Analysis Sensitivity Test in the $2\ell SS + 1\tau$ Channel for Asymmetric Leptoquark Production Mechanism

Casey Hampson
Department of Physics
Kennesaw State University, USA

André Sopczak
University of Prague, Czechia

Abstract

In this report, we briefly describe the phenomenology behind a novel method for pair producing asymmetric leptoquarks (LQ's), in which the final state LQ's are not charge conjugates of each other. Then, we outline the process of testing such a model from within ATLAS's Athena software, generating validation plots, and submitting offical Monte Carlo (MC) samples from the computing grid. From there, we describe the analysis done for the previous release (Rel.) 21 ntuples in the $2\ell SS + 1\tau$ channel with $t\bar{t}H$ production as signal, and how it was ported to the current Rel.22 ntuples. A case study was done with the $t\bar{t}H$ samples to test the new pipeline. All that remains is to input the LQ signal once it is produced.

CERN Summer Student Project Report
August 09, 2024

Contents

1 Introduction

The Leptoquark (LQ) is a proposed Beyond the Standard Model (BSM) particle that can couple to both leptons and quarks, thus allowing one to transform into another. Such a particle has been theorized for decades, where it was first found in many Grand Unification Theories (GUT) such as that by Georgi and Glashow in 1974 [5]. Experimental interest in LQ's has been generally unremarkable; however, in recent years, due to the increasing energies of particle collisions at the LHC giving rise to newer discoveries, interest in the LQ search has increased, as they can provide explanations for a number of discrepencies between Standard Model (SM) predictions and experimental results.

Once such example is the decay of the B-meson [7], in which a flavor-changing neutral current process such as $b \to s\ell\ell$ deviates from Standard Model (SM) predictions, signalling a violation of Lepton Flavor Universality (LFU). Another example is that LQ's can radiatively generate Majorana neutrino masses [8]. Clearly, LQ's are very attractive sources of new physics.

LQ's are found as either vector or scalar particles; we will focus solely on the scalar LQ's, as the novel pair production method we will introduce in this report has not been studied for the vector LQ's. There are five multiplets with varying representations under the standard model gauge group $SU(3)_c \times SU(2)_L \times U(1)_Y$, as shown in Table 1. Notably, all five are triplets under $SU(3)_c$, which is expected, as they couple to a quark and must therefore carry color. This also means that it is possible for there to be quark-quark interactions with a LQ, however we will not consider this for a similar reason as to why we aren't considering vector LQ's. Each multiplet has a number of charged eigenstates; the R_2 leptoquark, for instance, has two charged eigenstates: $R_2^{+5/3}$ and $R_2^{+2/3}$. This comes from the LQ's representation under the $SU(2)_L$, along with, the Gell-Mann-Nishijima formula which, in our normalization, given by:

$$Q = I_3 + Y, \tag{1}$$

where Y is the hypercharge of the LQ and I_3 is the third component of the weak isospin of the LQ.

As a scalar particle, the LQ's couple to the quark and lepton via a Yukawa interaction, with the strength of the interaction determined by the magnitude of the coupling constant y. However, here, the Yukawa coupling is actually a 3×3 matrix, whose indices correspond to the fermion generation. As an example, the $LQ - q - \ell$ interaction part of the S_1 Lagrangian is given here [1]:

$$\mathcal{L}_{\text{int}} = Y^{RR}_{1,ij}\bar{u}^c_i \ell_j S_1^\dagger + Y^{LL}_{1,ij}\left(\bar{Q}^{c\intercal}_i i\sigma_2 L_j\right) S_1^\dagger \tag{2}$$

Here, Q_i and L_i are left-handed quark and lepton $SU(2)$ doublets, and u_i and ℓ_i are right-handed $SU(2)$ singlets. The superscripts c and $\intercal$ mean charge conjugation and transposition of the $SU(2)$ doublets, respectively. The parentheses in the second term denote contraction in $SU(2)$ space. Note that there are two terms, corresponding to the two different coupling types for the S_1 LQ, one for coupling of left-handed quarks to left-handed leptons, and one for coupling of right-handed quarks to right-handed leptons. If we were to, for instance, select $Y^{RR}_{11} \neq 1.0$ and all others to zero, then the LQ would couple only right-handed, first-generation quarks to right-handed, first generation leptons (which is the electron, since there are no

$SU(3)_c \times SU(2)_L \times U(1)_Y$	Symbol	Q-L Chirality	F
$(\bar{\mathbf{3}}, \mathbf{3}, 1/3)$	S_3	LL	-2
$(\mathbf{3}, \mathbf{2}, 7/6)$	R_2	$RL,\ LR$	0
$(\mathbf{3}, \mathbf{2}, 1/6)$	$\tilde{R}_2$	RL	0
$(\bar{\mathbf{3}}, \mathbf{1}, 4/3)$	$\tilde{S}_1$	RR	-2
$(\bar{\mathbf{3}}, \mathbf{1}, 1/3)$	S_1	$LL,\ RR$	-2

Table 1: The representations of the scalar LQ multiplets under the standard model gauge group, the accepted symbols in the literature, the chirality types of the quark and lepton that the LQ couples to, and the fermion number of the LQ.

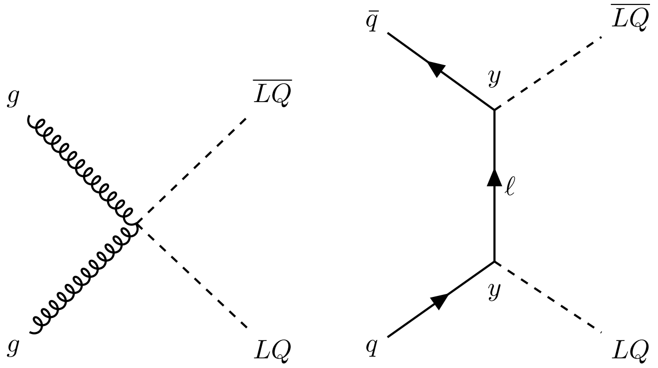

Figure 1: Diagrams contributing to conventional LQ pair production at the LHC at leading order.

right-handed neutrinos). In other words, the branching ratio $\beta(S_1 \to ue) = 1.0$ (note that due to charge conservation, the S_1 LQ cannot decay into an electron and a down-type quark).

