

Semester Thesis

Top jet training region for Normalized Autoencoders in the t-channel Semivisible Jet Analysis in CMS

Autumn Term 2024

Supervised by:

Florian Eble
Dr. Roberto Seidita
Prof. Dr. Annapaola de Cosa

Author:

Roshni Sahoo

Declaration of Originality

I hereby declare that the written work I have submitted entitled '**Top jet training region for Normalized Autoencoders in the t-channel Semivisible Jet Analysis in CMS**' is original work which I alone have authored and which is written in my words.

Author(s)

Roshni Sahoo

Supervisor(s)

Florian Eble
Roberto Seidita

Supervising lecturer

Annapaola de Cosa

With the signature, I declare that I have been informed regarding normal academic citation rules and that I have read and understood the information on 'Citation etiquette' (<https://www.ethz.ch/content/dam/ethz/main/education/rechtliches-abschluesse/leistungskontrollen/plagiarism-citationetiquette.pdf>). The citation conventions usual to the discipline in question here have been respected.

The above-written work may be tested electronically for plagiarism.

Zürich, 19.02.2024

Place and date



Signature

Abstract

The existence of Dark Matter (DM) provides strong evidence for physics beyond the Standard Model. The DM resides within a ‘dark sector’, which, if it is a strongly coupled sector, can result in jet-like collider objects in the Large Hadron Collider called ‘semivisible’ jets. Semivisible jets are collimated sprays of invisible dark hadrons and Standard Model hadrons. Normalized Autoencoders (NAE) can be used as anomaly detection algorithms to perform signal-agnostic searches for semivisible jets. This can be done by training the NAE to recognize ordinary Standard Model (SM) jets to tag potential semivisible jets as ‘anomalous’ objects. We designed a signal-free training region composed of top-quark jets arising from $t\bar{t}$ events and demonstrated that this is a suitable region to train the NAE directly on real data in order to classify semivisible jets from Standard Model top-quark jets.

Contents

1	Introduction	1
2	Research Background	2
2.1	Large Hadron Collider	2
2.2	Compact Muon Solenoid	2
2.3	Semivisible jets	3
2.4	Semivisible jet search strategy	6
2.4.1	Autoencoders	6
2.4.2	Normalized Autoencoders	7
2.4.3	Unsupervised semivisible jet tagging	8
2.4.4	Background jets	8
3	Analysis	10
3.1	$t\bar{t}$ events categorization	10
3.2	Top quark jet categorization	12
3.3	Data Samples	14
3.3.1	Generation	14
3.3.2	Pre-selection Region	15
3.3.3	Top jet Training Region	15
3.4	Training normalized autoencoder on $t\bar{t}$ jets	19
3.4.1	Training on leading $t\bar{t}+$ jets	19
3.4.2	Training on all $t\bar{t}$ jets in bins of p_T	20
4	Conclusion	23

1 Introduction

The Standard Model (SM) of particle physics has proven to be highly successful in explaining many experimental results and predicting various phenomena with great precision. However, there are still many open problems, including the origin of Dark Matter (DM), neutrino masses, etc., that the Standard Model cannot explain. The existence of Dark Matter provides strong evidence of physics beyond the Standard Model. The discovery of Dark Matter is one of the core missions of Large Hadron Collider (LHC) experiments, with DM searches primarily focused on Weakly Interacting Massive Particles (WIMPs). WIMPs leave the detector undetected and produce missing transverse energy \cancel{E}_T in the detector that can be inferred by transverse momentum conservation in every collision.

Hidden Valley models [1] provide an alternative approach to search for DM candidates. These models give rise to strongly interacting dark sectors where such a QCD-like strong interaction can cause fragmentation and hadronization of dark particles. Such a dark parton shower contains invisible DM particles and dark sector states that can decay into SM particles via a portal. One such example is a semivisible jet (SVJ) [2, 3], which are collimated sprays of SM hadrons interspersed with invisible stable dark hadrons. These SVJs result in a multijet+ \cancel{E}_T signature where one of the jets is closely aligned with the \cancel{E}_T .

The details of the dark parton shower depend on many unknown parameters, such as dark hadron masses, the dark confinement scale and the dark coupling constant, which could then be a function of other parameters in a fully specified Hidden Valley model. Thus, to avoid model-dependent searches of semivisible jets, anomaly detection techniques such as unsupervised machine learning can be used to detect SVJs. Such techniques can detect SVJs among a wide variety of background SM jets without prior knowledge about the signal (SVJ).

Autoencoders (AE) can be used to perform signal-agnostic searches for new physics [4, 5] where the AEs have been trained to recognize SM jets and can be used to tag anomalous objects such as SVJs. AEs have been proven to perform well for detecting SVJs in the presence of QCD background jets [6]. However, they perform poorly when discriminating SVJs from top-quark jets as background. Normalized Autoencoders (NAE) [7] can overcome the shortcomings of AEs and perform much better for signal-agnostic searches of semivisible jets in the presence of top-quark background jets [8]. First applications of the NAE for anomalous jet detection used jet images [9]; however, jet substructure variables are used instead in this study.

In this project, we aim to design a signal-free top training region so that the NAE can be trained to learn top-quark jets, which can then be used for discriminating

semivisible jets from background top-quark jets arising from $t\bar{t}$ events in the signal region, independent of the signal model.

2 Research Background

2.1 Large Hadron Collider

The Large Hadron Collider (LHC) is the world's largest and most powerful particle accelerator built by the European Council for Nuclear Research (CERN) from 1998 to 2008. The LHC is a circular collider with a 27-kilometer circumference. The collider tunnel contains two adjacent high-energy particle beamlines which travel in opposite directions around the ring. These beams are made to intersect at four points along the ring where particle collisions take place corresponding to the four particle detectors - ATLAS [10], ALICE [11], CMS [12] and LHCb as depicted in yellow in figure 1. These beams are guided around the accelerator ring using superconducting magnets. The LHC operates beams of very high energy, up to 6.8 TeV per beam, which results in a total collision energy of 13.6 TeV.

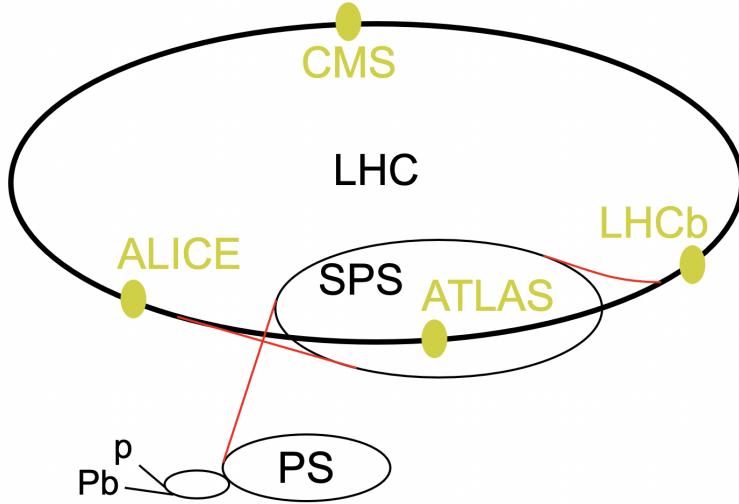


Figure 1: Layout of the Large Hadron Collider.

2.2 Compact Muon Solenoid

The Compact Muon Solenoid (CMS) is one of the four detectors in the LHC accelerator ring; it is a general-purpose detector. The CMS is a cylindrical detector

surrounding the collision point where high-energy protons collide. It consists of several layers, as shown in figure 2, the innermost layer being the silicon tracker, which detects the trajectories of charged particles that enter the detector. The next layer is an electromagnetic calorimeter (ECAL), which detects and measures the energy of photons and electrons, followed by a hadron calorimeter (HCAL), which detects and measures the energy of hadrons like mesons and baryons. Following the ECAL and HCAL is a superconducting magnet that bends the paths of charged particles so that their momenta can be measured from the radius of curvature. Finally, in the outermost layer is the muon detector. Muons are weakly interacting particles and pass undetected at the inner layers of the detector. The CMS experiment at CERN plays a vital role in potentially discovering physics beyond the Standard Model, such as SuperSymmetry and Dark Matter, and in further studying the properties of the Higgs boson, which was also discovered in the CMS and ATLAS experiments.

