Real-Time Anomaly Detection in High Energy Physics

Ryan Justin Atkin[1,2], **James Michael Keaveney**[1] **and Claire David**[2]

[1]Department of Physics, University of Cape Town, Cape Town, South Africa
[2]African Institute for Mathematical Sciences, Cape Town, South Africa

E-mail: ryan.atkin@uct.ac.za

Abstract. While searches for physics beyond the Standard Model (BSM) have yet to yield conclusive discoveries, they continue to motivate the development of more flexible, data-driven strategies. At the ATLAS experiment at the Large Hadron Collider (LHC), trigger systems are used to rapidly select potentially interesting proton-proton collisions for further analysis. Traditional triggers rely on pre-defined criteria, such as high-momentum particles, which may miss more subtle or unconventional signs of new physics. To overcome this limitation, machine learning algorithms are being developed to identify anomalous events in real time based on their overall detector signature, rather than specific features. Using unsupervised learning techniques, these algorithms learn to characterise typical collision patterns directly from the data, without input from Standard Model or BSM theory. Events that diverge significantly from these patterns are flagged as anomalous for further study. These events are stored for detailed offline analysis. This approach enables a broad and largely model-independent search for unexpected phenomena in the vast datasets of Run-3 and beyond. It could potentially reveal signals that targeted BSM searches might overlook.

1 Introduction

The Standard Model (SM) of particle physics describes all the known elementary particles and their interactions via three of the four fundamental forces of nature. The SM is a very successful theory whose predictions have been backed up by many experimental results. However, there are some phenomena the SM does not explain, alluding to the SM being an incomplete theory. These unexplained phenomena include, but are not limited to, the origin of dark matter [1], the dominance of matter over anti-matter in the universe [2], and the lack of a description of gravity. So far, direct searches for theories beyond the SM (BSM) have yet to provide any conclusive evidence for these new theories. As such, there is an increasing focus on other methods to find new physics. One of these methods is using Anomaly Detection (AD), which involves studying data that significantly deviates from the typical dataset. For example, consider a dataset that is completely dominated by hadronic objects. An event that has no hadronic objects and only electrons or muons would then clearly deviate from what all the other events contain. Using AD avoids any model assumptions, allowing the data to speak for itself, and provides the potential to find signals that the current searches cannot.

2 ATLAS trigger system

In order to study anomalous events, they first need to be found. In the ATLAS experiment [3] at CERN, triggers are used to select collisions that are likely to contain interesting physics. Given that there are around 40 million proton-proton collisions that occur within the ATLAS experiment every second, and that most of these are low-energy inelastic-scattering collisions, not all of the collisions can be saved to disk. Instead, two levels of trigger systems are used to reduce the data rate to a manageable amount, as well as select the most interesting collisions.
The first level trigger, or L1 trigger, is a hardware-based trigger with the trigger logic implemented on FPGA boards. The L1 trigger has to look at every collision to determine if there is anything interesting, and needs to make this decision within 2.5 μs. To keep within this short latency, only coarse information from the calorimeters and muon spectrometer are used to reduce the data rate from $\sim$40 MHz to $\sim$100 kHz.
The data that passes the L1 decision is sent to the High-Level Trigger (HLT). The HLT is a software-based trigger that performs a basic reconstruction of the objects, using all parts of the detector within a region where the L1 trigger fired, also known as the Region-of-Interest (RoI). The HLT triggers are seeded by corresponding L1 triggers, forming what are known as "trigger chains". For example, the muon HLT triggers will only look at data that fired one of the muon L1 triggers. Each collision can fire multiple different triggers though. So while the L1 trigger processes every collision, the HLT trigger only looks at a subset of the data. The HLT trigger then needs to make its decision within tens of milliseconds, reducing the data rate further from $\sim$100 kHz to $\sim$1 kHz. All data that pass at least one of the HLT triggers is sent off detector for permanent storage.
The current set of triggers all use standard physics signals to trigger on. Therefore, to trigger on anomalous collisions, new triggers are needed. In order to learn from the data itself and try to exploit unknown features in the data, machine learning techniques are used to create these new AD triggers.

