. A table showing the different masses is Table 1, these masses were chosen to show how they may effect the event statistics in a clear, simple way. With this tool, it was possible to make an initial 'cutflow' of different variables explained in 3. Cutflow refers to a series of selection criteria applied to the sample where the surviving events are shown in a table. This was to get a feel of how the samples would react to the different criteria before using it as a comparative tool. These flat ROOT trees were fed into Python collating programs called KiSelector and KiPlotter, allowing the three mass samples to be directly compared on the different selection variables in overlaid histograms and cutflow tables.

Next, RECO samples were introduced to begin cross-sample comparisons. These samples reflect the full detector modelling and are compiled into flat trees with different software to the TRUTH

[1]Available at https://gitlab.cern.ch/atlas-phys-susy-wg/SimpleAnalysis

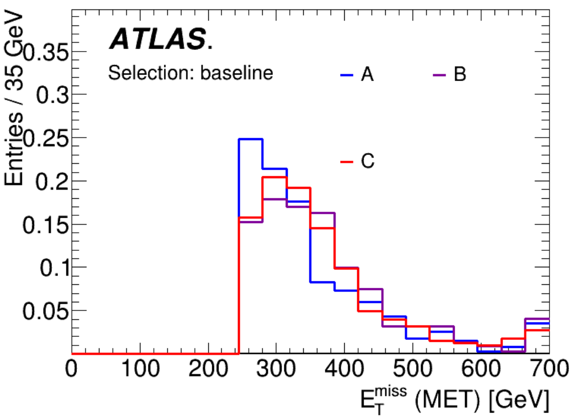
samples, separate to SimpleAnalysis. It is of note that a small preselection was undertaken after the samples were generated unlike for the TRUTH samples, which will need accounting for in cutflow comparisons. Therefore, a baseline selection of cuts was chosen to account for these differences and create a balanced level for fair comparison of later selections chosen to isolate the one lepton product tW channel events. Another Python collating program was used called KiComparer, which created overlaid histograms after each cut in the chosen cutflow to be able to compare the different samples after every step in the process to isolating the one lepton decay events.

5 Results

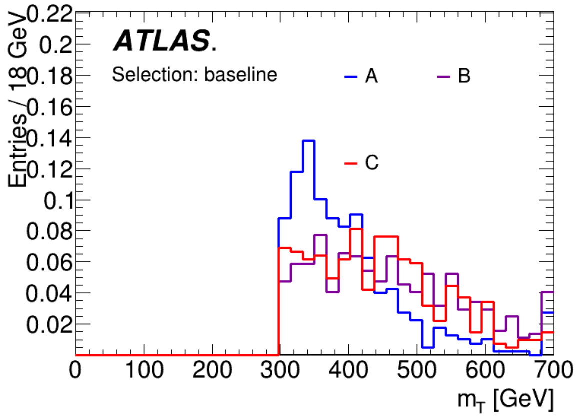
Given below are the various histograms and the cutflow table described in Section 4. This includes the TRUTH sample variables compared for the different mass points in Figure 3, the cutflow table for the one-lepton decay event selection in Table 2, the comparative histograms before the one lepton selection in Figure 4 and after the selection in Figure 5. Comparative histograms of the RECO and TRUTH samples for the specific variable am_{T2} are also given in Figure 6 for the different mass points.

	% of sample size after previous cut					
Sample Masses	m(H$^-$) = 600GeV m(a) = 250GeV		m(H$^-$) = 800GeV m(a) = 250GeV		m(H$^-$) = 800GeV m(a) = 350GeV	
	TRUTH	RECO	TRUTH	RECO	TRUTH	RECO
MET >250	33.8	34.8	52.4	53.6	48.9	48.7
m$_T$ > 300	35.0	33.8	40.7	37.4	43.2	39.3
am$_{T2}$ > 200	78.7	69.2	82.6	74.6	82.7	74.8
m(b$_1$, l) > 160	50.3	56.5	58.7	58.1	54.4	57.2