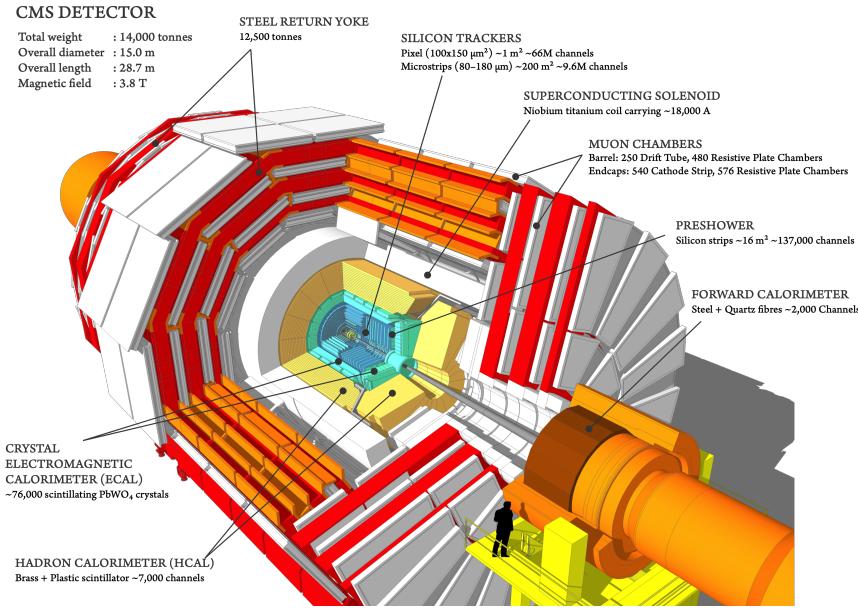


Figure 2: Layers of the Compact Muon Solenoid (CMS) [13].

2.3 Semivisible jets

Hidden Valley models hint at the possibility of a new type of ‘hidden’ particle that interacts weakly with ordinary particles and could potentially be detected at the LHC. Hidden valley models are those in which the Standard Model (SM) gauge group G_{SM} is extended by a non-Abelian gauge group G_ν , giving rise to new light particles. These new light particles don’t have charges in the SM but

can couple to SM particles via a mediator. Under the assumption that this new gauge group is $SU(N)$, they give rise to gauge bosons and confinement, just like in the case of Quantum Chromodynamics, the theory of strong interactions in the Standard Model. This confinement leads to hadronization into objects like hadrons and mesons under some confinement scale Λ_ν . Due to this confinement, collimated sprays of dark hadrons will be produced in the LHC which cannot be detected unless they can decay into Standard Model particles. These dark hadrons can decay into Standard Model quarks, which can then hadronize to form jets of SM hadrons. These hadronic jets interspersed with dark matter are termed semivisible jets (SVJ).

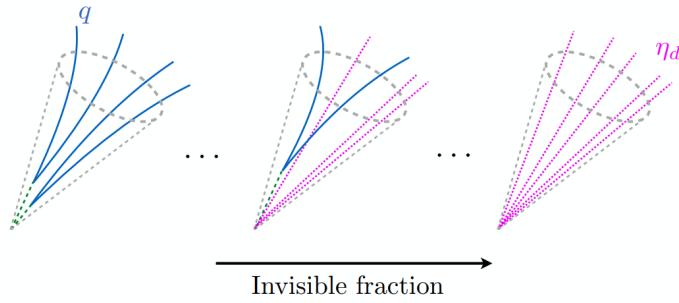


Figure 3: The leftmost image represents the case in which all the dark hadrons decay into SM quarks (green dashed lines), giving a signature that has been searched for in new physics dijet resonance searches. As we move towards the right, some fraction of dark hadrons is stable, resulting in SM jets interspersed with invisible dark hadrons, giving rise to so-called semivisible jets. All the hard hadrons are stable in the right figure, giving rise to a fully invisible jet. Image obtained from [3].

The central diagram in figure 3 shows that some fraction (r_{inv}) of dark hadrons can decay into Standard Model quarks, giving rise to semivisible jets. These stable and invisible dark hadrons are Dark Matter candidates. Semivisible jets have a multijet+ $\cancel{E_T}$ signature where one of the jets is closely aligned with $\cancel{E_T}$. The direction of the missing transverse energy $\cancel{E_T}$ relevant to this project is illustrated in the central diagram of figure 4.

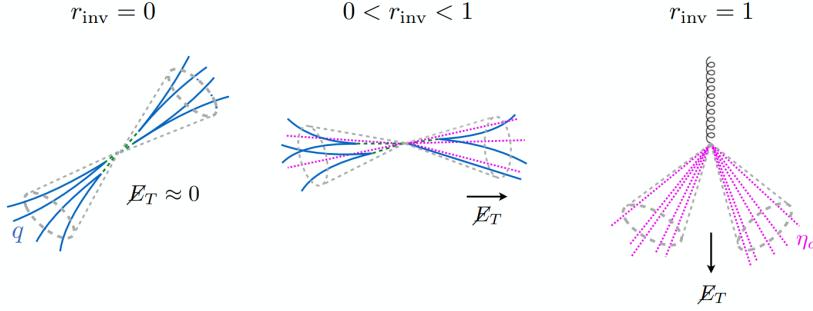


Figure 4: Missing transverse energy for different dark hadron fractions. The signatures on the left and right are addressed by dijet and WIMP searches, respectively. Our focus is on the signature in the middle.

There are two different channels via which the dark sector can interact with the Standard Model sector, namely the s-channel and the t-channel. In this project, we focus on the t-channel production of dark hadrons as illustrated in the right Feynman diagram of figure 5.

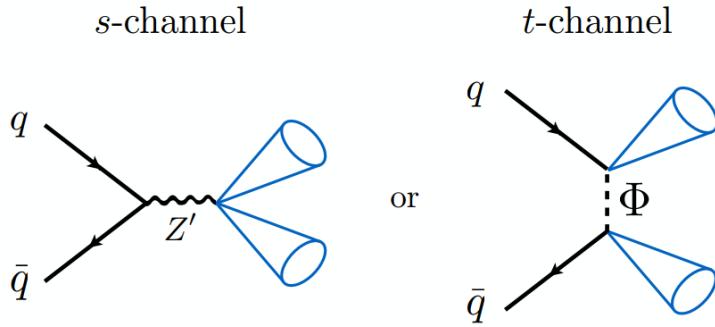


Figure 5: Different channels via which the Dark sector can interact with the SM sector. Image obtained from [3].

The important parameters of the model that influence the kinematic properties of the final state are,

- α_d - the dark strong coupling constant
- m_Φ - the mass of the t-channel mediator
- m_{dark} - the mass of the dark hadrons

- y_{dark} - Yukawa coupling between SM and dark quarks
- r_{inv} - the fraction of stable, invisible dark hadrons

2.4 Semivisible jet search strategy

2.4.1 Autoencoders

Autoencoders consist of two main components: an encoder f that transforms the input space into a lower dimensional latent space and a decoder g that maps the latent space back to output space with the same dimensions as the input space. This whole process is illustrated in figure 6. An autoencoder (AE) is designed to minimize the reconstruction error between the input and output such that if it encounters samples outside the training distribution, also known as ‘anomalies’, they will have higher reconstruction errors. Autoencoders serve the purpose of conducting model-agnostic searches of physics beyond the Standard Model. In this case, the AEs are training on SM jets, and the SVJs are the anomalies.