3 Generic Event-Level Anomalous Trigger Option (GELATO)

The triggers used for anomaly detection in the ATLAS experiment are the Generic Event-Level Anomalous Trigger Option (GELATO) triggers, designed by the ATLAS AD trigger group. To classify a collision as anomalous, an algorithm needs to be able to distinguish between a typical collision and one that looks different from this. To this end, both the L1 and HLT GELATO use a form of deep neural network known as Auto-Encoders (AEs). These networks aim to reconstruct the input data as best as possible. If the typical dataset makes up 99.9% of all the data, the AE will learn to reconstruct this well. When it is given the 0.1% that is anomalous to this typical dataset, the AE will struggle to reconstruct it and thus have a large reconstruction error. The GELATO L1 trigger uses a more complex version of the AE, known as a Variational Auto-Encoder General Adversarial Network (VAE-GAN) [4]. A diagram illustrating the structure of a VAE-GAN is given in Figure 1. There are three components to a VAE-GAN, the generic AE, the variational part and the adversarial part.
A generic AE (the top section of Figure 1) has two main components, the encoder and decoder. The encoder reduces the number of dimensions at each layer down into the latent space. This reduction tries to find an efficient and compressed representation of the data, while learning the important features and removing noise. The decoder expands the dimensions back out from the latent space to the same number of dimensions as the input. The AE is an unsupervised model trained by minimising the loss function $\mathcal{L}_{MSE}$, defined by the Mean Square Error (MSE), between the outputs of the decoder and the inputs to the encoder. This loss function is shown in Equation 1, where x_i and $\hat{x_i}$ are the inputs and outputs respectively, and n is the number of inputs/outputs. In a generic AE, collisions with a large MSE would be considered anomalous.
For the variational part of the VAE-GAN (middle section of Figure 1), an addition is made to the latent space of the generic AE. Instead of the final layer of the encoder mapping directly onto the latent space, it maps onto mean (μ_z) and standard deviation (σ_z) vectors that have the same dimensions as the latent space (n_z). The μ_z and σ_z are used to create an n_z-dim standard normal distribution, from which the latent space variables are sampled. With the inclusion of the variational part comes the addition of the Kullback-Leibler (KL) divergence loss, $\mathcal{L}_{KL}$, to the total loss function. The definition of $\mathcal{L}_{KL}$ is given in Equation 2. Here, μ_k and σ_k are the components of the μ_z and σ_z vectors, and β is a parameter that is varied to control how much of an impact the KL divergence has on the total loss function. The KL divergence encourages the latent space to follow a standard normal distribution, ensuring a well-behaved latent space.
The adversarial part of the VAE-GAN (bottom section of Figure 1) includes an additional discriminator D. This discriminator is a simple Deep Neural Network (DNN) that learns to differentiate between the "real" and "fake" signals, or inputs (x_i) and outputs ($\hat{x_i}$) respectively. The discriminator is trained using the loss function $\mathcal{L}_D$, as defined in the lower section of Figure 1. The total loss function of the VAE-GAN is altered by including the adversarial loss $\mathcal{L}_{Adv}$, shown in Equation 3. Here the λ is a parameter that is varied to control how much of an impact the GAN has on the total loss function, and $D(\hat{\boldsymbol{x}})$ is the output of the discriminator when given the output of the VAE-GAN as input. The decoder tries to fool the discriminator into classifying the output $\hat{x_i}$ as real, with the discriminator acting as an adversary to the decoder. This helps to regularise the output, creating more realistic data-like reconstructions. The combination of these three loss functions are used to define the total loss function