1.1 Conventional LQ Production

We will briefly comment on the current conventional LQ production methods and give some brief phenomonolgical descriptions before we continue to describe the new asymmetric method.

LQ production at the LHC consists of a number of different mechanisms, largely categorized as either pair production or single/resonant production. There are two main contributions to normal pair production, as shown in Figure 1. This involves a QCD-driven component, as well as a Yukawa-driven component in which a lepton is exchanged in the t-channel. Note that in both cases, the final state LQ's are charge conjugates of each other; in other words, an LQ and its anti-particle are produced.

Based on the contributions to the single production cross section, we have that the amplitudes are proportional to only a single Yukawa coupling. So, the form of the total cross section is given by

$$\sigma_{\text{single}} = f(m_{\text{LQ}})|y_i|^2, \tag{3}$$

where the function f is dependent on the mass of the LQ. On the other hand, the pair production cross section takes a more complicated form:

$$\sigma_{\text{pair}} = f_{\text{QCD}}(m_{\text{LQ}}) + f_{\text{int}}(m_{\text{LQ}})|y_i|^2 + f_{\text{t-chan}}(m_{\text{LQ}})|y_i|^4. \tag{4}$$

In this case, we have the first term arising due to the QCD component, in which there are no fermions and thus no Yukawa coupling dependence. The third term is due to the t-channel diagram as shown before, and therefore that term depends quartic-ly on the Yukawa coupling, and the middle term is the interference between the two. Evidently, this cross section becomes dominated by the QCD contribution for small magnitudes of the Yukawa coupling, but for larger couplings, the quartic term in the pair production cross-section can begin to dominate. However, the mass dependence also plays a larger part in the pair production terms, meaning that the cross section drops off faster for higher LQ masses when compared to single production, hence, in the higher mass regime, pair production no longer dominates.

2 Asymmetric Pair Production

In a recent study [4], a novel method for pair producing LQ's that are not charge conjugates of each other, called "asymmetric" production, has been put forward, with the possibility that its cross sections are of similar or higher order than those of the conventional single and pair production methods mentioned in the previous section, for suitable masses and Yukawa coupling magnitudes. This novel method also comes with

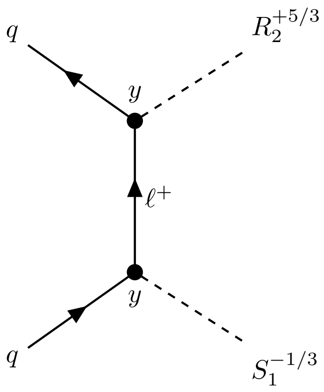

Figure 2: One possible contribution towards the asymmetric pair production of the S_1 and R_2 LQ's. Notably, their charges and fermion numbers are different.

a few added benefits, such as the possibility for quark-quark initial states as opposed to quark-antiquark initial states, which is particularly preferrable for the LHC due to lessened PDF suppression.

The main requirements to initiate this pair production is that the two leptoquarks in the final state couple to a lepton of the same chirality and flavor, and that their fermion numbers differ. As an example, the R_2 LQ can couple to the left-handed $SU(2)$ quark doublets and the right-handed $SU(2)$ lepton singlets (and vice-versa), and the S_1 LQ, as mentioned before, can couple to both the right and left-handed $SU(2)$ quark and lepton singlets. Because of the similar lepton coupling, we have the ability to asymmetrically produce an R_2 and an S_1 LQ in the final state from a quark-quark initial state, so long as the corresponding Yukawa coupling matrix elements are non-zero. The (only) feynman diagram for the leading order contribution for an S_1/R_2 asymmetric production is given in Figure 2.

2.1 The S_1 and R_2 Case

To examine further the S_1/R_2 scheme, we will focus only on the scenarios in which the LQ's couple to a right-handed lepton. We ignore left-handed lepton couplings in order to eliminate the possibility for neutrinos in the final state, as they will not be in the analysis channel we will describe in a later section. The relevent interaction terms in the Lagrangian are therefore:

$$\mathcal{L}_{\text{int}} = Y^{RR}_{1,ij}\bar{u}^c_i \ell_j S_1^\dagger + Y^{LR}_{2,ij}\left(\bar{Q}_i^\intercal \ell_j R_2\right) + \text{h.c.} \tag{5}$$

Now, if we expand out the second term into the charge eigenstates of the R_2 multiplet, we get

$$\mathcal{L}_{\text{int}} = Y^{RR}_{1,ij}\bar{u}^c_{R,i} e_{R,j} S_1^\dagger + Y^{LR}_{2,ij}\bar{u}_{L,j} e_{R,i} (R_2^{+5/3})^* + (Y_2 V_\dagger)^{LR}_{ij}\bar{d}_{L,j} e_{R,i}(R_2^{+2/3})^*, \tag{6}$$

where now we have added chirality subscripts for the fermion states, LQ charges as superscripts which are in terms of the charge of the positron, and asterisks to represent charge conjugations of specific LQ eigenstates. Further, since the previous Lagrangian was in terms of the weak eigenstates of the fermions,[1] when we induce spontaneous symmetry breaking, we need to introduce the Cabbibo-Kobayashi-Maskawa (CKM) matrix to relate the weak eigenstates of the down-type quarks to their mass eigenstates. Fortunately for us, the contents of this matrix is not important for our study, so, for simplicity, we will take it to be the identity. We will make the additional assumptions for simplicity that not only is there mass degeneracy within LQ multiplets, but also among different multiplets; i.e. in every process, the LQ's will be assumed to have the same mass.[2]

[1] Ordinarily, the weak eigenstates are denoted with primes, for instance, to indicate the difference between the weak and mass eigenstates. For the purposes of this study where this difference is not entirely important as well as for visual clarity, we chose to leave them out.