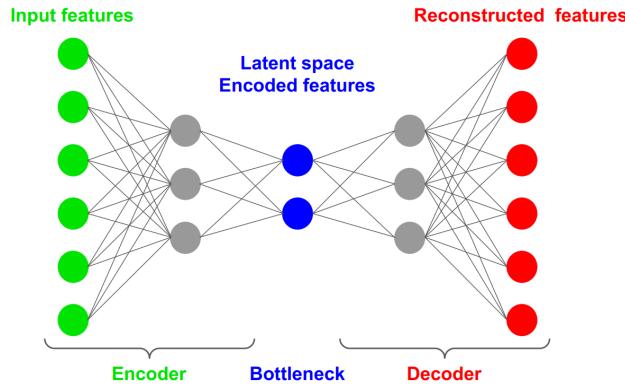


Figure 6: Neural Network Architecture.

The specific loss function that is used for an example x is of the form,

$$L(x) = \|g(f(x)) - x\| \quad (1)$$

where $\|\cdot\|$ is a distance.

AEs perform quite well for anomaly detection of SVJs in the presence of QCD background jets [6]. However, they have a relatively low Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) for SVJ vs top quark jets.

This happens due to the problem of out-of-distribution reconstruction (OOD) as depicted in figure 7.

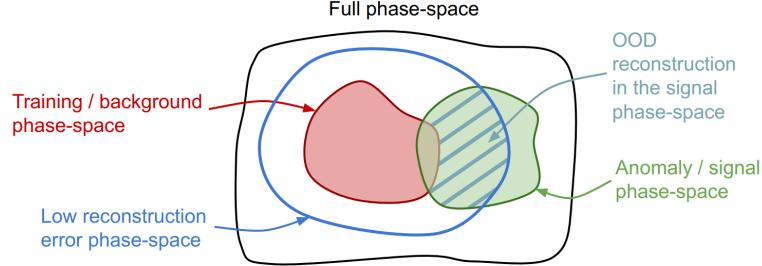


Figure 7: Problem of out-of-distribution reconstruction (OOD).

To understand OOD reconstruction, observe that in figure 7, the low reconstruction error phase-space (in blue) is larger than the training phase-space, which means that some signal/anomaly (SVJs in this case) samples that may be lying in the phase-space depicted by the shaded region will also be reconstructed with low error by the AEs, and this will lead to missing the detection of SVJs in the OOD, or in other words low anomaly detection performance. This problem can be tackled using Normalized Autoencoders (NAE), which suppress the OOD reconstruction in a fully unsupervised way.

2.4.2 Normalized Autoencoders

Normalized Autoencoders (NAEs) work in a way that ensures that the low reconstruction error phase-space (blue region in figure 7) matches as closely as possible to the training/background phase-space (red coloured region) such that the OOD samples will have high reconstruction error. The NAE model probability p_θ is defined using the reconstruction error E_θ such that it assigns high probability to the low reconstruction error samples.

$$p_\theta = \frac{1}{\Sigma_\theta} \exp(-E_\theta(x)) \quad (2)$$

Samples are generated according to the distribution p_θ using a Markov Chain Monte Carlo algorithm (MCMC) and are termed *negative samples* (x'). The loss function is the difference between the reconstruction error of the training/positive samples and that of the negative samples.

$$\mathbb{E}_{x \sim p_{\text{data}}} [L_\theta(x)] = \mathbb{E}_{x \sim p_{\text{data}}} [E_\theta(x)] - \mathbb{E}_{x' \sim p_\theta} [E_\theta(x')] \quad (3)$$

The first term in the loss function given in equation 3 is the ‘positive energy’, and the second term is the ‘negative energy’. The positive energy is simply the reconstruction error over the training data set, and the negative energy is the reconstruction error over the negative samples from the probability distribution p_θ . This loss function aims to bring p_θ as close as possible to p_{data} . Training is a min-max problem, i.e., the positive energy needs to be minimized while the negative energy has to be maximized. A known problem in such energy-based models is that the negative energy diverges. A solution to this problem should aim to prevent the divergence of the negative energy while keeping the energy difference close to 0. This is achieved by a modified loss function given by,

$$L = \log(\cosh(E_+ - E_-) + \alpha E_+) \quad (4)$$

where α is a hyper-parameter.

2.4.3 Unsupervised semivisible jet tagging

During training, ten independent NAEs are trained on the background dataset. The backgrounds of relevance for this project are top quark jets. To measure the distance between the positive and negative samples, the Energy Mover’s Distance (EMD), which is a measure of dissimilarity between two distributions, is computed. In figure 8, the bottom left plot is the EMD between negative and positive samples, and the top left plot has the reconstruction errors on negative (blue), background (red) and signal (green) samples. As the positive energy is minimized beyond a specific value, the EMD begins to increase. This means that the NAE cannot better reconstruct the training samples and suppress the OOD reconstruction simultaneously after this point. The best epoch is chosen just before the EMD increases, i.e., when the OOD reconstruction is minimal and training sample reconstruction is maximal. This is a fully signal-agnostic procedure to train an NAE without using an SVJ signal simulation.

2.4.4 Background jets

There is an overwhelming amount of background jets that the NAE has to be trained on so that it can accurately tag SVJs or any other anomalous jets in the signal region. The different possible background jets are,

- QCD multijet
- $t\bar{t}$ jets

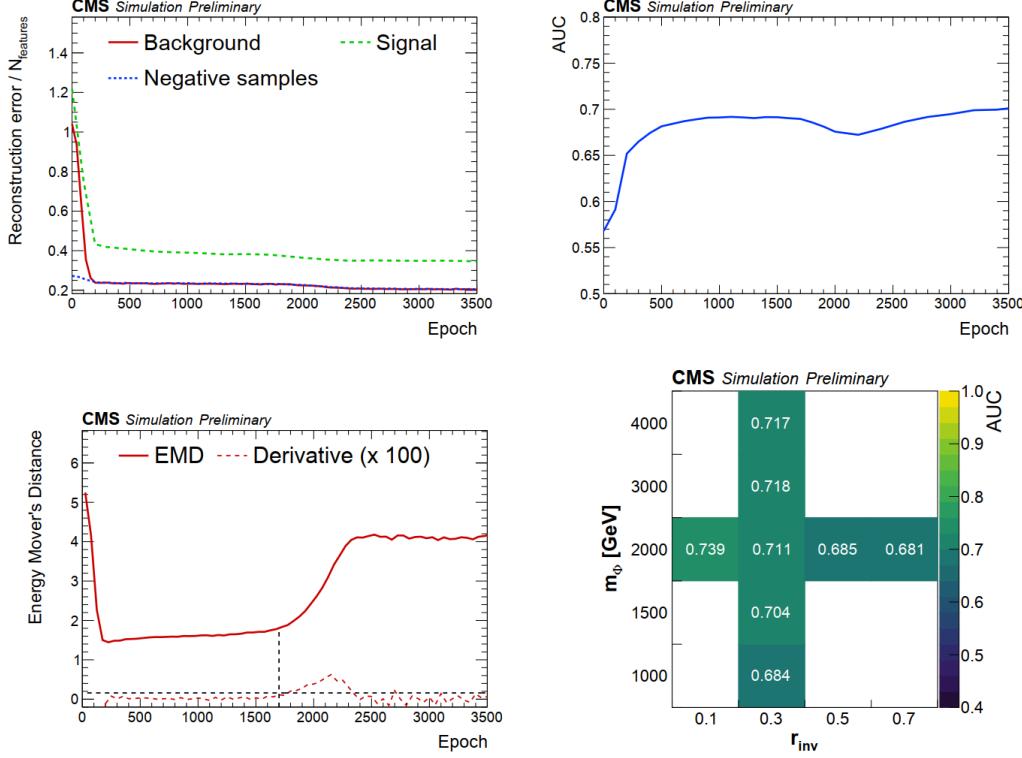


Figure 8: Example of a NAE training. Top left - Reconstruction error of background, signal and negative samples. Top right - AUC evolution. Bottom left - EMD of negative and positive samples as a function of epoch number for a given training. Bottom right - Average AUC of 10 independent trainings. Images obtained from [8].

