A method for inferring signal strength modifiers by conditional invertible neural networks

Máté Zoltán Farkas[1,*], *Svenja* Diekmann[1,**], *Niclas* Eich[1,***], and *Martin* Erdmann[1,****] on behalf of the CMS Collaboration

[1]RWTH Aachen University, III. Physikalisches Institut A, Otto-Blumenthal-Straße, 52074 Aachen, Germany

Abstract. The continuous growth in model complexity in high-energy physics (HEP) collider experiments demands increasingly time-consuming model fits. We show first results on the application of conditional invertible networks (cINNs) to this challenge. Specifically, we construct and train a cINN to learn the mapping from signal strength modifiers to observables and its inverse. The resulting network infers the posterior distribution of the signal strength modifiers rapidly and for low computational cost. We present performance indicators of such a setup including the treatment of systematic uncertainties. Additionally, we highlight the features of cINNs estimating the signal strength for a vector boson associated Higgs production analysis of simulated samples of events, which include a simulation of the CMS detector.

1 Introduction

Likelihood fits in high-energy physics (HEP) can be time-consuming to perform especially with increasing number of nuisance parameters. Compared to numeric maximum likelihood approaches posterior inference with conditional Invertible Neural Networks (cINNs) is extremely time-efficient. Hence, they appear as a candidate for physics parameter inference. Since the network structure is based on normalizing flows, the resulting network model is continuous and continuously differentiable in both directions, which makes them potentially applicable in differentiable analysis workflows. Outside of HEP, these networks have already been successfully applied in several challenges, including guided image generation, image colorization and scientific model inversion [1, 2]. Apart from these, cINNs have also proven to be successful in several physics applications, such as stellar parameter estimation, cosmic-ray source property determination and detector effect unfolding [3–5]. In this paper, cINNs are applied to signal strength modifier parameter inference for a vector boson associated Higgs production analysis at the CMS experiment using simulated samples which include a simulation of the CMS detector. The performance indicators and the treatment of statistical and systematic uncertainties are going to be presented. We describe the parameter reconstruction quality, and highlight the features of the obtained prediction-truth distributions.

*e-mail: mate.zoltan.farkas@cern.ch
**e-mail: svenja.diekmann@cern.ch
***e-mail: niclas.steve.eich@cern.ch
****e-mail: martin.erdmann@cern.ch

2 Vector Boson associated Higgs Production Analyis Strategy and Setup

Inference on the signal strength modifiers has been performed in the VH ($V = Z, W$) production processes at the CMS experiment on Monte Carlo (MC) samples. These samples have been generated with HERWIG++ [6] and PYTHIA [7]. Especially, the gluon-fusion ($gg \to ZH$) and the quark initiated processes ($qq \to ZH$ and $qq \to VH$) have been considered. In total, there are three signal processes (with three signal strength modifier parameters) and 13 corresponding background processes – such as Drell-Yan production process (DY) and vector boson fusion (VBF) – with a nuisance parameter each. In the analysis workflow, the relevant final state objects from the MC samples are selected and categorized. Each sample in each category is then further categorized with a feed-forward deep neural network (DNN). This DNN returns a probability score of each event belonging to a process group. These probability scores can be histogrammed and used as an observable for the maximum likelihood fit. For the cINN, these histograms are used as input for the conditions c.

3 Neural Network Setup

3.1 Introduction to cINNs

cINNs approximate the unknown posterior distribution $p(x|c)$ of the observable x given c using the network parameters ϕ, i.e. after training

$$p_\phi(x|c) \approx p(x|c) \tag{1}$$

holds. This is done is done by minimizing the Kullback-Leibler divergence between these distributions:

$$\mathrm{KL}(p\,||p_\phi) = \mathrm{E}_{x\sim p(x|c)}\left[\log\left(\frac{p(x|c)}{p_\phi(x|c)}\right)\right] = \log\Big(p_\phi(x|c)\Big) + const. \tag{2}$$

Eq. 2 can be rewritten using the network output variables $z = f(x)$ and the continuous differentiability of the network. Enforcing z to follow a normal distribution $\mathcal{N}(z|0,1)$ this maximum likelihood expression can be rewritten into the final form of the loss function L as

$$L = \mathrm{E}_{x\sim p(x|c)}\left[\frac{z^2}{2} - \log\left|\det\frac{\partial z}{\partial x}\right|\right], \tag{3}$$

which requires the network architecture to have feasible Jacobian determinants.

The network input x and c and network output z have a corresponding node. The inputs x are mapped to the normal distributed latent outputs z. The cINN model itself consists of Affine Coupling Blocks (ACBs) and permutation layers, which remove the correlations between the inputs x. At the global optimum of the loss function the latent space distribution becomes statically independent from the input data. The ACBs receive the conditions as an additional input. The normalizing flow is implemented in these blocks, which keeps the Jacobian term easy to evaluate and guarantees the invertibility of the network. As an ACB, the GLOW coupling block [8] is used, see fig. 1, which decomposes the forward and backward transformation into two step-wise mappings with a triangular Jacobian each. For this reason, the composite Jacobian is only dependent on the product of the diagonal entries of each matrix. The mappings s_i and t_i do not need to be invertible themselves and can be parameterized with feed-forward DNNs.

During training, the network learns the mapping from the inputs x to the fixed-shape latent output space z. Upon convergence, sampling the z from the latent space and propagating these samples backwards from the network, the samples for each input's posterior distribution $p(x|c)$ can be obtained.