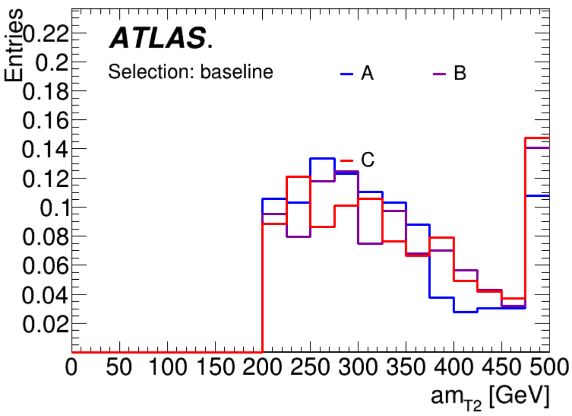
Table 2: The selection of variables chosen to isolate the one lepton decay events in a 'cutflow' table, taken after a preselection detailed in Section 6. The three samples are given in terms of the masses input for their generation. The two methods of generating the samples are given as TRUTH and RECO. The % values represent the size of the sample surviving each selection, compared to the sample size before the selection. It is used rather than bulk numbers because the samples vary in population size.



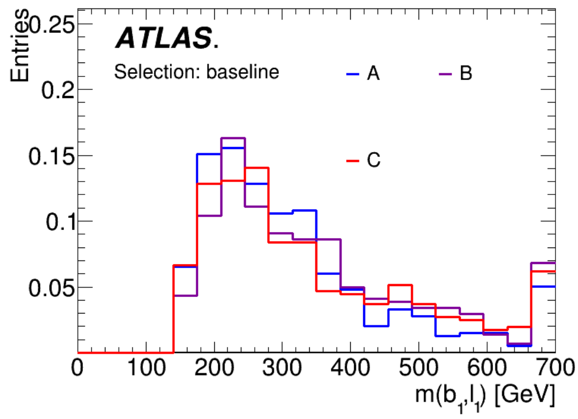
(a) Comparison for E_T^{miss}.



(b) Comparison for m_T.

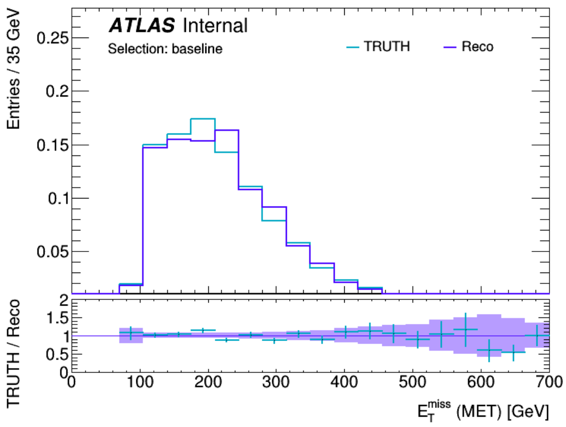


(c) Comparison for am_{T2}.

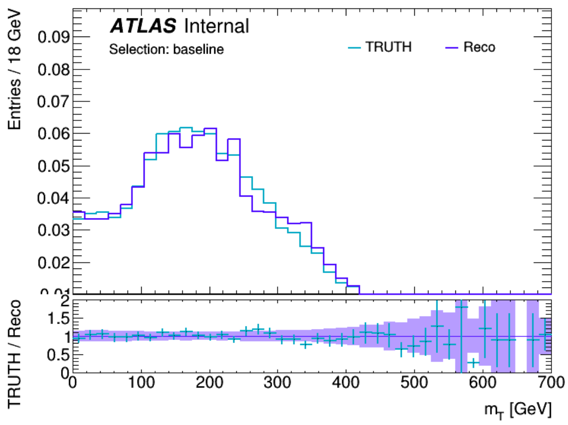


(d) Comparison for $m(b_1, l)$.