- $Z + \text{jets}$
- $W + \text{jets}$

In this work, we want to define a training region for $t\bar{t}$ jets and train our NAE on these background jets. Thus, in the following sections, we are focused on the different decay channels of the top quark (t).

3 Analysis

3.1 $t\bar{t}$ events categorization

Top quarks are heavy and short-lived, so they decay via the weak interaction before they can hadronize and form bound states. The top quark decays to a W boson and, most frequently, a bottom (b) quark. The W boson can further decay into a quark-antiquark pair or a lepton-neutrino pair. The decay channels are illustrated in figure 9.

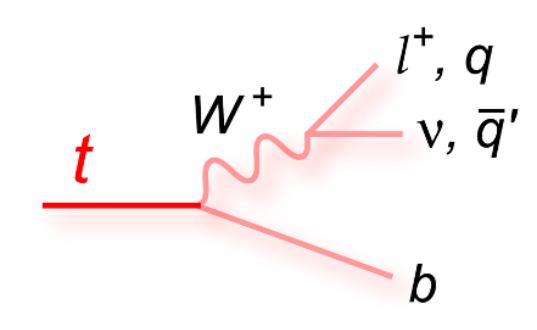


Figure 9: W boson decay channels

Since the W boson can decay into either quark-antiquark or lepton-neutrino pairs, this results in a top-antitop pair having three possible decay channels,

- **Fully hadronic** when both top quarks decay into u,d,s or c quarks
- **Semi-leptonic** when one top decays into a quark-antiquark pair and the other into a lepton-neutrino pair
- **Di-leptonic** when both top quarks decay into lepton-neutrino pairs

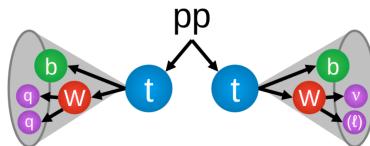


Figure 10: Semi-leptonic channel of $t\bar{t}$ decay where the left top quark decays into a W boson which further decays into a quark-antiquark pair. The right top quark similarly decays into a W boson and b quark, and the W further decays into a lepton-neutrino pair.

A top quark that decayed leptonically can be categorized further into,

- **Electron** and electron antineutrino
- **Muon** and muon antineutrino
- **Leptonic τ** decay - into tau neutrino + muon/electron + muon/electron antineutrino
- **Hadronic τ** decay - into charged or neutral pions

We could categorize $t\bar{t}$ events by looking at the PDG ID, status and mother index of GEN particles and classifying them into three categories - fully hadronic, semi-leptonic or di-leptonic. We further classified the semi-leptonic decays into four categories - electrons, muons, leptonic τ decay or hadronic τ decay. The fractions of each category have been summarized in table 1 and 2. The pre-selection region is described later in section 3.3.2.

Region	Fraction of $t\bar{t}$ events [%]		
	Hadronic	Semi-leptonic	Di-leptonic
Pre-selection	73.96	24.57	1.46
Pre-selection + $\cancel{E}_T > 200\text{GeV}$	10.59	82.59	6.73
Pre-selection + $\cancel{E}_T > 400\text{GeV}$	1.03	89.16	9.68
Pre-selection + $\cancel{E}_T > 600\text{GeV}$	0.78	89.92	9.30

Table 1: $t\bar{t}$ event categorization

Region	Fraction of semi-leptonic $t\bar{t}$ events [%]			
	Electron	Muon	Leptonic-tau	Hadronic-tau
Pre-selection	27.23	8.31	7.53	56.93
Pre-selection + $\cancel{E}_T > 200\text{GeV}$	22.38	8.19	8.93	60.50
Pre-selection + $\cancel{E}_T > 400\text{GeV}$	19.68	7.81	9.41	63.10
Pre-selection + $\cancel{E}_T > 600\text{GeV}$	16.38	7.76	9.48	66.38

Table 2: Semi-leptonic decay channel categorization

Semivisible jets are hadronic showers, so we veto isolated leptons from the events in the pre-selection region. However, not all leptons are reconstructed very well and are thus left behind in the events. Alongside these leptons, neutrinos/antineutrinos are generated, which cannot be detected and result in true missing transverse energy (MET) \cancel{E}_T . This is the reason for an increase in the number of semi-leptonic decays as the \cancel{E}_T is increased in the pre-selection region, as seen in table 1.

3.2 Top quark jet categorization

In the case where the top quark decays hadronically via the W boson, there will be three jets due to the production of one b quark and a quark-antiquark pair (u,d,s or c) from the W boson. In the case of the di-leptonic or semi-leptonic channels, there will be one jet due to the b quark produced along with the W boson. In the case of hadronic top quark decay, there are three possible cases based on the transverse momentum p_T of the top quark,

- Unmerged hadronic top - 3 *light flavour jets*
- Partially merged hadronic top - 1 *boosted W jet* and 1 *b jet*
- Fully merged hadronic top - *boosted top jet*

In addition to these three cases, QCD jet production initiated by a quark (excluding the top quark) or a gluon may also occur.

We developed an algorithm to check which category the top-quark jets fall into. This was done by identifying the light flavour quarks (u,d,s,c) coming from a W boson. The ΔR was computed between all hard truth-level partons and the reconstructed jet. ΔR is the distance in the azimuthal angle-pseudorapidity ($\eta - \phi$) plane. If all three truth-level quarks from the top jet have $\Delta R < 0.8$, then the reconstructed jet is classified as a boosted top jet. If only two truth-level quarks from the W boson have $\Delta R < 0.8$, then the jet is classified as a boosted W jet. If neither of the above two conditions is satisfied, it is an unmerged hadronic top jet. The reconstructed jets in the events were categorized depending on how many jets from $t\bar{t}$ satisfied the match condition. We divided the jets into five different categories, in bins of the transverse momenta p_T , summarized in tables 3 & 4 for the training and pre-selection region, respectively.

p_T bins	Fraction of top-quark jets [%]				
	QCD	Boosted top	Boosted W	b	Light flavours
$p_T < 200$ GeV	58.85	5.78	1.49	21.53	12.35
$200 < p_T < 400$ GeV	58.73	5.24	1.96	23.54	10.54
$p_T > 400$ GeV	55.34	13.77	1.77	21.42	7.71