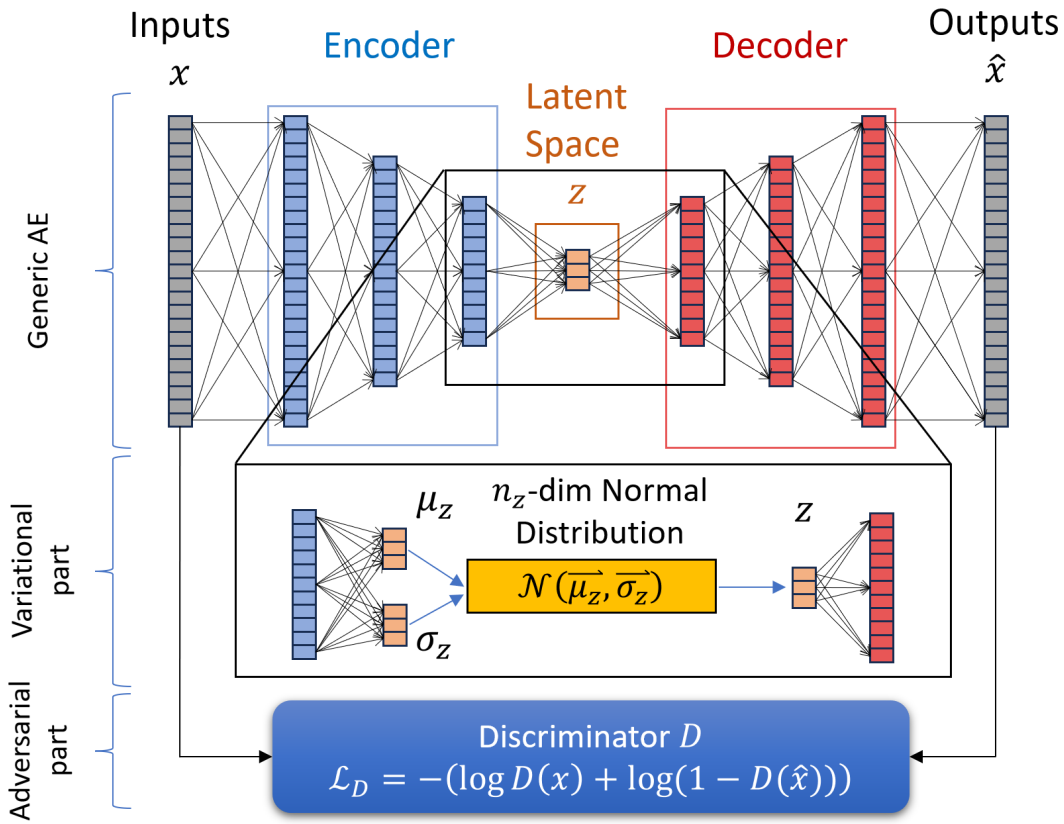

Figure 1: Illustration of the structure of a Variational Auto-Encoder General Adversarial Network (VAE-GAN). The upper section shows the generic Auto-Encoder (AE). The middle section shows the variational latent space, which includes a sampling from a standard normal. The lower section represents the GAN, or discriminator, showing the loss function used to train the discriminator.

of the VAE-GAN, as in Equation 4.

$$\mathcal{L}_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{x_i} - x_i)^2 \tag{1}$$

$$\mathcal{L}_{KL} = \beta\frac{1}{2}\sum_{k=1}^{n_z}\left(\mu_k^2 + \sigma_k^2 - \ln\sigma_k^2 - 1\right) \tag{2}$$

$$\mathcal{L}_{Adv} = \lambda \log D(\hat{\boldsymbol{x}}) \tag{3}$$

$$\mathcal{L}_{Reco} = \mathcal{L}_{MSE} + \mathcal{L}_{KL} - \mathcal{L}_{Adv} \tag{4}$$

3.1 GELATO L1

Given the required latency of the L1 triggers, and the hardware constraints, the L1 GELATO can't use the full structure of the VAE-GAN when it is implemented on the FPGA boards. Instead, the full VAE-GAN structure is used when training the L1 GELATO, but only the encoder and the μ_z vector are implemented onto the FPGA boards for use in the ATLAS experiment. Due to this simplification of the algorithm, the metric used to determine whether a collision was anomalous or not (also known as the AD score) is the clipped KL divergence $\mathcal{L}_{KL}^{clipped} = \beta\frac{1}{2}\sum\mu_k^2$. This is why the use of a VAE-GAN was necessary as it ensured that the clipped KL score would be good enough to find the anomalous collisions. This allows for a quick and efficient algorithm to be implemented on the FPGAs. The inputs to the GELATO L1 consist of 44 variables. These cover the transverse momentum p_T, pseudorapidity η, and angular direction ϕ of 15 objects: six jets, four taus, 4 muons and the missing transverse energy E_T^{miss}.