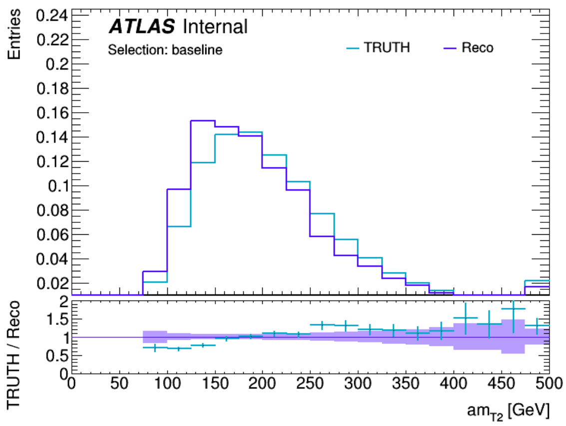
Figure 3: Histograms representing the TRUTH sample variations for four variables. The three distributions represent three different mass samples: 'A' corresponds to $m(H^-) = 600GeV$,$m(a) = 250Gev$, 'B' corresponds to $m(H^-) = 800GeV$,$m(a) = 250GeV$ and 'C' corresponds to $m(H^-) = 800GeV$,$m(a) = 350Gev$. The plots are normalised as only the shape of distributions is important, the sample size is unimportant.



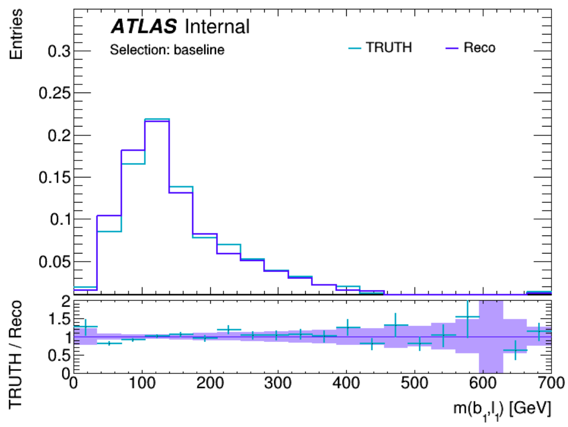
(a) Comparison for E_T^{miss}.



(b) Comparison for m_T.

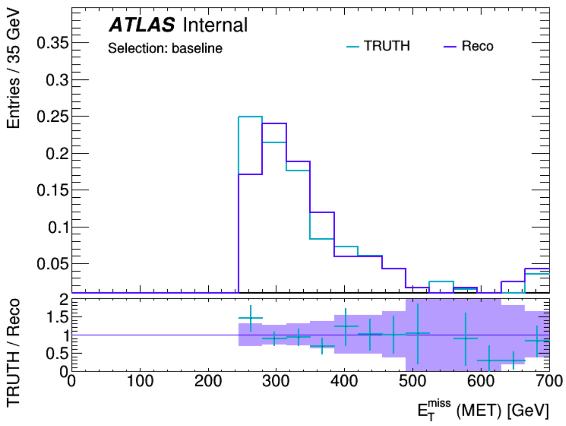


(c) Comparison for am_{T2}.

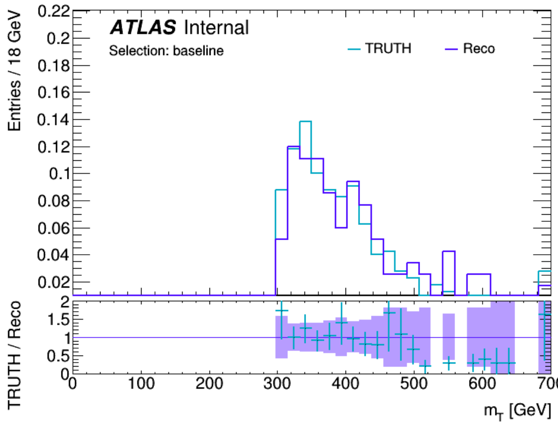


(d) Comparison for $m(b_1, l)$.