Table 3: Top jet category fractions in Training Region

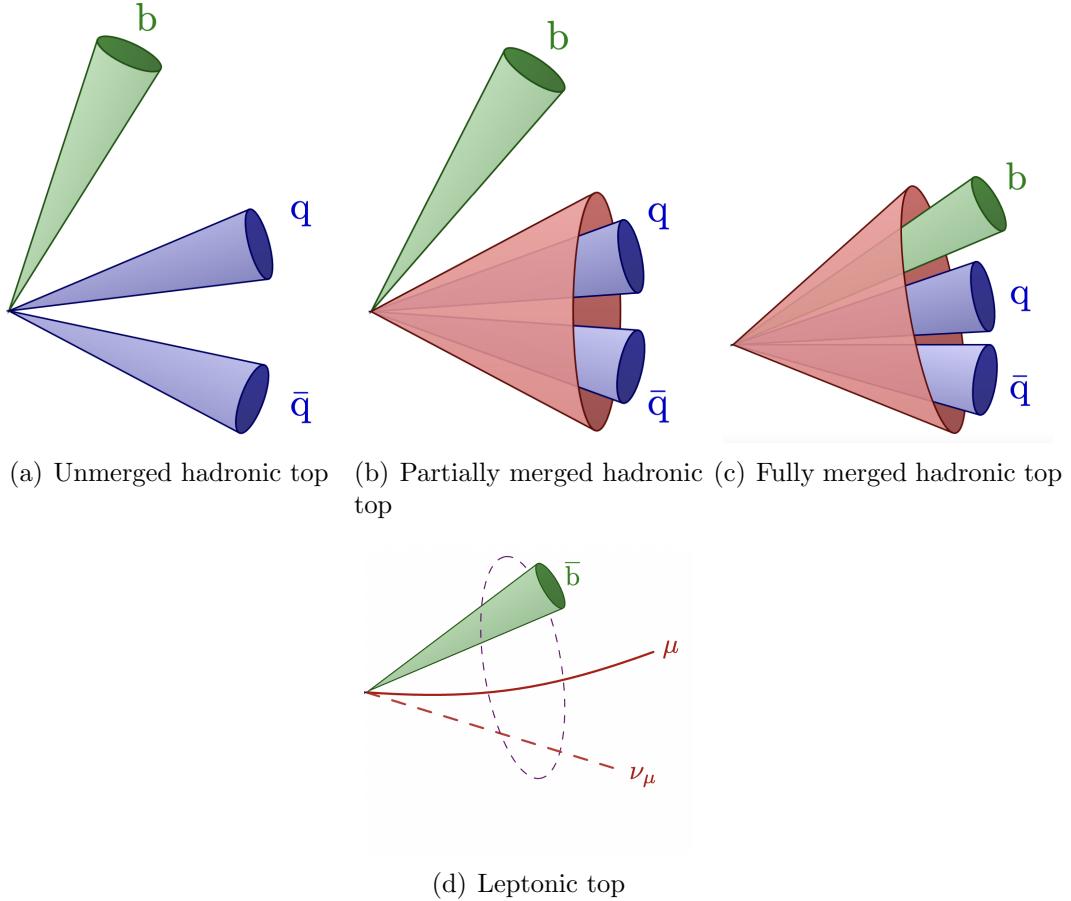
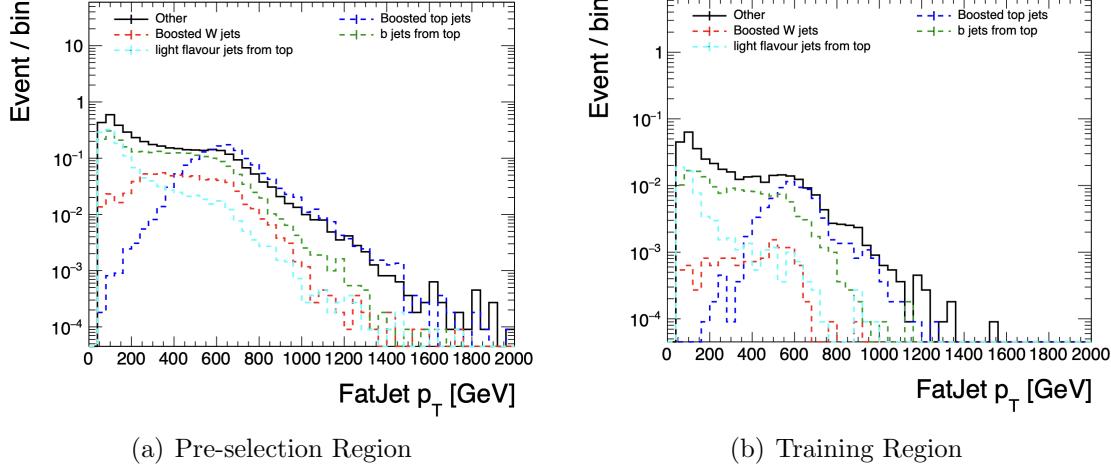


Figure 11: Illustration of different top jet categories

p_T bins	Fraction of top-quark jets [%]				
	QCD	Boosted top	Boosted W	b	Light flavours
$p_T < 200$ GeV	44.63	8.5	5.84	24.33	16.7
$200 < p_T < 400$ GeV	44.81	7.77	7.79	25.78	13.85
$p_T > 400$ GeV	39.35	17.16	7.29	24.02	12.18

Table 4: Top jet category fractions in Pre-selection Region

Figure 12: FatJet p_T distribution of different jet categories

3.3 Data Samples

In this work, we are interested in the anomalous detection of SVJs vs top quark jets. The considered top quark jets dataset contains a mix of b jets, light quark jets from fully unmerged hadronic top decay, boosted W jets from partially merged hadronic top decay, boosted top jets from fully merged hadronic top decay as well as initial state radiation gluon jets from the $t\bar{t}$ production, explained in detail in section 3.2.

3.3.1 Generation

The events used are simulated data. The background $t\bar{t}$ events as well as the signal events are generated at the parton-level at $\sqrt{s} = 13$ TeV with **MadGraph5_amc@nlo** [14], final state quarks are hadronized with **PYTHIA** [15] and events are reconstructed with the CMS detector using **GEANT4** [16]. For the simulation of hadronization in the dark sector, the Hidden Valley module in Pythia8 is employed.

This study uses AK8 jets with $R = 0.8$ from the jet axis because semivisible jets are expected to be wider than SM jets. The ten independent NAEs used for training take eight jet substructure (JSS) input variables, mapped to a normal distribution: jet major and minor axes, first energy flow polynomial, energy correlation function, transverse momentum dispersion, soft drop mass, 2- and 3-subjettiness. The NAE architecture is a simple, fully connected neural network with 10, 10, 6, 10, 10 neurons. The background events are split into 85% for

training and 15% for evaluation. The training set is further divided into 70% training examples and 15% validation examples. Signal events are not used for training.

3.3.2 Pre-selection Region

A set of loose selection criteria is applied to the training and testing samples to remove obvious backgrounds, for example, events with four isolated leptons, while keeping most of the signal. The set of selections made at the pre-selection level is summarized in table 5. The $\cancel{E_T}$ filters are made such that events with artificial MET due to detector mismeasurement can be rejected. One or more isolated leptons are vetoed in this region, as we do not expect SVJs to have them. However, it was found that many semi-leptonic events were still present in the pre-selection region, also known as ‘lost’ leptons. The fraction of semi-leptonic events for the pre-selection region with $\cancel{E_T}$ filters is summarised in table 1.

Selection	Rel. Eff [%]
Trigger	6.76
$\cancel{E_T}$ filters	99.9
Lepton Veto	35.5
$S_T > 1300$ GeV	9.34
$n_{\text{FatJet}} \geq 2$	99.9
Total abs. eff. [%]	0.24

Table 5: Cuts for $t\bar{t}$ jets in pre-selection region

3.3.3 Top jet Training Region

The training region is designed to be a signal-free region with SM jets with the same jet substructure as SM jets in the signal region to train the NAE directly on data later. In this study, we want to design a signal-free top quark jet training region for which we use semi-leptonic $t\bar{t} +$ jets events and use the lepton as a tag and jets from the opposite side as a probe. SVJs are hadronic jets, which means they are not expected to have mini-isolated leptons, and so selecting only semi-leptonic $t\bar{t}$ events rejects signal events in the training region. We want to avoid training directly on data in the pre-selection region as it may have signal jets, thus reducing the performance of the NAE as it would learn the signal jets. So, we train in the training region, which has SM jets with the same JSS as the SM jets in the signal region; however, there are no signal jets. All the samples were binned in H_T to increase the statistics.

A set of selection criteria is applied to the training and testing samples, which is summarized in table 6.

