Use of a Normalizing Flow model for generating Drell-Yan events in the ATLAS collaboration at the LHC

Peter Fitzhugh[1,*], *Xiangyang* Ju[2,**], *Rosy* Nicolaidou[1,3,***] on behalf of the ATLAS Computing Activity

[1]IRFU, CEA, Université Paris-Saclay, Gif-sur-Yvette; France
[2]Scientific Data Division, Lawrence Berkeley National Laboratory, Berkeley CA; United States of America
[3]CERN, Geneva; Switzerland

Abstract. The search for the dimuon decay of the Standard Model (SM) Higgs boson represents a typical bump-hunting physics analysis performed at the Large Hadron Collider (LHC). It looks for a tiny peak created by new physics, or the Higgs boson in this case, on top of a smoothly falling SM background in the two-muons invariant mass spectrum $m_{\mu\mu}$. The background events are estimated from a data-driven side-band fit with a floating factor for normalization and a pre-determined function for the background spectrum whose parameters are constrained from systematic uncertainties. The criteria for determining the background function are based on the spurious signal, which measures the residual signal events obtained from a signal-plus-background fit to background-only simulated events. Therefore, these simulated events must have enough statistics, an order of billions of events, so that their statistical fluctuations are negligible compared to the expected number of signal events. However, generating Drell-Yan events with high-order QCD calculations and detailed detector simulation is computationally expensive. Our study uses a normalizing flow model trained on simulated events by the ATLAS experiment to generate billions of events with GPUs for the spurious signal study. Preliminary results show that the normalizing flow model accurately describes both the muon kinematic variables that is trained on and the existing correlations among these variables. This procedure can be easily adapted to other LHC bump-hunting analyses requiring high statistics of simulated events.

ATL-SOFT-PROC-2023-037
21/09/2023

1 Introduction

In 2012, the long-searched for Higgs boson, a particle deeply connected with the theoretical foundations of the Standard Model of particle physics, was discovered at the LHC, by the ATLAS [1] and CMS [2] collaborations. Since this discovery, the two collaborations have been actively studying the properties of this new particle, so far consistent with those of the Standard Model Higgs boson. While the interactions of the Higgs boson with the force carriers and the third-generation of fermions have been observed by ATLAS [3] and CMS,

*e-mail: peter.michael.fitzhugh@cern.ch
**e-mail: xiangyang.ju@cern.ch
***e-mail: rosy.nicolaidou@cern.ch

the lower limit has only been set for the $H \to \mu\mu$ decay channel by CMS out of all the second-generation fermions, but at a significance of 3 sigmas [4]. The $H \to \mu\mu$ decay channel thus offers the best opportunity to measure the Higgs boson interactions with a second-generation fermion at the LHC.

The search for the $H \to \mu\mu$ decay looks for a resonance peak on top of a falling background in the two-muons invariant mass distribution, $m_{\mu\mu}$. It is a challenging search because of the tiny signal-over-background ratio, which, for example, is at the level of 0.2% in the region of $120 < m_{\mu\mu} < 130$ GeV after the event selections in the ATLAS Run 2 analysis [5]. The dominant background is the Drell-Yan (DY) process, $Z/\gamma^* \to \mu\mu$. The statistical uncertainty of the DY background becomes a limiting factor in studying the background modelling at the level required by the small expected $H \to \mu\mu$ signal. For the needs of the above mentioned analysis, about 20 billion $Z/\gamma^* \to \mu\mu$ events were generated with GPUs in High-Performance Computers.

The High Energy Physics community has extensively studied deep generative models to simulate collision events. Popular generative models include variational autoencoders, normalizing flows (NF), generative adversarial networks, and emerging score-based models (examples of references are given in Refs. [6, 7]). In this study, a NF model is used (although other models were also tested), and it was found to be easy to train and efficient in achieving high accuracy in generating the $Z/\gamma^* \to \mu\mu$ events.

2 Normalizing Flow Model

A normalizing flow is a technique that transforms a simple base density distribution $\pi(\vec{z})$ to a more complex target density distribution $p(\vec{x})$ using a bijective, differentiable function known as a bijection $\vec{x} = f(\vec{z})$. A normalizing flow uses a chain of bijections to construct the final bijection, which allows the modelling of complex target distributions. A bijection is preferably chosen as a simple function to make the normalizing flow learnable and computationally efficient. The coefficients of the function are parameterized by neural networks (NN), often by the MultiLayer Perceptrons (MLPs). Using the change of variables technique in mathematics, a normalizing flow can estimate the target density distribution with the input vector $\vec{x}$. Therefore, we can use the Adam optimizer to update the learnable weights in the NN $\vec{w}$ to minimize the negative log-likelihood function $\mathcal{L}(\vec{w}|\vec{x})$.