The encoder has 2 hidden layers, with the first having 32 dimensions and the second 16 dimensions, which is then reversed in the decoder. The latent space has only three dimensions. The threshold on the clipped KL divergence score used to separate anomalous from typical data is determined by requiring a unique rate of 1 kHz. The unique rate is the amount of data passing the GELATO L1 per second that do not pass any other L1 triggers.

3.2 GELATO HLT

Given that the HLT has a longer latency than the L1 triggers, the algorithm has more time to perform inference, allowing a more computationally intensive algorithm to be implemented in the ATLAS experiment. The GELATO HLT therefore uses a full generic AE (top section of Figure 1) for both the training and implementation of the trigger. Since the output of the decoder is used to calculate the score, there is no need for the use of the variational and GAN parts to ensure a well behaved latent space.

The inputs to the GELATO HLT are also slightly different, using 47 inputs covering the p_T, η and ϕ 16 objects: six jets, three electrons, three muons, three photons, and E_T^{miss}. The encoder is larger as well, with four layers with dimensions of 100, 100, 64 and 32, with these reversed for the decoder. The latent space has a dimension of four. In order to reduce the noise, a minimum p_T selection is applied to the objects. All jets need to have $p_T > 50$ GeV, while the electrons, muons and photons need $p_T > 30$. If an object fails these requirements, they are set to zero. Both the loss function and the AD score for the GELATO HLT are the masked MSE. This is just Equation 1, but only using the non-zero x_i and $\hat{x_i}$. The GELATO HLT threshold is determined such that the unique data rate is 10 Hz.

3.3 Training and testing data

The data used to train both the GELATO L1 and HLT was the Enhanced Bias (EB) [5] dataset. This is real data collected by the ATLAS experiment during the 2024 data taking period. A minimal set of L1 triggers, spanning a wide range of energies and objects, was used to select this data. Each event in the dataset has a dedicated weight mostly based on the prescales of the triggers that are used to "unbias" the data. This provides a dataset that contains rare events while maintaining an unbiased spectrum of events. The EB dataset is also used to determine the thresholds on the AD scores based on the unique rate of EB data with AD scores above the threshold.

A broad selection of Monte Carlo (MC) samples are used as signal models to demonstrate the model independent nature of the GELATO triggers. They represent both Standard Model (SM) and Beyond the Standard Model (BSM) processes, and are also known to struggle with the standard ATLAS triggers. The MC samples are summarised in Table 1 [6].

Data	
Enhanced Bias	Data collected in 2024 by the ATLAS experiment using a minimal set of L1 triggers
SM MC processes	
$Z \to \nu\nu$ (b filter)	Z decaying to $\nu\nu$ with a b filter
VBF $hh \to bbbb$	VBF hh decaying to $b\bar{b}b\bar{b}$
BSM MC processes	
HAHM ggF: $h \to Z_d Z_d \to 2l2\nu$	Hidden Abelian Higgs Model (HAHM) with $m_Z d$ = 28 GeV produced via ggF h
HNL$\to e\mu\nu$	Heavy Neutral Lepton (HNL) with m = 7.5 GeV, $c\tau$ = 1 mm
ggF $h \to$SUEP$\to$full-had	Soft Unclustered Energy Pattern (SUEP) produced via ggF h
VBF $h \to aa \to 4b$	aa produced via VBF h with m_a = 55 GeV, τa = 1 ns
ggF $h \to aa \to 4b$	aa produced via ggF h with m_a = 16 GeV, τa = 10 ns

Table 1: Summary of the data used for training the GELATO triggers, and the MC samples used for testing the GELATO triggers [6]. The MC samples are split by whether they are SM or BSM processes.

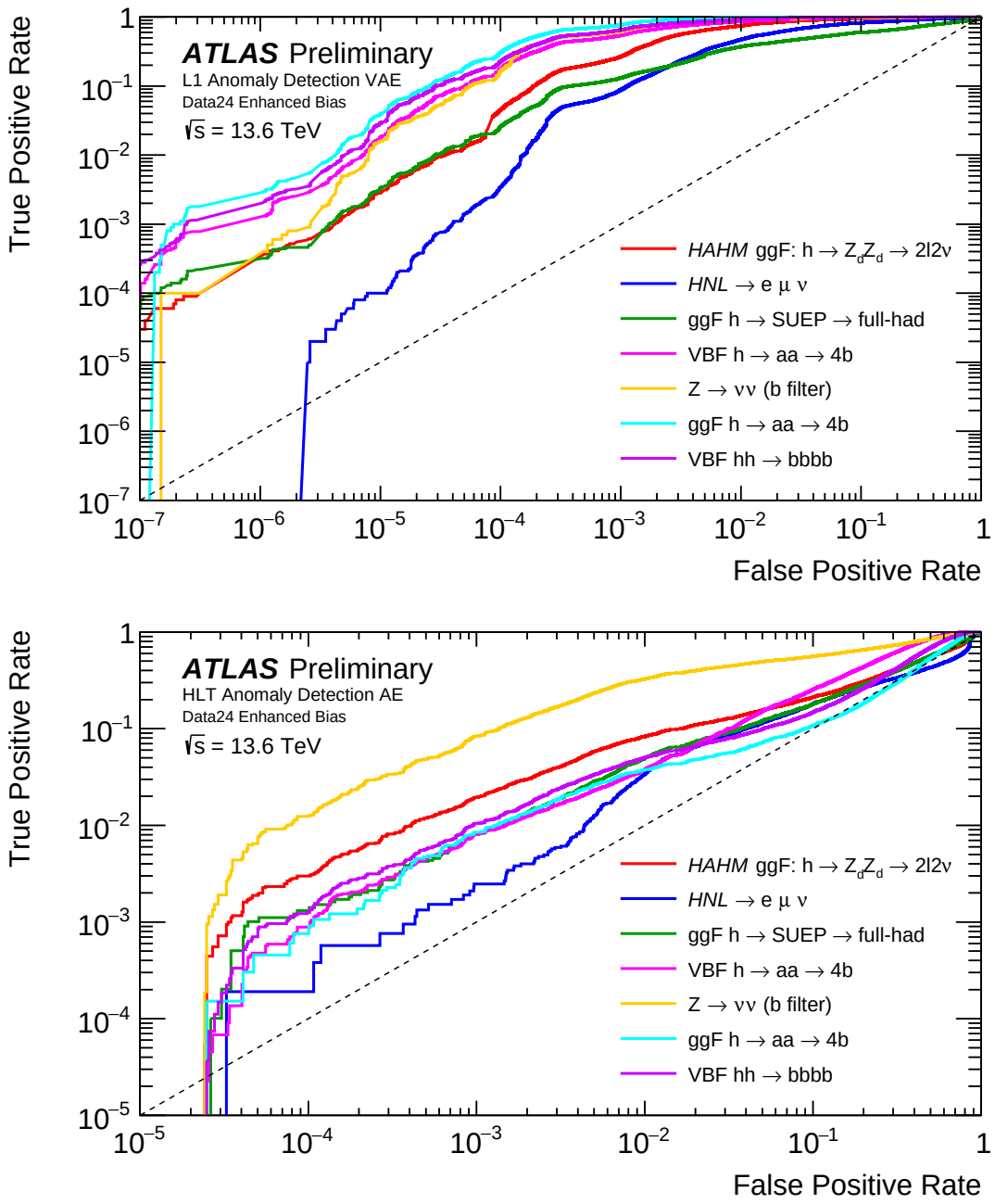

Figure 2: The ROC curves for the GELATO L1 (top) and GELATO HLT (bottom) [6]. The y-axes show the TPR defined as the weighted yield of MC signal events passing the GELATO triggers at various AD score thresholds. The x-axes shows the FPR defined as the weighted yield of EB background events passing the GELATO triggers at various AD score thresholds. The diagonal dashed lines represent using a 50% probability to classify the event as either anomalous or typical. The coloured lines are the different MC samples used as signal proxies, covering both SM and BSM processes. All events used in the calculations for the lower plot passed the GELATO L1. The descriptions of the different MC samples are given in Table 1.

4 Discussion

The Receiver Operating Characteristic (ROC) curves for both the GELATO L1 and HLT are shown in Figure 2. For these ROC curves, the MC processes are considered the different signal processes, while the EB dataset is considered the background. For every AD score obtained in the signal datasets, the weighted fraction of those samples above that AD score determines the True Positive Rate (TPR). The same is done for the EB dataset, which gives the False Positive Rate (FPR). For the given AD scores, the TPR is plotted against the FPR to evaluate the signal-background discrimination of the GELATO triggers. If the GELATO triggers classified purely randomly, the diagonal dashed lines in Figure 2 would be obtained. So, if the ROC curve of a MC process is above this dashed line, it means the GELATO trigger can distinguish between the signal and background. Given that all the MC processes are above the dashed line for both the GELATO L1 and HLT, these new triggers show strong discriminating power between signal and background across a wide variety of possible processes.

A study of the correlations of the input variables at the HLT stage, with the GELATO L1 and HLT scores, was performed for the events that passed the GELATO L1. It was found that the GELATO L1 and HLT AD scores were not strongly correlated to any single variable, showing that the algorithms are learning more information about the data. The correlation of the AD scores with the η and ϕ of all the objects was negligible, all being around 1 or 2 % correlation. However, the correlation between the GELATO AD scores with the ϕ of E_T^{miss} was found to be around 10%. The highest correlations of the AD scores were with the highest p_T objects of each object class. Interestingly though, when looking at the correlations of the events that passed the GELATO HLT only and no other HLT triggers, there were higher correlations for the sub-leading objects compared to the leading objects. This suggests that while very high p_T objects are considered more anomalous, they are likely to be triggered by some of the other triggers. The uniquely anomalous events thus likely have lower p_T objects with higher multiplicities.

5 Conclusion

In order to find new and interesting physics at the LHC in a model agnostic way, the first anomaly detection triggers have been developed for the ATLAS experiment, called the GELATO triggers. These are machine learning based algorithms, using a generic auto-encoder or a VAE-GAN depending on the level of trigger. The GELATO L1 has already been implemented in the ATLAS experiment, while the GELATO HLT has been added to the trigger lists but is still under resource cost and data rate studies. For future plans, a way to calibrate the GELATO triggers needs to be determined. The common triggers use well-defined trigger-object correspondences for calibration, but with anomalous data, there are no obvious corresponding objects. In conjunction with determining a calibration method, an offline analysis is ongoing to examine these anomalous events in greater detail and assess their potential to reveal novel physics. The GELATO triggers represent a shift towards more model-agnostic physics searches and will only improve in future.

References

[1] Planck Collaboration, "Planck2015 results: I. overview of products and scientific results," *Astronomy and Astrophysics*, vol. 594, p. A1, Sep. 2016. [Online]. Available: http://dx.doi.org/10.1051/0004-6361/201527101

[2] M. B. Gavela, P. Hernandez, J. Orloff, and O. Pene, "Standard model cp-violation and baryon asymmetry," *Modern Physics Letters A*, vol. 09, no. 09, pp. 795–809, Mar. 1994. [Online]. Available: http://dx.doi.org/10.1142/S0217732394000629

[3] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider," *Journal of Instrumentation*, vol. **3**, no. 08, p. S08003, 2008. [Online]. Available: http://stacks.iop.org/1748-0221/3/i=08/a=S08003

[4] E. Govorkova et. al., "Autoencoders on field-programmable gate arrays for real-time, unsupervised new physics detection at 40 mhz at the large hadron collider," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 154–161, Feb. 2022. [Online]. Available: http://dx.doi.org/10.1038/s42256-022-00441-3

[5] ATLAS Collaboration, "Trigger monitoring and rate predictions using Enhanced Bias data from the ATLAS Detector at the LHC," CERN, Geneva, Tech. Rep., 2016, all figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-DAQ-PUB-2016-002. [Online]. Available: https://cds.cern.ch/record/2223498

[6] ATLAS Collaboration., "Public combined trigger plots for collision data," https://twiki.cern.ch/twiki/bin/view/AtlasPublic/CombinedTriggerPublicResults, 2025.