Selection	Rel. Eff [%]
Trigger (t-channel + lepton)	28.98
$S_T > 1100$ GeV (trigger safety)	4.64
\cancel{E}_T filters	99.90
n mini-isolated lepton = 1	30.37
Veto additional lepton	94.77
Jet cleaning	–
n AK4 jets ≥ 4	86.48
n b-tagged AK4 jets = 2	21.59
Cleaning tag side AK8 jet	–
n AK8 jets ≥ 1	99.82

Table 6: Selection criteria for top training region

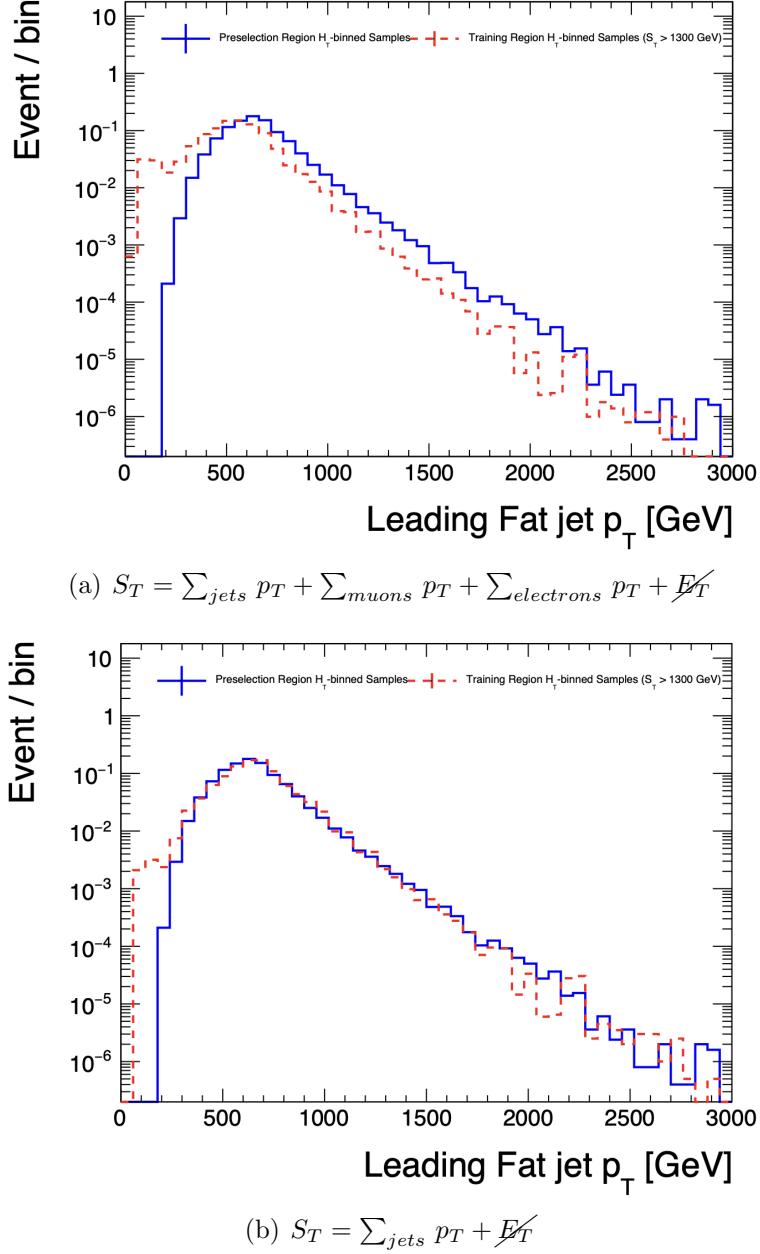
The jet substructure (JSS) variables are dependent on the jet p_T , so we have to make a S_T selection. $S_T > 1300$ GeV is well motivated as this selection is applied to define the pre-selection region. However, we make a cut at 1100 GeV instead because, upon a comparison of the JSS variables in the training and pre-selection region, their probability distributions match each other better. Now, regarding the definition of the S_T , we use the formula,

$$S_T = \sum_{jets} p_T + \cancel{E}_T \quad (5)$$

Initially, we used a different definition of S_T , which also included the electron and muon p_T ,

$$S_T = \sum_{jets} p_T + \sum_{muons} p_T + \sum_{electrons} p_T + \cancel{E}_T \quad (6)$$

However, the phase-space of the training and pre-selection region don't match very well, as seen in figure 13(a). Instead, with the S_T definition in equation 6, the histograms match much better, so we use this to make the selections.

Figure 13: Histograms of leading FatJet p_T

The JSS variables for different choices of S_T cut can be seen in figure 14, and we can see that in the case of $S_T > 1100$ GeV, the distribution of the JSS variables for the training and pre-selection region match better compared to any other S_T cut.

3.3 Data Samples

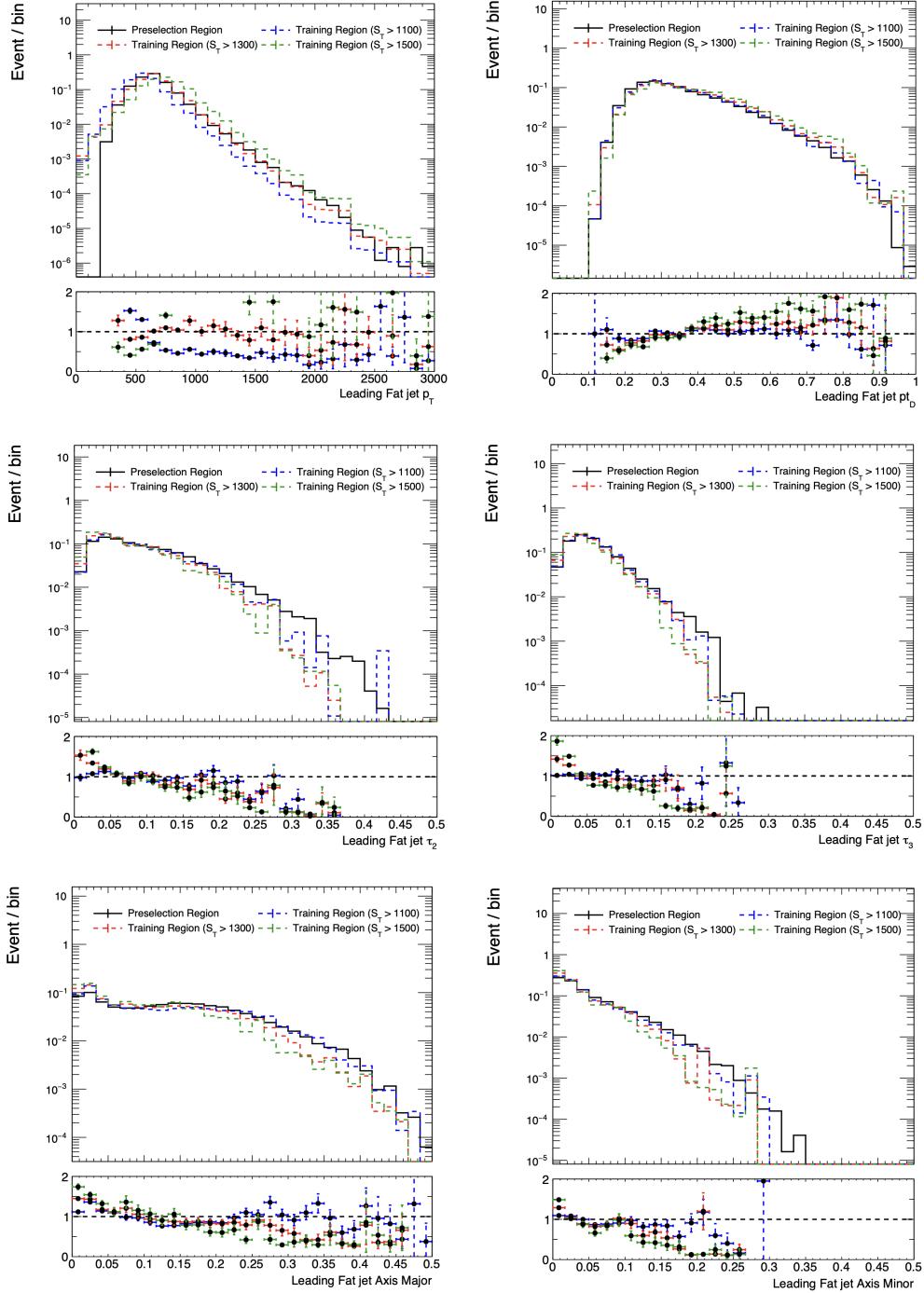


Figure 14: Histograms of JSS of leading jet for different S_T cuts

3.4 Training normalized autoencoder on $t\bar{t}$ jets

In this study, 10 independent NAEs were trained in order to keep statistical fluctuations low, and their average performance was evaluated. The NAEs were trained on top-quark jets, which are the background jets in this study.