Our model implementation is based on a type of normalizing flow known as the Masked Autoregressive Flow (MAF) [8]. In MAF, the bijection transforms the base density distribution by sequentially transforming each dimension based on the previously transformed dimensions. The autoregressive feature is ensured by masking the learnable weights in the NN. It depends on the ordering of the input vector and is slow for sampling. We added a permutation bijection to each MAF to minimize the ordering effect.

3 Simulated Data

Drell-Yan events are simulated with Sherpa 2.2.1 using NLO-accurate matrix elements for up to two partons and LO-accurate matrix elements for up to four partons calculated with the Comix and OpenLoops libraries and the NNPDF3.0 NNLO set. They were matched to the Sherpa parton shower using the MEPS@NLO prescription. The effects of pile-up are included in the simulation to mimic the LHC Run 2 collision conditions. All events were processed through the ATLAS detector simulation based on Geant4.

The same event selections as the ATLAS Run 2 analysis [5] were applied. The analysis then categorizes the remaining events according to the jet multiplicities. Our strategy is to train a different NF model for different jet multiplicities. In this study, only the results on the 0 jet category are shown. The target density distributions are the six muon kinematic variables (transverse momentum of each of the muon p_{T}, pseudorapidity η, azimuthal angle ϕ)[1] and the two-muons invariant mass, $m_{\mu\mu}$. Since muons may radiate a photon, losing a significant fraction of energy, up to one final-state photon per event is included in the $m_{\mu\mu}$ calculation. After all event selections, there are 1 million events left, 90% for training and 10% for testing.

4 Training Setup

In this work, we train the MAF model to generate the $Z/\gamma^* \rightarrow \mu\mu$ events. The input vector $\vec{x}$ contains the above mentioned seven variables, all scaled to be within the region of $[-1, 1]$. And the base density distribution is a seven-dimensional Gaussian distribution. The MAF model transforms the Gaussian distribution to the muon kinematic variables.

The hyper-parameters of the model are described as follows. The MLPs inside each MAF module consist of two layers of dense networks with a layer size of 128 and a ReLU activation. We chained 20 MAF modules. The model is trained for 1000 epochs with a batch size of 512. Instead of using a constant learning rate, we employed a learning rate scheduler that decays the learning rate from 10^{-3} to 10^{-5} following a power-law distribution; doing so smoothed the training loss distribution and boosted the performance. Those parameters were similar to those used in Ref. [9]. No hyperparameter optimization was performed.

The best model is chosen for testing. We use the "Wasserstein Distance" WD [10], a measure of the dissimilarity of two probability distributions, to monitor the model performance during training. The WD for a given variable is evaluated between its target and generated distributions. We define the mean Wasserstein Distance, $\overline{WD}$, as the arithmetic mean of the WD values for the seven variables. After each epoch, the $\overline{WD}$ is evaluated. Figure 1 shows the $\overline{WD}$ for each epoch and the minimum $\overline{WD}$ up to that epoch. The model that yields the minimum $\overline{WD}$ is selected.

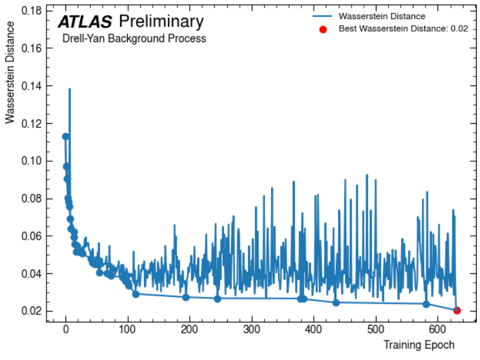

Figure 1: Wasserstein Distance as a function of the epoch number throughout the training process. The current best Wasserstein distance at each epoch is shown by the dots. The best Wasserstein Distance was found at epoch 630.