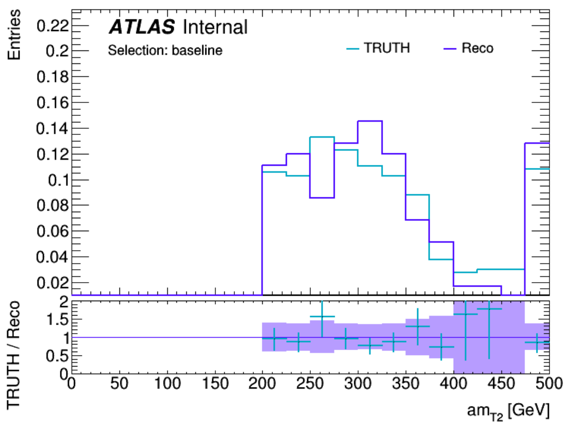
Figure 4: Four comparative histograms of the four variables used in selecting for one lepton decay events. The histograms are again normalised as only the distribution shapes are of interest and each histogram is given with an associated error plot to give more information of the differences compared to the sample size in each histogram bin. These distributions are plotted after the preselection detailed in Section 6, but before the selection for the events of interest.



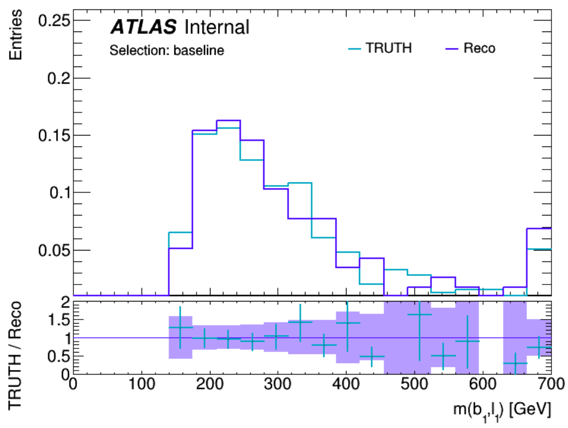
(a) Comparison for E_T^{miss}.



(b) Comparison for m_T.

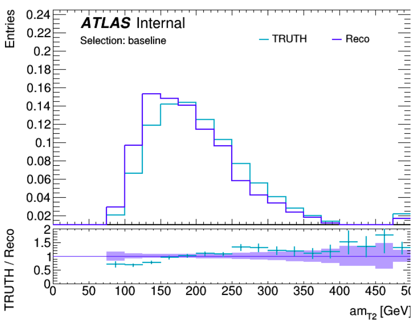


(c) Comparison for am_{T2}.

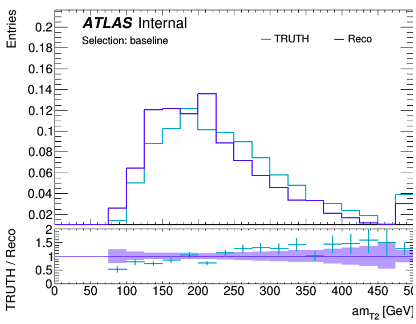


(d) Comparison for $m(b_1, l)$.

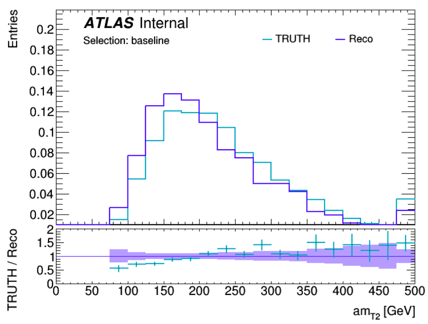
Figure 5: Four comparative histograms of the four variables used in selecting for one lepton decay events. The histograms are again normalised as only the distribution shapes are of interest and each histogram is given with an associated error plot to give more information of the differences compared to the sample size in each histogram bin. These distributions are plotted after both the preselection and the selection for the events of interest.