[2] As will be mentioned later, will end indeed up considering asymmetric masses by requesting a few extra grid points with final state LQ's of different masses, but only for future supplementary study.

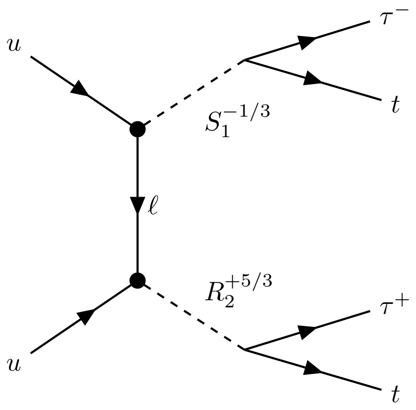

Figure 3: The contribution to $pp \to S_1 R_2 \to t\tau t\tau$ at leading order.

2.2 The $2\ell SS + 1\tau$ Channel

In the interest of analysis of this new pair production method, we need to pick a channel to study. Our group specializes in the $2\ell SS + 1\tau$ channel, which says that there are two same-sign leptons and one hadronically decaying tau in the final state. One way we can analyze this channel is to choose our Yukawa couplings such that the LQ's we pair produce both decay into heavy particles, like taus and top quarks. In order to achieve this, we would be looking at having

$$Y_{i3}^{XR} \neq 0, \text{ where } X = L, R \text{ and } i = 1, 2, 3. \tag{7}$$

This will allow the LQ to couple to only right-handed, third-generation leptons, which is only the tau. Since we also want a decay into a top quark, we need $Y_{33}^{XR} \neq 0$, but we will let the LQ couple to all quark generations to reduce PDF suppressions. Further study can be done to determine whether the reduced branching ratio $\beta(LQ \to t\tau)$ as a result of considering all quark generations out-weighs this reduced PDF suppresion benefit. Table 2 contains the cross sections for various masses and Yukawa couplings, where the magnitude of each Yukawa coupling matrix element is identical. In principle, it is entirely possible to specify that the Y_{33}^{XR} matrix element be greater than the others in order to isolate the decay $LQ \to t\tau$; this is something we also delegate to a future study.

m_{LQ}	y	σ
1500 GeV	0.1	1.23×10^{-7}
	0.5	1.232×10^{-7}
	1.0	7.676×10^{-5}
2000 GeV	0.1	0.001 201
	0.5	1.403×10^{-8}
	1.0	8.829×10^{-6}
2500 GeV	0.1	0.000 141 8
	0.5	1.023×10^{-6}
	1.0	1.769×10^{-5}

Table 2: The cross sections in picobarns for various mass and Yukawa coupling magnitudes. This corresponds to the Yukawa matrix elements chosen in Equation (2.2).

The only way to asymmetrically pair-produce the S_1 and R_2 LQ's and have them decay into top quarks and taus is shown in Figure 3. The S_1 LQ will subsequently decay into a tau and a top quark and the R_2 LQ will subsequently decay into an anti-tau and a top quark.

One example decay chain for the taus and top quarks are:

$$S_1 \to t\tau^- \to W^+ b\tau^- \to qqb\tau^- \tag{8}$$

$$R_2 \to t\tau^+ \to W^+ b\tau^+ \to \ell^+ \nu b\tau^+, \tag{9}$$

which gives us a satisfactory final state for our analysis channel.

3 Methodology

As this is a novel production method, there are a number of steps we must take in order to begin analyzing this process. We will make use of the ATLAS experiment's Athena software, which contains a number of helpful tools and wrappers around popular event generators, simulators, and reconstruction algorithms. There are three main steps for the production of the signal in the ATLAS experiment: testing and production of validation plots, approval from subgroup conveners, and submission of a ticket to generate events on the grid.

3.1 Testing and Production of Validation Plots

To start to generate events for our chosen process, we will use MadGraph5_aMC@NLO, a popular program that is used for event generation and cross-section calculation. It comes equipped with many tools for doing these calculations within the regime of the standard model, but for anything outside of this, it requires a supplementary "model" that describes all of the details of the BSM process we want to simulate. The typical pipeline for this is to use the FeynRules package within Mathematica, from which a Universal FeynRules Output (UFO) model is produced that can be imported into MadGraph5_aMC@NLO. Fortunately for us, models have already been made for the scalar LQ's; see Refs [2,3]. A scenario with multiple LQ's present requires a simple combination of the individual models, and it is one such combination model that we will use.

The Athena software in the ATLAS experiment provides the `Gen_tf.py` "transform" script, which wraps around MadGraph5_aMC@NLO and Pythia8 (among other generators) and takes in a "JobOptions" file, a pseudo-Python script. The JobOptions file contains all the information relevant for event generation and parton showering. For instance, the desired process is specified using MadGraph5_aMC@NLO syntax, Yukawa couplings, masses, and the number of events are specified, and filters can be added which apply generator-level kinemetic cuts. In the command line, the name of the output EVNT file is specified, which contains the raw event data after the showering is complete.

The EVNT file cannot be directly parsed without considerable effort, so we need to produce so-called "TRUTH derivations". This is done with another one of Athena's transform scripts called `Derivation_tf.py`, and requires only a few command-line inputs such as the input file and the name of the xAOD-formatted output file. There are a few different types of TRUTH formats which contain varying levels of the full truth record, from a direct copy of the EVNT data in xAOD format to a lite version that contains only the absolute necessary information for an analysis.

Once the TRUTH derivations have been made, we can then apply a simple analysis using Athena's EventLoop framework, which is a tool that let's one easily parse through xAOD files and extract relevant information (among many others). Inside this analysis program, we can grab all of the kinematics for all of the final state particles, including the LQ's and their intermediate decay products. All that remains is to use ROOT to place them into histograms, which is a trivial exercise. Four example validation plots are shown in 4.

3.2 Approval and Submission to the Grid

To make statistically valid predictions, we need to produce a very large number of events, more than could be reasonably produced even on a decently powerful single machine. Additionally, if the events are not produced within ATLAS (i.e. on a personal computer), their validity may be called into question. So, our events will need to be produced on ATLAS's central grid, which splits up the jobs and lets them all run

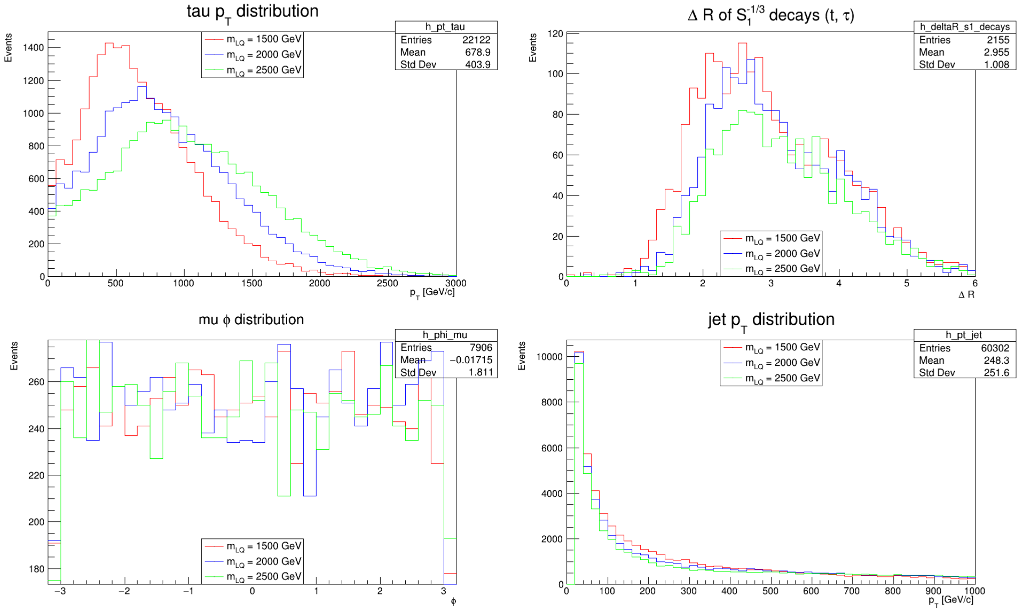

Figure 4: Validation plots generated from test runs for the LQ model. Top left shows the p_T distribution for the τ which came from the decay of the LQ's; top right shows the ΔR for the decay products (t and τ) of the S_1; bottom left shows the ϕ distribution for the final state μ's; and the bottom right shows the p_T distribution for the produced jets.

in parallel on high-performance machines to improve efficiency. Before this can be done, approval must be granted by the relevent convening groups. In our case, this is the Lepton+X group, of which Tau+X is a subgroup. For documentation purposes, this was done here: `https://indico.cern.ch/event/1439002/`. After approval is granted, we must submit a request with details about this approval, the relevant JobOptions files for each mass (or some other parameter) point, and a few additional details. Then, the ATLAS MC coordinators will run the generation, simulation, and reconstruction. This request was done here: `https://its.cern.ch/jira/browse/ATLMCPROD-11359`

4 Analysis

There are a number of SM processes that contribute also to the $2\ell SS{+}1\tau$ channel, including Higgs production, heavy boson ($W^{\pm}$, Z) production, top quark production, and several others; the full table of background (and signal) processes that are considered for the analysis is given in Table 4 in Appendix A. Because of this, we need to find a way to separate these ordinary SM "backgrounds" from our LQ pair production "signal", and the best way to achieve this is with machine learning. This is because there are dozens of relevent features that are present in the ntuples created from the event generation described in Section 3, and many machine learning models have an edge over conventional fitting algorithms when it comes to many-dimensional scenarios such as this.

This process is done in a number of different ways based on the experiment – we will briefly list our steps here then fully describe them in the subsequent sections. First, we make some preliminary plots/tables using TRExFitter to show the raw number of events with and without weighting/selection criteria. Then, we produce so-called "small" ntuples which use the same selection criteria to slim down the raw ntuples and make further analysis quicker. At the same time, we take the full dataset and transform it to numpy arrays so that it can be used to train a machine learning model. After the model is trained, the small ntuples are fed through it to generate probabilities in what are called "friend" ntuples. Lastly, TRExFitter can generate full distributions and other statistical measures using the friend-ntuples.

4.1 $t\bar{t}H$ Analysis in the $2\ell SS + 1\tau$ Channel

Previous analysis for the $2\ell SS + 1\tau$ channel had been done using release (Rel.) 21 ntuples, but the current release is now Rel.22. There are numerous differences between the different releases, such as different variables, different algorithms to produce those variables, general data quality, and more. Most notable are the different variables, as things like weights and selection criteria often need to be significantly changed due to variables that are renamed or no longer present. We briefly summarize some more details with the branch changes in Section 4.1.1.

Our main task here is to port the Rel.21 code to be compatible with the Rel.22 data. From there, while waiting for the LQ samples to be produced, we can rerun the previous $t\bar{t}H$ analysis to make some comparisons and check the quality of both the new data and the modified code. It is also a good start for further $t\bar{t}H$ analysis in this release.

4.1.1 Branch Changes from Rel.21 to Rel.22

Not only was the previous analysis using Rel.21 ntuples, but there were a few variables present in the ntuples that had been custom-made specifically for Higgs multi-lepton analysis. There were a good number of these variables that had to be removed from the selection criteria, as well as from the feature list that was fed into the machine learning model. A full list of all the branches that were present and used in the Rel.21 analysis but not present in the Rel.22 analysis is given in Section C. Note that this list contains all variables that were not themselves present, but it is possible some of them were able to be renamed or easily replaced. For instance, we were able to make the following replacement:

$$lep\_isolationLoose\_VarRad\_X \rightarrow lep\_Iso\_Loose\_VarRad\_X$$

where $X = 0, 1$, among a few others.

Further, there were a number of prompt lepton veto (PLV) variables that were completely removed. Prompt versus non-prompt leptons are those that emerge from the vertex of interest as opposed to one that emerges from some secondary vertex as a result of the decay of some heavy intermediary particle like a $W^{\pm}$ or Z boson that is irrelevent for the study. It is very useful to be able to remove the non-prompt leptons. These PLV variables are in the process of development from the IFF group, which is attempting to unify the efforts towards prompt and fake lepton classification across many of the analysis groups. As a result, there is not much availability for these classifications as of now.

4.1.2 Pre-fit Yields and Small Ntuples

Before starting any major data processing, it is helpful to know the raw number of events we have, as well as how many we have in the signal region, the $2\ell SS + 1\tau$ region. To do this, we need to modify the selection TRExFitter uses to incorporate all the criteria we want. The full selection string we used is:

```
Selection: XXX_2LSS_SELECTION && XXX_LEPTON_PROMPT_SELECTION &&
↪  XXX_EXCLUSION_Z_PEAK && XXX_TRIGGER_SELECTION && sumPsbtag85 > 5
```

The variables starting with “XXX” are found in a `replacement.txt` file as they are often quite long, but their contents are largely self-explanatory given their names. The variable `XXX_2LSS_SELECTION` ensures that there are two leptons of the same sign in the final state, as well as ensuring that there is a hadronically decaying tau. `XXX_EXCLUSION_Z_PEAK` provides some additional kinematic constraints by ensuring that the invariant mass of the two leading leptons is not close to the Z resonance peak. `XXX_LEPTON_PROMPT_SELECTION` used to contain a number of PLV variables to remove non-prompt leptons as well as some other generic identification variables. While those had to be removed, there remained still some useful variables in this bit of the selection, such as *lep_ambiguityType_X*, which assists in a different but still slightly related method of lepton identification. Lastly, the variable `sumPsbtag85` is a measure of btagging. Many $t\bar{t}$ events have relatively low numbers of b-jets, so making this cut helps remove that background.

On top of this pre-selection, we also apply a set of weights. Often times, in a simulation, the algorithms and whatnot are not perfect, leading to certain processes being underrepresented compared to how they would be in a real experiment. As a result, we apply these weights to try and match as closely to experiment as possible. This criteria involves things like run years, as well as certain sample numbers. The full criteria for the weight and the pre-selection is given in Appendix D.

	Non-weighted		Weighted	
Sample	Before Selection	After Selection	Before Selection	After Selection
$t\bar{t}H$	1015311	23037	785.2	17.8
$t\bar{t}W$	902540	10759	2524.5	30.3
$t\bar{t}Z(Z/\gamma*)$	2164697	14500	2116.6	14.0
$t\bar{t}$	350927	457	47738.8	61.8
VV	17373668	566	74970.7	1.3
Others	4148134	1329	508810.8	5.6
Totals	25955277	50648	636946.6	130.7

Table 3: The yields for all the sample regions before and after selection, with and without weighting.

Now that the selection and weighting criteria have been established, we can apply it and run TRExFitter to determine yields for all of our sample regions. Table 3 shows these, listing the yields for the samples before and after the pre-selection and weighting. We can also produce some simple distributions for the regions we want to examine, such as the leading lepton p_T or the number of jets. Those plots are given in Figure 5. The full set of the validation plots are given in Section 9.

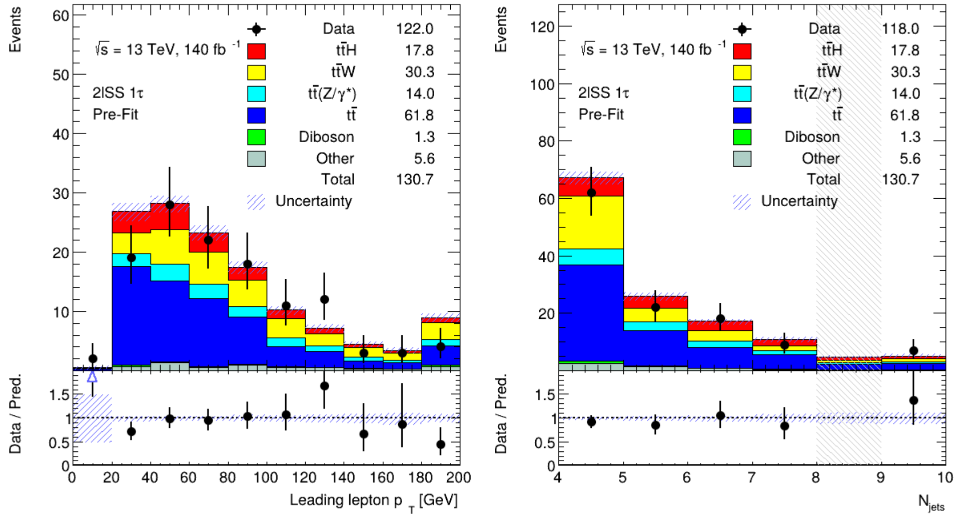

Figure 5: Distributions for the leading lepton p_T and the number of jets.

Lastly, To create the small ntuples, we skim through the raw ntuples and filter out events that don't pass our selection, essentially getting rid of all the uninteresting events. This is done with a simple script that is parallelized using HTCondor. Otherwise, if it is ran just on an ordinary `lxplus` node, it will not only take forever but also probably be killed.

4.1.3 Model Training and Friend Ntuples

We will use PyTorch to train our machine learning model, and since PyTorch cannot read ROOT files out of the box, we must transform all of our small ntuples into numpy arrays, which can then be read into PyTorch. This is also takes a long time, and is done by submitting an HTCondor job. The actual contents of the script itself are not important, only that it translates the data into PyTorch-readable format. Once this is complete, we can choose our model to train on the data. For simplicity and speed due to time constraints, we chose one of the simplest models, the ResNet-6. For information on how it works, see Ref [6], for instance.

The model was trained on a large number of features. Just as with the previous step, in which there were a number of missing branches that had to be taken into account in forming the selection criteria, there were a number of training features that were missing in this stage. The full list of these features is given in List C.2. The model itself trained relatively quickly on a local machine with 16 GB of ram and an NVIDIA RTX 4050. Figure 6 contains the R.O.C. curves for classification of events as $t\bar{t}H$ (signal) and $t\bar{t}W$, the most important background.

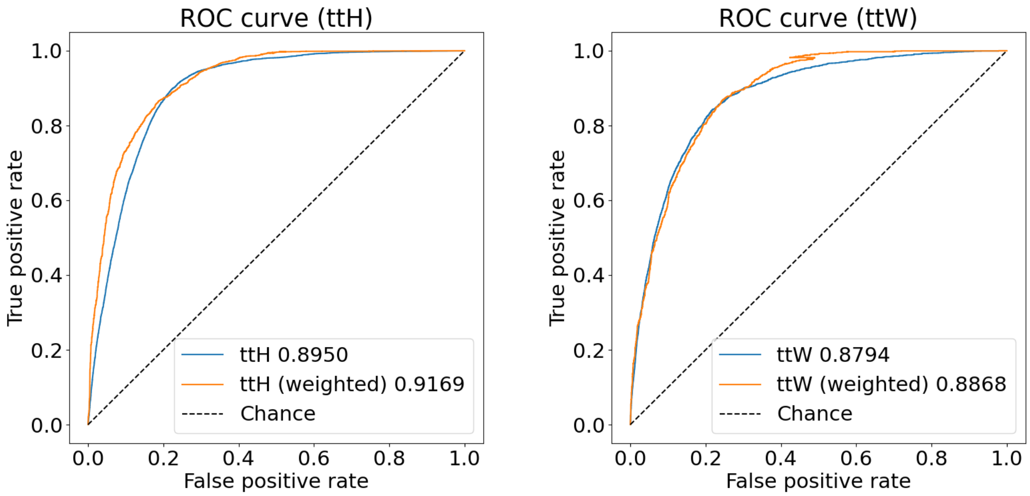

Figure 6: R.O.C. curves for classification of events as either $t\bar{t}H$ signal, or $t\bar{t}W$, the most important background.

The next step is to produce the friend ntuples, which contain a single branch that gives the probabilities for events to be classified as $t\bar{t}H$. This is the alternative to copying the ntuples and simply adding the branch to the existing files, which saves space. All that needs to be done to produce the ntuples is to pass the small ntuples into the network and save the output; these are the friend ntuples.

4.2 Probability Distributions and Statistical Uncertainties

With the friend ntuples created, we can now run TRExFitter again to produce probability distributions and other plots related to the statistical uncertainties. This is done by passing a few extra flags into the run command for TRExFitter, as well as adding another region to the configuration with the option `UseFriend: True`. After doing this, we get a plot like that shown in Figure 7.

There are a number of things to note. First of all, the $t\bar{t}$ background that was unusually prominent in the previous validations plots appears to separate very easily as it completely disappears after the first two bins, which is a very good sign. Additionally, and perhaps most importantly, it is quite clear that the $t\bar{t}H$ has been very nicely separated. To more closely analyze how well our model performed, we can make a simple measure of the statistical uncertainly associated with the significance of the model classification. This is shown in Figure 8. Also in the figure is the output from the Rel.21 analysis [9], and we can immediately see that the new ntuples, even with the reduced criteria and simpler model, preform very well, even better than the previous analysis.

4.3 LQ Analysis

Unfortunately, the requested LQ samples were not completed in time for an analysis to be run on them. The formalities associated with gaining approval from the various convening groups as well as the requesting process took up far more time than was initially expected. Further, I had to learn all about the theory behind LQ's and how to operate the Athena software. Despite this, it was still very much worthwhile to spend the time porting the Rel.21 $2\ell SS + 1\tau$ analysis code into Rel.22 and testing it on the existing $t\bar{t}H$

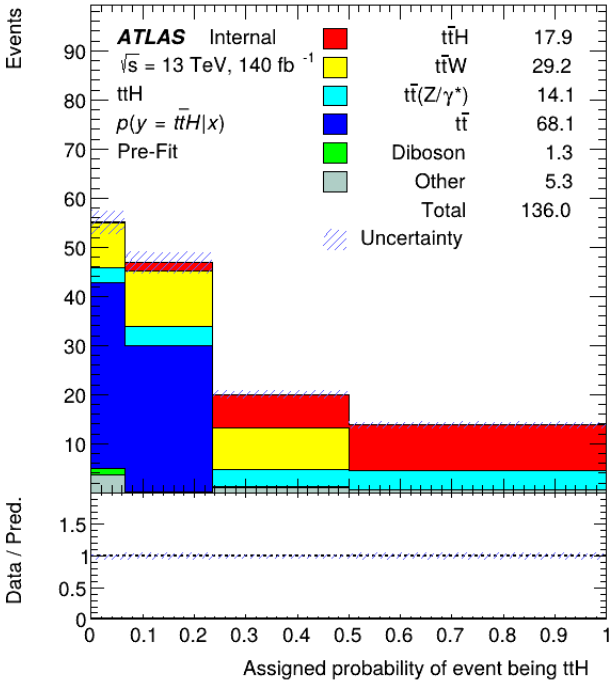

Figure 7: Distribution for the probability that a given event will be classified as the $t\bar{t}H$ signal.

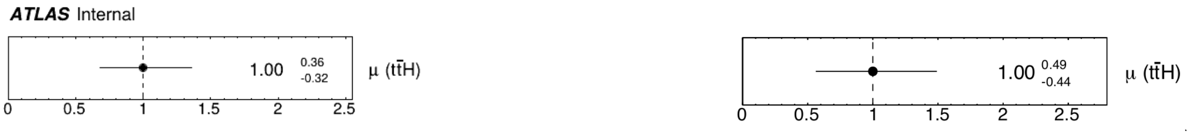

Figure 8: The statistical uncertainties (for only the statistics, not the systematics) for the significance of our model performance. On the left is our model, and on the right is the performace of the previous Rel.21 analysis model, which was a relatively complex transformer; see Ref. [9]

samples. We found that even with a simple neural network, we were able to get some relatively good results, which bode well both for the quality of Rel.22 samples but also for the future analysis of the LQ samples.