3.4.1 Training on leading $t\bar{t}+$ jets

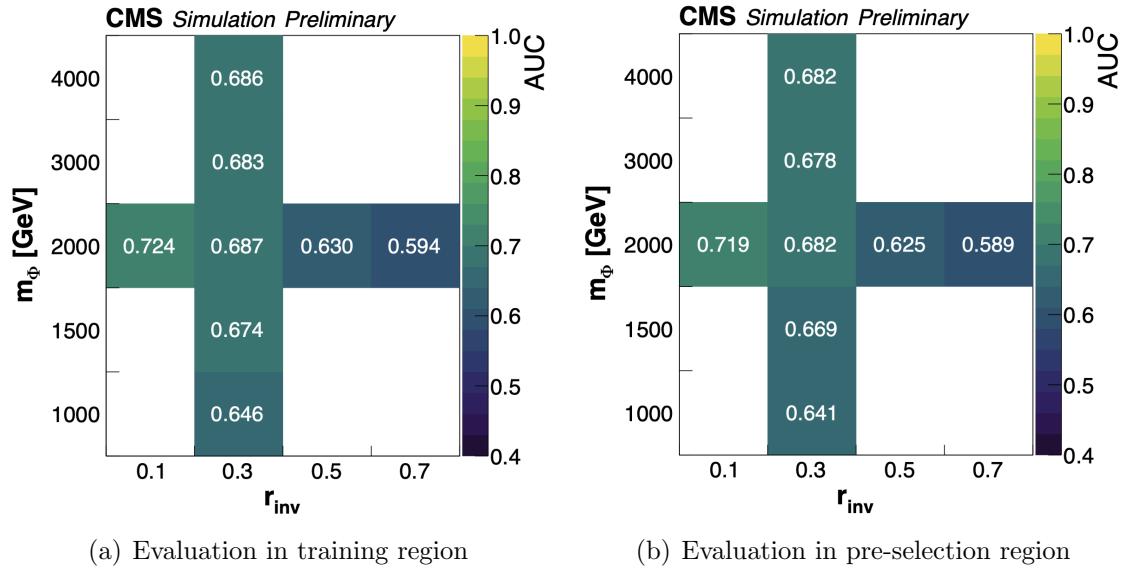


Figure 15: Training on leading $t\bar{t}+$ jets in the training region. Evaluation in the training region (left) and pre-selection region (right).

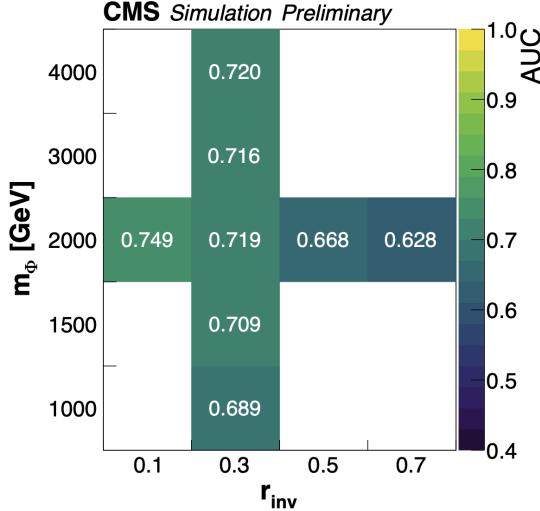


Figure 16: Training and evaluation on leading $t\bar{t}+$ jets in the pre-selection region

After training the 10 independent NAEs on $t\bar{t}+$ jets in the training region and evaluating them in the pre-selection region, we see that the average AUCs match well. This means that the NAEs were able to learn the correct JSS probability distribution to be able to classify SVJs from $t\bar{t}$ jets in the pre-selection region.

3.4.2 Training on all $t\bar{t}$ jets in bins of p_T

We do the training in bins of p_T , which helps reduce differences between training and inference regions due to correlations with jet p_T . We selected three bins to train the NAEs,

- $p_T < 200$ GeV
- $200 < p_T < 400$ GeV
- $p_T > 400$ GeV

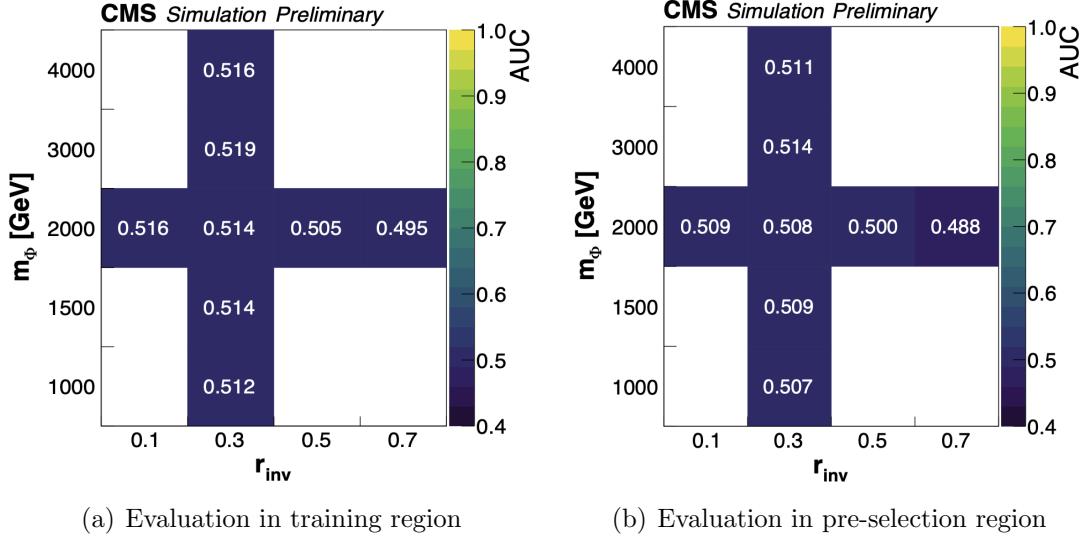


Figure 17: Training on all $t\bar{t}$ jets with $p_T < 200$ GeV in training region. Evaluation in training region (left) and pre-selection region (right).

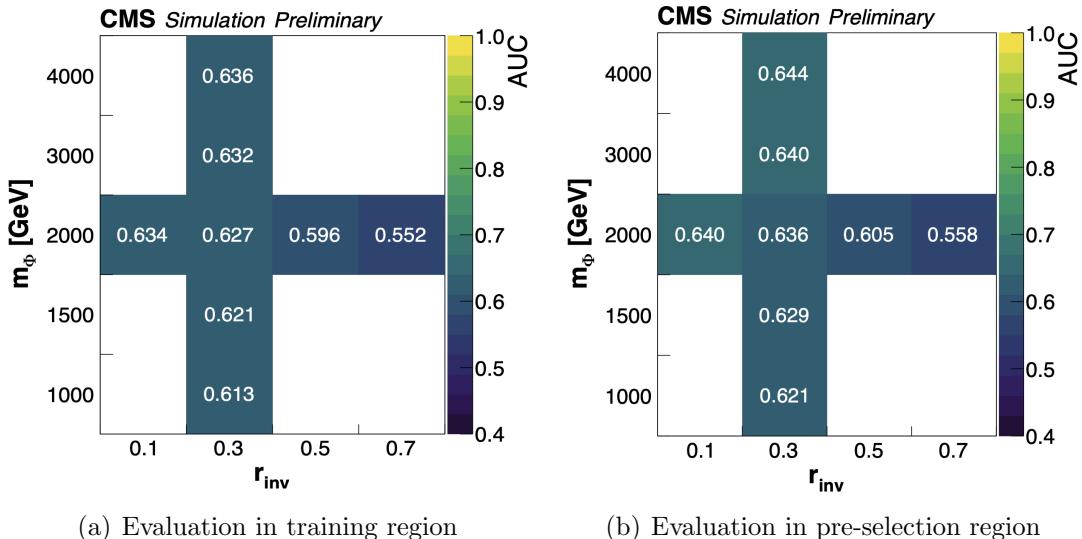


Figure 18: Training on all $t\bar{t}$ jets with $200 < p_T < 400$ GeV in training region. Evaluation in training region (left) and pre-selection region (right).

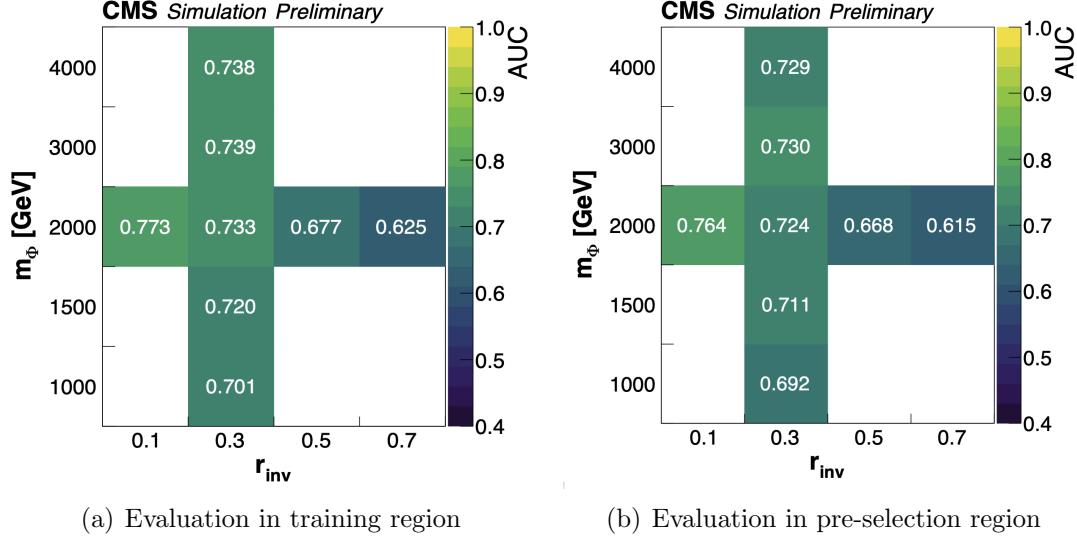


Figure 19: Training on all $t\bar{t}$ jets with $p_T > 400$ GeV in training region. Evaluation in training region (left) and pre-selection region (right).

In the region with $p_T < 200$ GeV, the NAE does not learn very well to classify between top-quark jets and SVJs. This might be because it is difficult for the NAE to distinguish between the signal and background phase-space, and it cannot perform any better than having an average AUC of ~ 0.5 . However, in the other two bins, the NAE is able to learn the JSS distribution quite well and perform the classification of semivisible jets from top-quark jets in both the training and pre-selection regions.

4 Conclusion

We designed a signal-free training region for top-quark jets by selecting semi-leptonic $t\bar{t}$ events. The jet substructure variables in the pre-selection and training regions were compared for different values of S_T cuts in order to determine the best value of S_T for the training region, which in this case was 1100 GeV. 10 independent NAEs were trained on leading $t\bar{t}$ jets in the training region and evaluated for the pre-selection region. The average AUC for the 10 independent training models was computed by taking the best epoch, which is chosen at a point just before the EMD increases, i.e., when the OOD reconstruction is minimal, and the training sample reconstruction is maximal. It was concluded that the NAE can learn the correct JSS distribution to classify semivisible jets from top-quark jets in the pre-selection region. Furthermore, the NAEs were trained in three bins of p_T , considering all jets in the events falling in these bins, in the training region and evaluated in the pre-selection region. The JSS distributions depend on the jet p_T , so training in bins of p_T helps reduce JSS differences between training and inference regions. It was found that the NAE is able to learn the JSS probability distribution very well for $200 < p_T < 400$ GeV and $p_T > 400$ GeV, but not so well for $p_T < 200$ GeV. The latter could be because the NAE is not able to distinguish between the background and signal phase space very well. Designing a signal-free training region for top-quark jets will, in the future, allow us to train the NAE directly on real data to detect anomalous jets (SVJ) from background SM jets (top-quark jets in this case).

References

- [1] M. J. Strassler and K. M. Zurek, “Echoes of a hidden valley at hadron colliders,” *Physics Letters B*, vol. 651, no. 5-6, pp. 374–379, 2007.
- [2] T. Cohen, M. Lisanti, and H. K. Lou, “Semivisible jets: dark matter undercover at the LHC,” *Physical review letters*, vol. 115, no. 17, p. 171804, 2015.
- [3] T. Cohen, M. Lisanti, H. K. Lou, and S. Mishra-Sharma, “LHC searches for dark sector showers,” *Journal of High Energy Physics*, vol. 2017, no. 11, pp. 1–32, 2017.
- [4] T. Heimel, G. Kasieczka, T. Plehn, and J. Thompson, “QCD or what?,” *SciPost Physics*, vol. 6, no. 3, p. 030, 2019.
- [5] M. Farina, Y. Nakai, and D. Shih, “Searching for new physics with deep autoencoders,” *Physical Review D*, vol. 101, no. 7, p. 075021, 2020.
- [6] F. Canelli, A. de Cosa, L. Le Pottier, J. Niedziela, K. Pedro, and M. Pierini, “Autoencoders for semivisible jet detection,” *Journal of High Energy Physics*, vol. 2022, no. 2, pp. 1–17, 2022.
- [7] S. Yoon, Y.-K. Noh, and F. Park, “Autoencoding under normalization constraints,” in *International Conference on Machine Learning*, pp. 12087–12097, PMLR, 2021.
- [8] F. Eble, “Unsupervised tagging of semivisible jets with normalized autoencoders in CMS,” tech. rep., 2023.
- [9] B. M. Dillon, L. Favaro, T. Plehn, P. Sorrenson, and M. Krämer, “A normalized autoencoder for LHC triggers,” *SciPost Physics Core*, vol. 6, no. 4, p. 074, 2023.
- [10] G. Aad, X. S. Anduaga, S. Antonelli, M. Bendel, B. Breiler, F. Castrovilli, J. Civera, T. Del Prete, M. T. Dova, S. Duffin, *et al.*, “The ATLAS experiment at the CERN large hadron collider,” 2008.
- [11] K. Aamodt, A. A. Quintana, R. Achenbach, S. Acounis, D. Adamová, C. Adler, M. Aggarwal, F. Agnese, G. A. Rinella, Z. Ahammed, *et al.*, “The ALICE experiment at the CERN LHC,” *Journal of Instrumentation*, vol. 3, no. 08, p. S08002, 2008.
- [12] C. Collaboration, S. Chatrchyan, G. Hmayakyan, V. Khachatryan, A. Sirunyan, W. Adam, T. Bauer, T. Bergauer, H. Bergauer, M. Dragicevic, *et al.*, “The CMS experiment at the CERN LHC,” *Jinst*, vol. 3, p. S08004, 2008.

- [13] T. Sakuma, “Cutaway diagrams of CMS detector,” in *J. Phys.: Conf. Ser.* **513** 022032, no. CMS-OUTREACH-2019-001, 2019.
- [14] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations,” *Journal of High Energy Physics*, vol. 2014, no. 7, pp. 1–157, 2014.
- [15] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, “An introduction to pythia 8.2,” *Computer physics communications*, vol. 191, pp. 159–177, 2015.
- [16] G. Collaboration, S. Agostinelli, *et al.*, “Geant4—a simulation toolkit,” *Nucl. Instrum. Meth. A*, vol. 506, no. 25, p. 0, 2003.