[1] ATLAS uses a right-handed coordinate system with its origin at the nominal interaction point (IP) in the center of the detector and the z−axis along the beam pipe. The x−axis points from the IP to the center of the LHC ring, and the y−axis points upward. Cylindrical coordinates (ρ,ϕ) are used in the transverse plane, ϕ being the azimuthal angle around the z−axis. The pseudorapidity is defined in terms of the polar angle θ as $\eta = -\ln\tan(\theta/2)$.

5 Results

We use the best MAF model to generate variables used in the ATLAS analysis. Those variables include the p_{T}, η, and ϕ of the leading and sub-leading muon, the two-muons invariant mass which were used to train the model, but also derived variables from the two-muons system kinetmatics like the value of the cosine of the lepton decay angle $\cos\theta^*$ in the Collins-Soper frame [11], and the p_{T} and η of the two-muons system. As directly trained to learn the muon kinematic variables and their invariant mass, the MAF model generates the seven distributions that agree with the ATLAS simulated events within the statistical uncertainties as shown in Figure 2. The small discrepancy in the ϕ distribution of the subleading muon can be improved by further tuning of the hyperparameters.

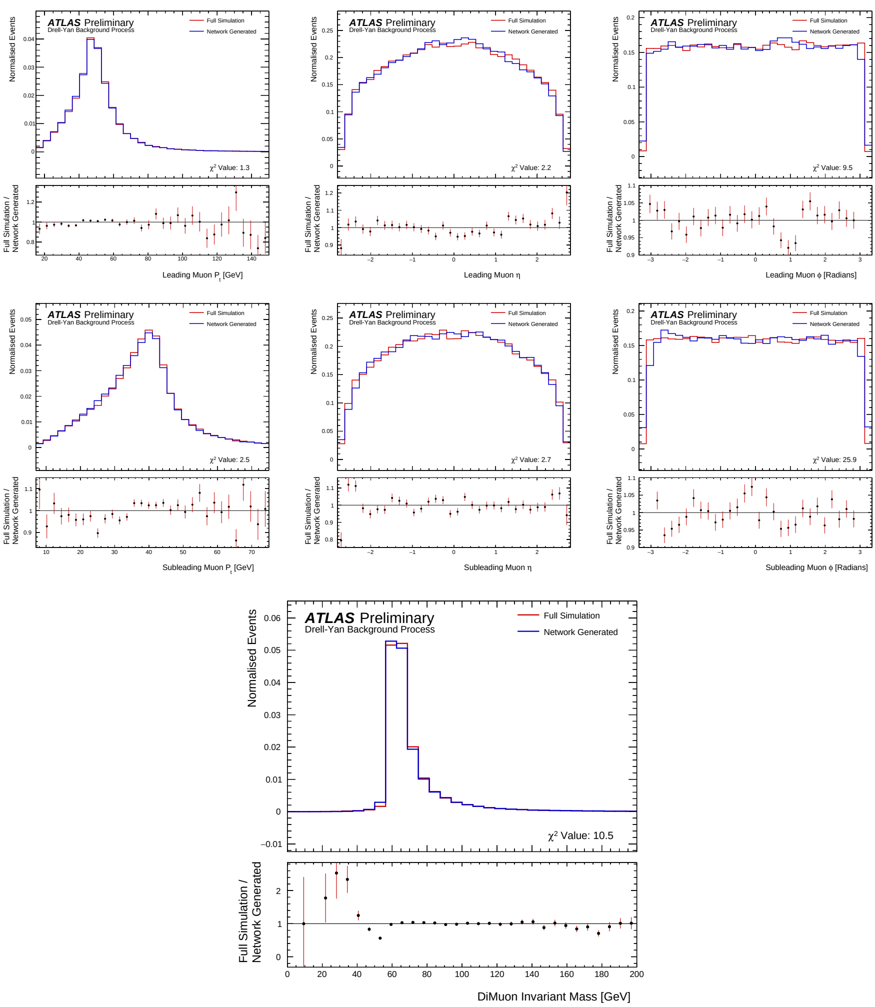

Figure 2: Comparison of the p_{T}, η, and ϕ of leading muon (top row), sub-leading muon (middle row) and the invariant mass of the two-muons (bottom row) between the ATLAS simulated events in red and the MAF generated events in blue.