5 Discussion

For this summer student project, we investigated a novel method of leptoquark pair production which involves two leptoquarks in the final state that are not charge conjugates of each other. Additionally, there is the added benefit of a quark-quark initial state, which is desirable at the LHC. We outlined the steps taken to test the model and submit an official MC request on the grid for larger samples. A $t\bar{t}H$ analysis from Rel.21 ntuples was ported to Rel.22, which involved changing/removing a large number of branches/features from the pre-selection and training processes. This analysis was then done for the existing $t\bar{t}H$ Rel.22 samples, and relatively good agreement was found even for a simple neural network.

The next steps for the project would be to generate the full set of LQ samples for mass points in the range 1500-2500 GeV and for Yukawa couplings of 0.5 and 1.0. Then, the current $t\bar{t}H$ signal samples would be moved to background, and the new LQ samples become the new signal. The whole pipeline should work without any additional changes, so further hyperparameter tuning in terms of the machine learning model and selection criteria can be done. For instance, as mentioned before, PLV variables, which are helpful in removing fakes, can be added once they (or a viable substitute) are developed for Rel.22. This will assist in reducing backgrouns like $t\bar{t}$ which are seen to be quite relevent in the final plots.

Acknowledgements

I would like to extend my utmost graditude to the CERN Summer Student Program for putting on such a wonderfully organized summer. Additionally, I would like to thank Dr. Junjie Zhu, Dr. Myron Campbell, and Steven Goldfarb for organizing the US side of things, as well as providing advice and assistance through the program. I would like to thank my supervisor Dr. André Sopczak for providing such an interesting project and assisting me throughout. Additionally, I would like to thank all the members of the Tau+X and LPX groups for providing valuable feedback and words of encouragement in my various presentations.

Appendix

A List of DSIDs Considered for Analysis

Process	DSIDs
LQ	545824, 545825, 545826
$t\bar{t}H$	346343, 346344, 346345
$t\bar{t}W$	700168
$t\bar{t}W\_EW$	700205
$t\bar{t}Z$	504330, 504334, 504342
$t\bar{t}$	410470
VV	364250, 364253, 364254, 364255, 364283, 364284, 364285, 364286, 364287, 363355, 363356, 363357, 363358, 363360, 363489
Others	410560, 410408, 410646, 410470, 304014, 345705, 345706, 34572, 364242, 364243, 364244, 364245, 364246, 364247, 364248, 364249, 410081, 364156, 364157, 364158, 364159, 364160, 364161, 364162, 364163, 364164, 364165, 364166, 364167, 364168, 364169, 364170, 364171, 364172, 364173, 364174, 364175, 364176, 364177, 364178, 364179, 364180, 364181, 364182, 364183, 364184, 364185, 364186, 364187, 364188, 364189, 364190, 364191, 364192, 364193, 364194, 364195, 364196, 364197

Table 4: Table of all the DSID for the background and signal processes that are considered for the analysis. The “Others” category contains tZ, WtZ, tW, and ttt samples.

B Additional Region Plots

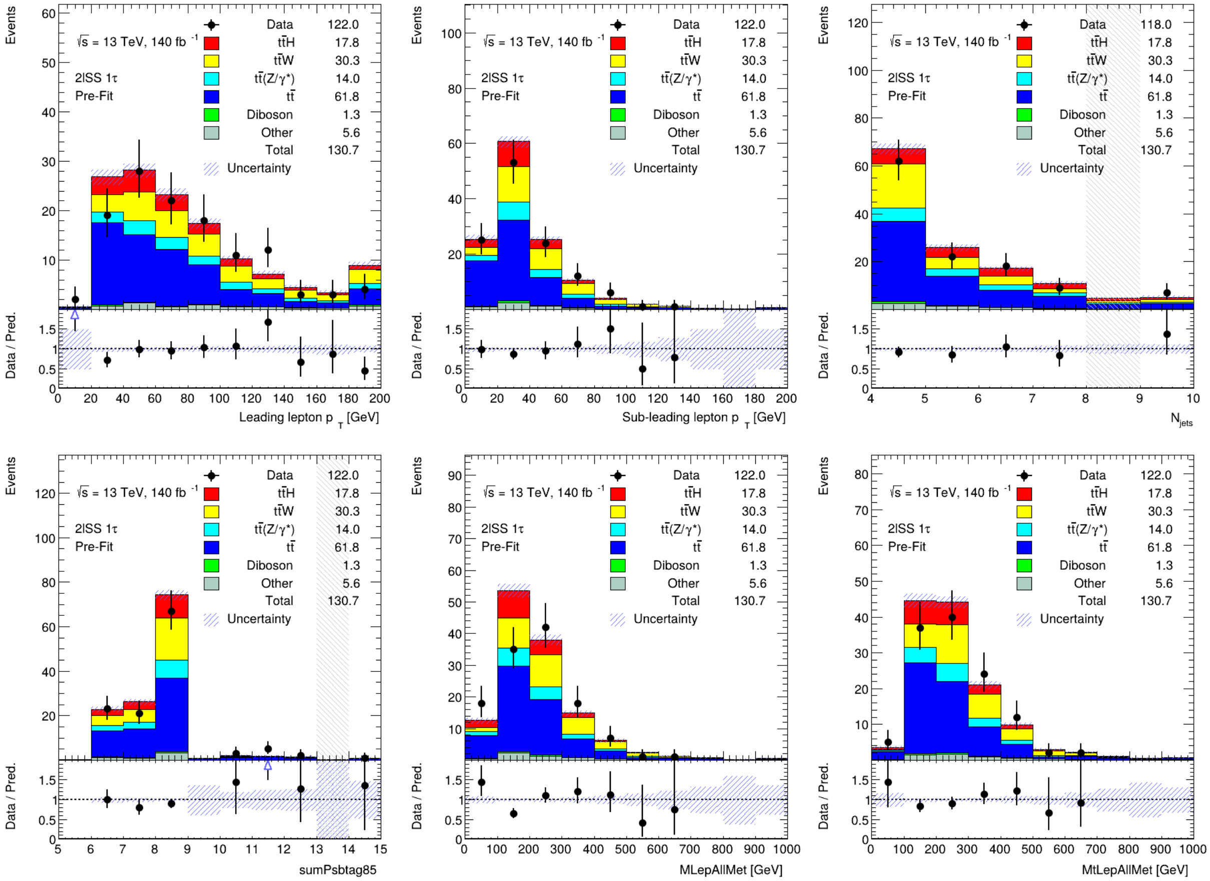

Figure 9: Additional regional plots.

C Rel.21 Branches Missing/Renamed from Rel.22

C.1 Pre-Selection Criteria

- *dilep_type*
- *nJets_OR_DL1r_85*
- *custTrigMatch_LooseID_FCLooseIso_SLTorDLT*
- *lep_Mtrktrk_atConvV_CO_X*
- *lep_Mtrktrk_atPV_CO_X*
- *passPLIVTight_X*
- *lep_isolationLoose_VarRad_X*
- *lep_chargeIDBDTResult_recalc_rel207_tight_X*

C.2 Training Features

- *HT_fwdJets*
- *HT_inclFwdJets*
- *Mb0*
- *Mb1*
- *Mlb*
- *MtLep1Met*
- *Ptll01*
- *bTagSF_weight_DL1r_77*
- *bTagSF_weight_DL1r_85*
- *best_Z_Mll*
- *best_Z_other_Mll*
- *best_Z_other_MtLepMet*
- *dEta_maxMjj_frwdjet*
- *eta_frwdjet*
- *lep_Mtrktrk_atConvV_CO_0*
- *lep_Mtrktrk_atConvV_CO_1*
- *lep_Mtrktrk_atPV_CO_0*
- *lep_Mtrktrk_atPV_CO_1*
- *lep_RadiusCO_0*
- *lep_RadiusCO_1*
- *lep_chargeIDBDTResult_recalc_rel207_tight_0*
- *lep_nInnerPix_0*
- *lep_nInnerPix_1*
- *lep_nTrackParticles_0*
- *lep_nTrackParticles_1*
- *lep_sigd0PV_0*
- *lep_sigd0PV_1*
- *minDeltaR_LJ_0*
- *minDeltaR_LJ_1*
- *minDeltaR_LJ_2*
- *minOSMll*
- *minOSSFMll*
- *mjjMax_frwdJet*
- *nFwdJets_OR*
- *nJets_OR_DL1r_77*
- *nJets_OR_DL1r_85*
- *passPLIVTight_0*
- *passPLIVTight_1*
- *taus_DL1r_0*
- *taus_passJVT_0*

D Full Pre-Selection and Weight Criteria Strings

```
Selection: l2SS_1tau && ((lep_Pt_0>=10e3 &&
↪  lep_Pt_1>=10e3)&&(fabs(lep_Eta_0)<=2.5&&fabs(lep_Eta_1)<=2.5) &&
↪  ((abs(lep_ID_0) == 13 && lep_isMedium_0 && lep_Iso_Loose_VarRad_0) ||
↪  (abs(lep_ID_0) == 11 && lep_isTightLH_0 && lep_Iso_Loose_VarRad_0 &&
↪  lep_ambiguityType_0 == 0)) && ((abs(lep_ID_1) == 13 && lep_isMedium_1 &&
↪  lep_Iso_Loose_VarRad_1) || (abs(lep_ID_1) == 11 && lep_isTightLH_1 &&
↪  lep_Iso_Loose_VarRad_1 && lep_ambiguityType_1 == 0))) && ((abs(lep_ID_0)==11
↪  && lep_isTightLH_0) || (abs(lep_ID_0)==13 && lep_isMedium_0)) &&
↪  ((abs(lep_ID_1)==11 && lep_isTightLH_1) || (abs(lep_ID_1)==13 &&
↪  lep_isMedium_1)) && (abs(Mll01-91.2e3)>10e3) &&
↪  custTrigMatch_CombinedWPs_SLTorDLT && (sumPsbtag85 > 5)
```

```
XXX_MC_WEIGHT: ((36207.66*(RunYear==2015 || RunYear==2016) +
↪  44307.4*(RunYear==2017) + 58450.1*(RunYear==2018))*(1/138965.16) *
↪  weight_pileup * weight_mc * xs/totalEventsWeighted)
```

References

[1] Andreas Crivellin and Luc Schnell. Complete lagrangian and set of feynman rules for scalar leptoquarks. *Computer Physics Communications*, 271:108188, February 2022.

[2] Ilja Doršner, Svjetlana Fajfer, and Ajla Lejlić. Novel leptoquark pair production at lhc. *Journal of High Energy Physics*, 2021(5), May 2021.

[3] Ilja Doršner and Admir Greljo. Leptoquark toolbox for precision collider studies. *Journal of High Energy Physics*, 2018(5), May 2018.

[4] Ilja Doršner, Ajla Lejlić, and Shaikh Saad. Asymmetric leptoquark pair production at lhc. *Journal of High Energy Physics*, 2023(3), March 2023.

[5] H. Georgi and S. L. Glashow. Unity of All Elementary Particle Forces. *Phys. Rev. Lett.*, 32:438–441, 1974.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[7] Gudrun Hiller and Martin Schmaltz. R_K and future $b \to s\ell\ell$ physics beyond the standard model opportunities. *Phys. Rev. D*, 90:054014, 2014.

[8] Uma Mahanta. Neutrino masses and mixing angles from leptoquark interactions. *Physical Review D*, 62(7), September 2000.

[9] Vladyslav Yazykov. Optimization of the tth selection including systematics using machine learning with atlas data, 2023. Presented 05 Sep 2023.