(a) Comparison for sample A.



(b) Comparison for sample B.



(c) Comparison for sample C.

Figure 6: Three comparative histograms for the variable am_{T2} for the samples of different H^- and a masses. The masses for each sample are detailed in Figure 3. The samples are all derived after a preselection cut, similar to Figure 4.

6 Discussion

The histograms in Figure 3 show an expected higher distribution in energy for the samples generated with the higher Higgs mass (samples B and C,) with more energy from the Higgs likely to transfer to the decay products. This is much less obvious in Figure 3, plot (d) for $m(b_1, l)$, due to the constraint of the top quark mass as mentioned in section 3 for the variable. There is still however a small difference as samples B and C are still spread at slightly higher average energy than sample A. These differences are promising as an initial sanity check, the TRUTH samples straight from generation uphold expectations from physical arguments. As the interest of this paper is of optimising the TRUTH samples to be comparable to full detector simulations, the RECO sample variations isn't of importance (as we know these behave as physically expected on generation) so haven't been considered.

Table 2 represents the selection cutflow after a baseline selection to isolate the one lepton product tW channel events. It is filled with comparative percentages against the sample size before each cut because the sample sizes are all different, so cannot be compared in bulk number. The baseline selection used in order of selection criterion, containing all variables described in section 3 with reasons for each are: number of b-tagged jets = 1, missing transverse energy (MET) > 100GeV, number of leptons detected = 1, $\Delta\phi_{min} > 0.4$ and number of jets detected above 20GeV > 1. Again these aren't present in the table as they don't (and aren't) expected to give any relevant information for this study, other than to ensure a level that gives the further selection a fair comparison between the TRUTH and RECO samples due to the RECO sample preselection.

It's clear from Table 2 that the two sample derivations show good agreement; for most comparisons of the two sample derivations the discrepancy is within 5%, other than am_{T2} which is within a slightly weaker 10%, and $m(b_1, l)$ which is somewhere between, with only one of the mass points (sample A) having agreement weaker than 5%. This shows promise as a candidate to replace RECO samples, as it suggests the sample behaves in a similar way after a similar preselection cut, which could lead to their use after the finer details are optimised, such as the Gaussian smearing. For am_{T2}, it was worth looking at the sample distributions for the different mass points to investigate any notable differences that would explain the discrepancy.

The am_{T2} plots in Figure 6 clearly show a higher energy distribution for the variable in the blue TRUTH samples across all three of the mass points. This suggests the discrepancy is mass-independent; one of the methods to fix this could be to introduce some form of linear correction to the generating software in TRUTH not involving the masses of H^- or the mediator a. However, this would likely also affect the distribution for the other measured variables, so a larger analysis after correction would have to be taken to account for these side-effects. The am_{T2} variable is very niche to the one lepton analysis, as explained in Section 3, so this considerable amount of work for little gain is likely not important unless there is particular interest in the variable. The fact that all the other variables have agreement within 5% (with exception for one sample) suggests a small difference between how the TRUTH and RECO samples are generated as to only affect am_{T2}, one suggestion is any difference in how the b-tagged jets are counted, which would require further investigation.

It is also useful to compare the distributions of the TRUTH and RECO samples for the selected variables before the one lepton selection in Figure 4 and after in Figure 5. The MET and m_T distributions are largely similar for the two simulation techniques in both figures, the error plots are also scattered with no clear trend above or below a ratio of 1. The am_{T2} and $m(b_1, l)$ variables however show a slight discrepancy where the TRUTH samples give slightly higher distributions over the histogram bins, also visible in the error plots where the ratio plots are more often above 1. This difference largely disappears in Figure 5, showing most of the TRUTH generated events with higher-than-average values in plots (c) and (d) of Figure 4 are removed after the full selection. More investigation should be done on the reason for this and whether the MET and m_T selections have an explainable effect, or whether it's just because of the small overall sample populations after the full selection, which is unlikely to produce obvious data discrepancies.