Figure 3 shows the comparison of the variables calculated from the muon kinematic variables. The MAF-generated distributions agree well with the ATLAS simulated events with statistical uncertainties, indicating that the model captures the correlations between the muon kinematics accurately. To confirm the hypothesis, we directly compared the correlation coefficients between these variables for the MAF-generated events and the ATLAS simulated events in Figure 4. They agree well, with less than a 1% level of discrepancy.

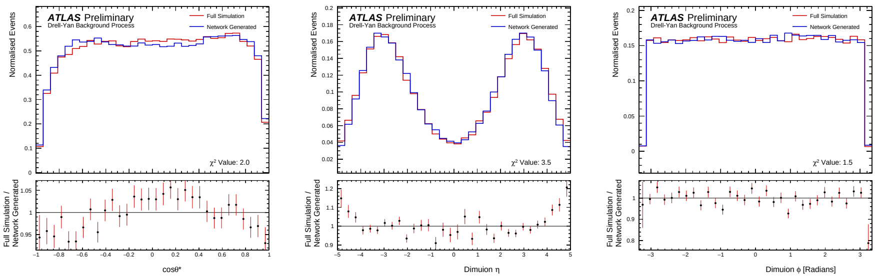

Figure 3: Comparison of $\cos\theta^*$ (variable as described in the text), the two-muons system ϕ and η distributions between the ATLAS simulated events in red and the MAF generated events in blue.

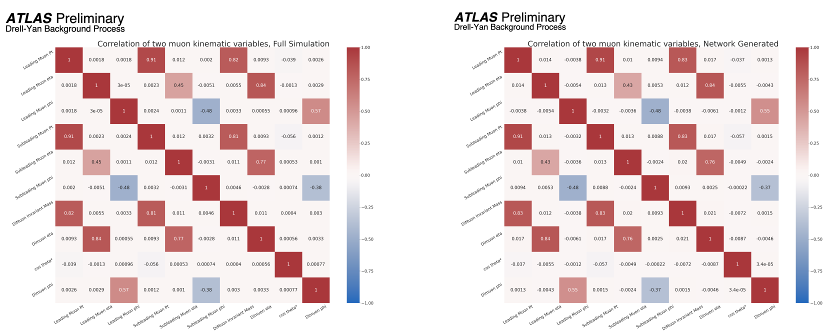

Figure 4: Correlation distribution of the seven trained-on variables and three derived variables as described in the text for the ATLAS simulated events (left) and the MAF generated events (right).

6 Conclusion

This study showcases the first attempt to generate detector-level events for the first time in ATLAS and shows promising results in simulating $Z/\gamma^* \to \mu\mu$ events. The normalizing flow model can accurately generate muon kinematic variables that are trained on but also derived observables. The correlations between the two-muons kinematics are very well captured by the model.

Training the model to reach the desired accuracy took several hours. But once the training is finished and the best model is chosen, it only takes a few minutes to generate millions of events on a modest GPU[2]. This method has the potential to increase the collaboration's ability to generate a larger number of statistics which would be a strong asset for analysis' during Run 3 at ATLAS.

This Project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 800945 - NUMERICS - H2020-MSCA-COFUND-2017.

References

[1] ATLAS Collaboration, Phys. Lett. B **716**, 1 (2012), arXiv:1207.7214 [hep-ex]
[2] CMS Collaboration, Phys. Lett. B **716**, 30 (2012), arXiv:1207.7235 [hep-ex]
[3] ATLAS Collaboration, JINST **3**, 438 (2008), S08003
[4] CMS Collaboration, JHEP **01**, 62 (2021), arXiv:2009.04363 [hep-ex]
[5] ATLAS Collaboration, Phys. Lett. B **812**, 135980 (2021), arXiv:2007.07830 [hep-ex]
[6] A. Butter, T. Plehn (2020), arXiv:2203.07460 [hep-ph]
[7] S. Badger et al. (2022), arXiv:2203.07460 [hep-ph]
[8] G. Papamakarios, T. Pavlakou, I. Murray (2017), arXiv:1705.07057 [stat.ML]
[9] A. Xu, S. Han, X. Ju, H. Wang (2023), arXiv:2303.10148 [hep-ex]
[10] C. Villani, *Optimal transport: old and new*, Vol. 338 (Springer Science & Business Media, 2008)
[11] J.C. Collins, D.E. Soper, Phys. Rev. D **16**, 2219 (1977)

[2]GPU Used: Integrated graphics of an Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz