EPJ Web of Conferences **337**, 01077 (2025) https://doi.org/10.1051/epjconf/202533701077
CHEP 2024

Normalizing Flows for Physics Data Analyses

Jan Gavranovič[1,2,*] and *Borut Paul* Kerševan[1,2,**]

[1]Jožef Stefan Institute, Jamova 39, Ljubljana, 1000, Slovenia
[2]Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, Ljubljana, 1000, Slovenia

Abstract. Monte Carlo simulations are a crucial component when analysing the Standard Model and New physics processes at the Large Hadron Collider. The goal of this work is to explore the performance of generative models for complementing the statistics of classical MC simulations in the final stage of data analysis by generating additional synthetic data that follows the same kinematic distributions for a limited set of analysis-specific observables to a high precision. Machine learning generative models were adapted for this task and their performance was systematically evaluated using a well-known benchmark sample containing the Higgs boson production beyond the Standard Model and the corresponding irreducible background. The best performing model was chosen for further evaluation with a set of statistical procedures and a simplified physics analysis. By implementing and performing a series of statistical tests and evaluations we show that a machine-learning-based generative procedure can can be used to generate synthetic data that matches the original samples closely enough and that it can therefore be incorporated in the final stage of a physics analysis with some given systematic uncertainty.

1 Introduction

In data analysis of New physics searches at the Large Hadron Collider (LHC) experiments, the use of Monte Carlo (MC) simulation is essential to accurately describe the kinematics of the known background processes in order to determine an eventual discrepancy with the measured data and attribute such a deviation to a certain new physics signal hypothesis.

Describing LHC data precisely through MC simulations involves several key steps (see e.g. [1]). Eventually, the final data analysis optimizes the data selection ("filtering") procedure to maximize measurement accuracy and new physics discovery potential (statistical significance) and determines the potential presence of new physics using statistical tests on the final data selection (for a nice overview see Ref. [2]).

Both the filtering and final statistical analysis are based on comparing the MC signal and background predictions with the real data by using several $O(10)$ kinematic variables. Obviously, the statistics of the simulated events limits the prediction accuracy of the background and signal events - ideally, the number of simulated events would exceed the data predictions by several orders of magnitude to minimize the impact of finite MC statistics on the systematic uncertainty of the final measurement. While the simulated background events can

*e-mail: jan.gavranovic@ijs.si
**e-mail: borut.kersevan@ijs.si

typically be shared between several analyses at a LHC experiment, the simulation of a chosen signal process, and the subsequent choice of the relevant kinematic observables is very analysis-specific.

To address the limitation posed by insufficient MC statistics in constraining the potential of physics analysis, a promising approach is the utilization of Machine Learning (ML), specifically focusing on deep generative modelling. The general idea is to create large numbers of events at a limited computing cost using a learning algorithm that was trained on a comparatively smaller set of MC-simulated events. The study presented here focuses mainly on autoregressive models and normalizing flows [3]. The data space in the final stage of a physics analysis is relatively low-dimensional, while the precision requirements are high. This is a perfect setting for such models, which have a needed complexity and necessary transformations to model such data samples in a computationally feasible way.

We focuses on the approach of developing a generative ML procedure for a finite set of analysis-specific reconstructed kinematic observables [4]. The generative ML algorithm is thus trained on a set of MC-simulated and reconstructed events using the kinematic distributions used in the final analysis. The requirement is to be able to extend the statistics of the existing MC using this procedure by several orders of magnitude with the generation being fast enough that the events can be produced on-demand without the need for expensive data storage. In other words, the ML algorithm will learn to model the multi-dimensional distributions with a 'surrogate' model probability density $p(\mathbf{x})$ of $O(10)$ observables $\mathbf{x}$ that can be used in the final stage of a given physics analysis.

The use of deep generative models for statistical amplification of MC samples in high-energy physics is a relatively new field, and promises to be a powerful tool in the future. This work aims to provide a systematic evaluation of: what are the actual precision requirements in terms of statistical analysis of a new physics search, and in view of these evaluates the performance of a few custom implementations of the recently trending ML models and a set of statistical tools for their evaluation.

2 The reference MC dataset

The study uses the publicly available simulated LHC-like HIGGS dataset [5] of a new physics beyond the Standard model (BSM) Higgs boson production and a background process with identical decay products in the final state and very similar kinematic features, to illustrate the performance of ML data generation in high-dimensional feature spaces.

The signal process is the fusion of two gluons gg into a heavy neutral Higgs boson H^0 that decays into heavy charged Higgs bosons $H^\pm$ and a $W^\mp$ boson. The $H^\pm$ then decays to a second $W^\pm$ boson and to a light Higgs boson h^0 that decays to b quarks. The whole signal process can be described as:

$$gg \to H^0 \to W^\mp H^\pm \to W^\mp W^\pm h^0 \to W^\mp W^\pm b\bar{b}\,.$$

The background process, which features the same intermediate state $W^\mp W^\pm b\bar{b}$ without the Higgs boson production is the production and decay of a pair of top quarks, $gg \to t\bar{t} \to W^\mp W^\pm b\bar{b}$, to a semi-leptonic final state. Events were generated assuming 8 TeV collisions of protons at the LHC with masses set to $m_{H^0} = 425$ GeV and $m_{H^\pm} = 325$ GeV.

Ignoring azimuthal angles ϕ due to the detector symmetry (giving a uniform distribution), and focusing only on continuous features results in an 18-dimensional feature space of low level (e.g. p_T) and high level features (e.g. invariant masses).

We use two different preprocessing methods: in the initial testing phase we use the logit transformation followed by standardization, and in the final evaluation we use the quantile transformation.

3 Normalizing Flows

Let $\mathbf{u} \in \mathbb{R}^D$ be a random vector with a known probability density function $p_{\mathrm{u}}(\mathbf{u}) : \mathbb{R}^D \to \mathbb{R}$. Distribution $p_{\mathrm{u}}(\mathbf{u})$ is called a base distribution and is usually chosen to be something simple, such as a normal distribution. Given data $\mathbf{x} \in \mathbb{R}^D$, one would like to know the distribution $p_{\mathrm{x}}(\mathbf{x})$ it was drawn from. The solution is to express $\mathbf{x}$ as a transformation T of a random variable $\mathbf{u}$, distributed according to a distribution $p_{\mathrm{u}}(\mathbf{u})$, in such a way that

$$\mathbf{x} = T(\mathbf{u}) , \quad \mathbf{u} \sim p_{\mathrm{u}}(\mathbf{u}) , \tag{1}$$

where T is implemented using ML components, such as a neural network. The transformation T must be a diffeomorphism, meaning that it is invertible and both T and T^{-1} are differentiable. Under these conditions, the density $p_{\mathrm{x}}(\mathbf{x})$ is well-defined and can be calculated using the usual change-of-variables formula

$$p_{\mathrm{x}}(\mathbf{x}) = p_{\mathrm{u}}\left(T^{-1}(\mathbf{x})\right)\left|\det J_T\left(T^{-1}(\mathbf{x})\right)\right|^{-1} = p_{\mathrm{u}}\left(T^{-1}(\mathbf{x})\right)\left|\det J_{T^{-1}}(\mathbf{x})\right| , \tag{2}$$

where J_T is a $D \times D$ Jacobian matrix of partial derivatives.

Invertible and differentiable transformations are composable, which allows one to construct a flow by chaining together different transformations. This means that one can construct a complicated transformation T with more expressive power by composing many simpler transformations:

$$T = T_K \circ \ldots \circ T_1 \quad \text{and} \quad T^{-1} = T_1^{-1} \circ \ldots \circ T_K^{-1} . \tag{3}$$

A flow is thus referring to the trajectory of samples from the base distribution as they get sequentially transformed by each transformation into the target distribution.

A trained flow model provides event sampling capability by Eq. (1) and density estimation by Eq. (2). The best description of the unknown probability density $p_{\mathrm{x}}(\mathbf{x})$ is obtained by fitting a parametric flow model $p_{\mathrm{x}}(\mathbf{x}; \boldsymbol{\theta})$ with free parameters $\boldsymbol{\theta}$ to a target distribution by using a maximum likelihood estimator computing the average log-likelihood over N data points.

4 Performance evaluation of the ML techniques

Histograms of event distributions of kinematic observables used in the HIGGS sample, for which the corresponding events were ML-generated, are shown in Fig. 1. One can observe that the models reliably reproduce the original distributions. We have evaluated all the models performance using divergence measures between probability distributions in one dimension and a classifier two sample test (refer to [4] for more details). The MADEMOG model was selected to be used in the statistical studies presented due to its overall best performance. This model is in fact simpler than the other models and is used as a standalone autoregressive model without the need for a flow architecture. With a more complex dataset we would expect the flow models to perform better.

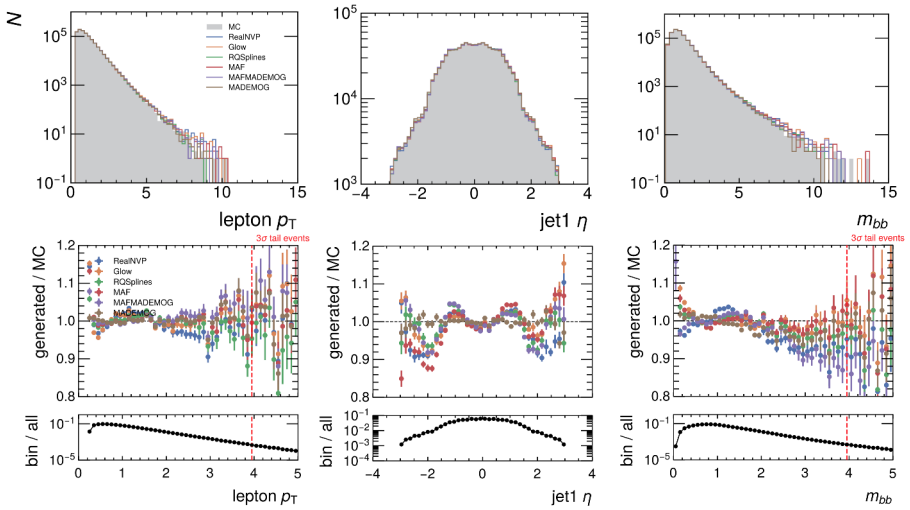

Figure 1. Distributions of ML-generated events using tested models, trained on the HIGGS dataset on its kinematic observables. The original MC distribution from this dataset is shown in grey. Only three selected variables are shown.

5 ML performance evaluation in a physics analysis

In order to provide a relevant quantitative evaluation of the impact of replacing MC-simulated distributions with ML-generated ones in an analysis environment, a simplified analysis setup was constructed, that matches a typical analysis of a new physics search in an LHC experiment and its statistical evaluation. In the following studies, the selected model trained on the HIGGS dataset was used to generate new background samples for the analysis.

In order to emulate a typical analysis procedure in an LHC experiment, a further step was introduced, namely the generative ML algorithm was trained and applied on background events in the HIGGS sample without any kinematic cuts on its observables, while the statistical analysis studies were done after an additional cut on a new ML classifier, trained to separate the signal from background events.

In this study the background yield is chosen and then the integrated luminosity with a given cross section is calculated. The evaluation was done in the background range $B \in [10^4, 10^6]$, corresponding to a luminosity increase up to three orders of magnitude ($L \in [10\ \text{fb}^{-1}, 1000\ \text{fb}^{-1}]$). The signal yield ($S$) fraction with respect to the background(B), i.e. S/B, was set to have a value of S/B=5% of the background yield in the performed tests. An inclusive relative systematic uncertainty (β), which in a real analysis would comprise both theoretical and experimental uncertainties, was set to β =10%, with the relevant subset of the studies performed also at two further values of β =5% and β =20% to cover the representative range of typical LHC analyses.

The data prediction was constructed by adding the MC-simulated samples of background and signal with a chosen signal content, giving a so-called Asimov dataset, which perfectly matches the MC-samples and is often used in the performance validation of a statistical analysis in an LHC experiment. For the purpose of the studies, the MC-simulated background is then replaced with the corresponding ML-generated samples, the production of which was previously described.

The signal prediction is retained as the MC one because in LHC analyses, the signal simulation, which generally needs a comparatively small number of generated events, is typically not as much of an issue as the background. The final analysis selection is centred around the signal prediction, retaining most of the signal statistics while only selecting the low-statistics tails of the background samples.

With the aim of closely matching a typical statistical analysis, as done in LHC experiments, the HistFactory model from the `pyhf` [6] statistical tool was used, and different standard procedures of evaluation of the agreement between data and simulation predictions were implemented. The statistical analysis was performed using two different possibilities for choosing the optimal variable, resulting in a binned distribution w.r.t. the m_{bb} in the first case and classifier score in the second case, which aims to give an optimal separation between the shape of the background and signal predictions.

The binned distributions of samples used in this statistical analysis are shown in Fig. 2 for the two relevant observables. One can see that the agreement between the Asimov data and the simulation prediction using the ML-generated background sample seems to be good, which is an encouraging starting point for a detailed analysis.