7 Conclusion

To conclude, over the course of this paper differently generated particle collision event samples were compared using various tools such as a 'cutflow' table (Table 2) and various overlaid histograms of the samples measured by particular variables in Figures 3,4-6. This was done as an early check of Monte Carlo generated TRUTH samples to see if they are a feasible replacement of samples generated with the full detector simulations, RECO samples, as less computationally expensive replacements. This would fill in some data sampling gaps in studies like that in Figure 8 of [7], in the specific context of one lepton decay events from the tW channel of the DMt signal for dark matter production in association with a single top quark, as detailed in [7]. These events were selected with specific criteria from the bulk data samples, properly detailed in Section 3 and Section 6.

The resulting comparative studies show potential for the TRUTH samples to be used in replacement of RECO samples after further development on the finer details of the generating techniques is taken. Firstly, the TRUTH samples agree with physical arguments in Figure 3, where there is a discernable difference in the histogram distributions such that the higher $m(H^-)$ values have slightly higher distributions for energy-related values like MET or m_T. Table 2 shows good agreement within 5% for almost all the selection variables relevant to the one lepton analysis apart from am_{T2}, which is evidence that the samples behave similarly when the preselection on the RECO samples is taken into account. Suggestions for fixing the minor discrepancy suggested in this paper include introducing a linear shift in the generating code in TRUTH, however this would require further study into the effect this would have on the other variables of interest measured for the samples. Due to the fact am_{T2} is only relevant to the one lepton analysis, it's unlikely that this difference is worth the considerable amount of further work.

8 References

[1] Albert Bosna. “The Distribution And Kinematics Of Neutral Hydrogen In Spiral Galaxies Of Various Morphological Types”. PhD thesis. NASA/IPAC Extragalactic Database: Groningen University, 1978.

[2] Y. Akrami et al. “Planck 2018 results. I. Overview and the cosmological legacy of Planck”. In: (2018). arXiv: 1807.06205 [astro-ph.CO].

[3] Douglas Clowe et al. “A direct empirical proof of the existence of dark matter”. In: *Astrophys. J.* 648 (2006), pp. L109–L113. DOI: 10.1086/508162. arXiv: astro-ph/0608407 [astro-ph].

[4] Jaco de Swart, Gianfranco Bertone, and Jeroen van Dongen. “How Dark Matter Came to Matter”. In: (2017). [Nature Astron.1,0059(2017)]. DOI: 10.1038/s41550017-0059,10.1038/s41550-017-0059. arXiv: 1703.00013 [astro-ph.CO].

[5] Martin Bauer, Ulrich Haisch, and Felix Kahlhoefer. “Simplified dark matter models with two Higgs doublets: I. Pseudoscalar mediators”. In: *JHEP* 05 (2017), p. 138. DOI: 10.1007/JHEP05(2017)138. arXiv: 1701.07427 [hep-ph].

[6] Georges Aad et al. “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. In: *Phys. Lett.* B716 (2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex].

[7] Priscilla Pani and Giacomo Polesello. “Dark matter production in association with a single top-quark at the LHC in a two-Higgs-doublet model with a pseudoscalar mediator”. In: *Phys. Dark Univ.* 21 (2018), pp. 8–15. DOI: 10.1016/j.dark.2018.04.006. arXiv: 1712.03874 [hep-ph].

[8] Deborah Pinna et al. “Single top quarks and dark matter”. In: *Phys. Rev.* D96.3 (2017), p. 035031. DOI: 10.1103/PhysRevD.96.035031. arXiv: 1701.05195 [hep-ph].

[9] B. Meirose. “Fake E-slash(T) from Calorimeter Effects”. In: 2009. arXiv: 0909.4152 [hep-ex].