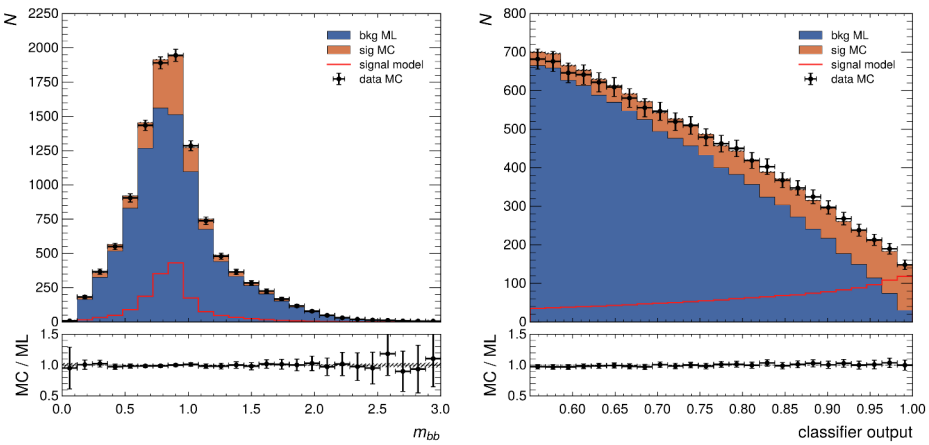

Figure 2. The m_{bb} and classifier score distributions after the classifier cut. The "MC data" (crosses) represents the Asimov data set composed from MC signal and background prediction and matches quite well the combined MC signal (orange) and ML background (blue) prediction. The red histogram shows the separate shape of the signal prediction.

Fig. 3 shows how well the expected value of the signal strength μ is reproduced in the statistical evaluation. The estimated value of μ with its uncertainty is shown w.r.t. the increase in integrated luminosity. It is evident that the statistical estimation quite reliably reproduces the expected value of $\mu = 1$ for a (small) injected signal at the fraction $S/B = 5\%$ of the background. One can observe a small bias, which of course persists with increasing integrated luminosity for the ML-generated sample. The (biased) values are well within the uncertainty, but it is of course clear that background mis-modelling, present when using ML-generated background, leads to biases and possible sensitivity loss in an analysis with relatively tiny signal presence.

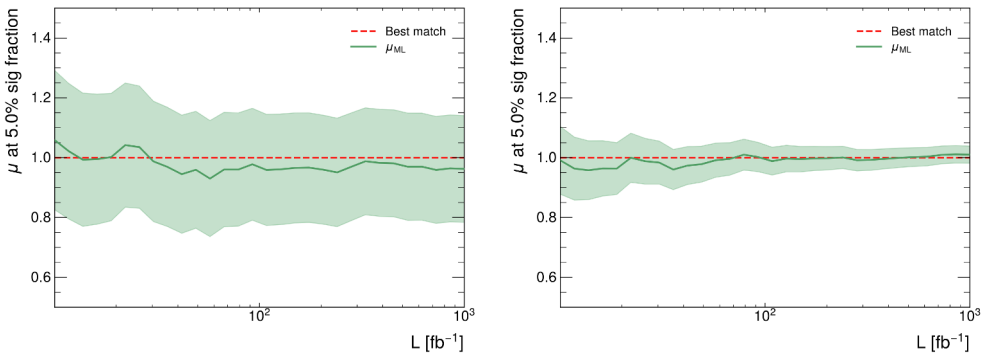

Figure 3. Fitted parameter of interest, the signal scale μ, as a function of integrated luminosity L for ML-generated background is shown for the profile likelihood fit using the m_{bb} variable (left) or the classifier score (right). The statistical error on the estimated μ value is shown with the error band around the central line of best values obtained from the likelihood fit. It is evident that the statistical estimation quite reliably reproduces the expected value of $\mu = 1$ for a (small) injected signal at $S/B = 5\%$ fraction of the background.

As the final step in this physics analysis study, one aims to evaluate the upper limits on the signal strength (μ_{UL}), together with the uncertainty estimates using the profile-likelihood-based test statistics, as is done in LHC analyses. The dependence of the extracted upper limit as a function of integrated luminosity is shown in Fig. 4 for different values of integrated luminosity and ML-generated number of events.

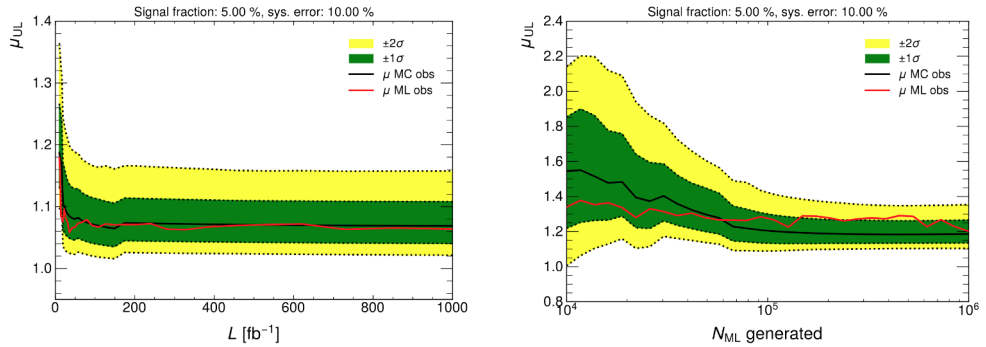

Figure 4. Upper limits on the signal strength μ for the likelihood fit to the classifier score distribution as a function of integrated luminosity L and the number of ML-generated events. The discrepancies (biases) are deemed acceptable and, one can observe that a sufficiently large value of ML-generated events is essential to get an acceptably unbiased result.

An alternative presentation of the estimated dependence of the upper limit μ_{UL} on the number of ML-generated events, as shown in Fig. 4, is given in Fig. 5, which shows the dependence of the difference $\Delta\mu_{\mathrm{UL}}$ between the upper limit values when using the MC and ML background predictions on the number of ML-generated events, as well as the injected systematic uncertainty, which is set to 5%, 10% and 20%. It is nicely visible, how the upper limit uncertainty converges to the fixed injected systematic uncertainty value, when the systematic uncertainty due to finite simulation (ML) statistics becomes negligible.

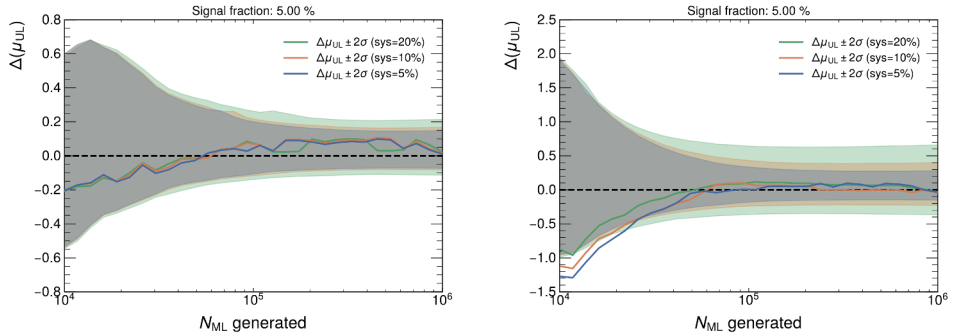

Figure 5. The difference $\Delta\mu_{UL}$ of the upper limits on the signal strength μ, as predicted by using the MC or ML-generated background, for the likelihood fit to the m_{bb} distribution (left) and the classifier score (right), as a function of the number of ML-generated events. In addition, uncertainty is also varied.

6 Discussion and outlook

We have investigated the possibility of using deep generative models for analysis-specific ML-based generation of events used by the final stage of a given particle physics analysis, where the state-of-the-art generative ML algorithms are trained on the available MC-simulated samples. The extended custom event samples can thus serve to extend the MC statistics, thereby minimizing the analysis uncertainty due to the statistics limitations of MC events or, equivalently, to smooth and minimize the uncertainties on the predicted kinematic event distributions used in the final statistical analysis of the data.

Results show that the ML-generated samples can indeed be used in a physics analysis as a surrogate model for the background prediction. Nonetheless, to further minimize the impact of the background ML mis-modelling, one would need to work on implementing techniques that go even beyond the current commercial state-of-the-art approaches, similar to the one used in this work, and to understand how to optimally adapt them for this use case in high energy physics.

References

[1] ATLAS Collaboration, The atlas simulation infrastructure. The European Physical Journal C, 70(3):823-874, 2010. https://doi.org/10.1140/epjc/s10052-010-1429-9.

[2] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. The European Physical Journal C, 71:1-19, 2011. https://doi.org/10.1140/epjc/s10052-011-1554-0.

[3] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference, 2021. https://arxiv.org/abs/1912.02762.

[4] Jan Gavranovič and Borut Paul Kerševan. Systematic evaluation of generative machine learning capability to simulate distributions of observables at the large hadron collider. Eur. Phys. J. C 84, 911 (2024). https://doi.org/10.1140/epjc/s10052-024-13284-6.

[5] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. Nature communications, 5(1):1-9, 2014. https://doi.org/10.1038/ncomms5308.

[6] Lukas Heinrich, Matthew Feickert, Giordon Stark, and Kyle Cranmer. pyhf: pure-python implementation of histfactory statistical models. Journal of Open Source Software, 6(58):2823, 2021. https://doi.org/10.21105/joss.02823.