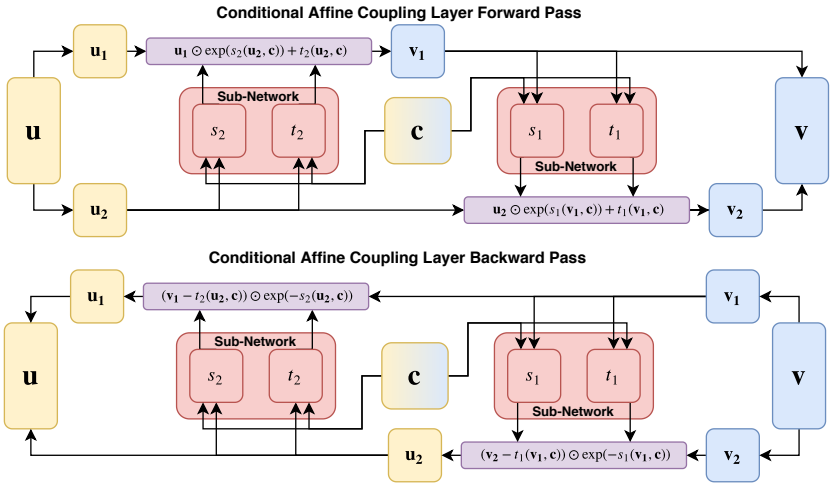

Figure 1. The GLOW affine coupling block. The outputs of the previous block u for each block are split and are processed successively ensuring the invertibility of the block. The conditions c are used as an input for both sub-networks. [3]

3.2 Synthetic Training Sample Generation

3.2.1 Parameter Prior Selection

Bayesian posterior inference requires the choice of such priors which reflect the expectations. For the signal strength modifier inference, each process has been scaled with a scaling coefficient μ drawn from the prior distributions to reflect the variations in the signal and background strength. Both statistical and systematic uncertainties have been treated in the training sample generation process.

For the signal processes, each of these μ has been drawn from a Γ distribution. This particular choice of priors enables a more refined sampling in the $\mu < 10$ region, where signal modifiers are the most expected. At the same time, the long decay of the Γ prior enables to maintain sensitivity for any $\mu \lesssim 100$. For the normalizing uncertainties, the priors have been chosen differently. The background and the luminosity nuisance parameter priors have been set to a lognormal distribution with mean 1. The latter nuisance parameter has a narrower prior and has been applied to each process equally.

3.2.2 Statistical and Shape-Changing Systematic Uncertainty Modelling

For the shape-changing uncertainties, only the templates of the 1σ up and down variations are accessible. For this reason, interpolation and extrapolation between these histograms are required. To model these uncertainties, a histogram template morphing technique [9] has been applied for each corresponding nuisance parameter.

Statistical effects originate both from the limited MC sample size and from the expectation following a Poisson distribution. After the histogram template morphing, each process bin content has been varied following a Poisson distribution to represent the former statistical effect. For the latter each resulting summed bin content has been varied similarly.

A sketch of the network setup is shown in fig. 2. In total, 16 physics signal and background modifier parameters have been used as an input with an added luminosity nuisance parameter.

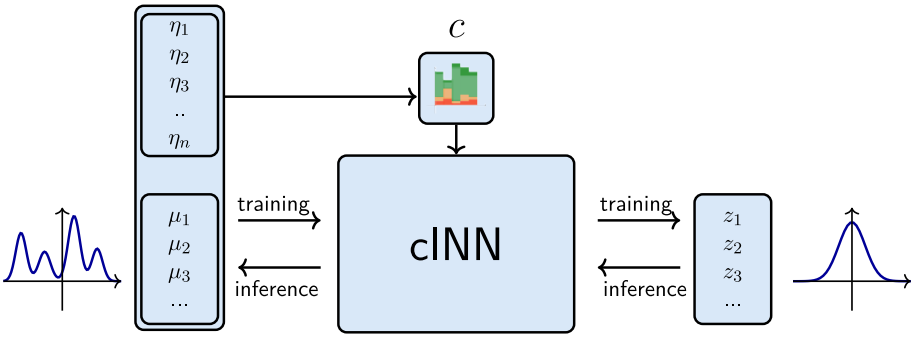

Figure 2. Sketch of the conditional invertible neural network structure, the inputs, the nuisance parameters and the latent space parameters z.

3.3 cINN Setup

The network has been trained on a dataset of 1.5 million samples and inference has been performed on 150 000 test samples. The model has been trained until both loss function values converged to a similar local optimum. The cINN model itself was implemented in PyTorch [10] using the Framework for Easily Invertible Architectures (FrEIA) library [1]. The network consists of 12 alternating GLOW ACBs and permutation layers. The feed-forward DNNs encoding the mappings s_i and t_i in the sub-networks in the GLOW blocks have 3 layers with 128 nodes with ReLU activations each. The learning rate has been gradually decreased from 10^{-3} to 10^{-5} via a cosine learning rate scheduler. As an optimizer Adam [11] has been used and a model with the lowest validation loss has been evaluated further. The model was trained for 11000 epochs.

4 Signal Strength Modifier Parameter Inference

Compared to the maximum likelihood fit, parameter inference with the trained network is swift. The performance of the network can be characterized by comparing the latent space distribution to the expected normal distribution, where any deviations from the latter lead to biases in the predictions. Inherent model biases in the predictions can also be studied through the calibration curves as well. The prediction quality can be studied by comparing the network parameter predictions to the true Monte Carlo value. Finally, comparing the model posteriors to their priors yields information about the reconstruction quality.

4.1 Latent Space Distribution

The latent space distributions for three selected output nodes are shown in fig. 3 for the test samples. The network model manages to reproduce the expected normal distribution shown in red. No strong deviations or biases can be observed in any of these distributions. Hence, the sampling from $\mathcal{N}(0, 1)$ is justified to obtain the approximated posterior samples.

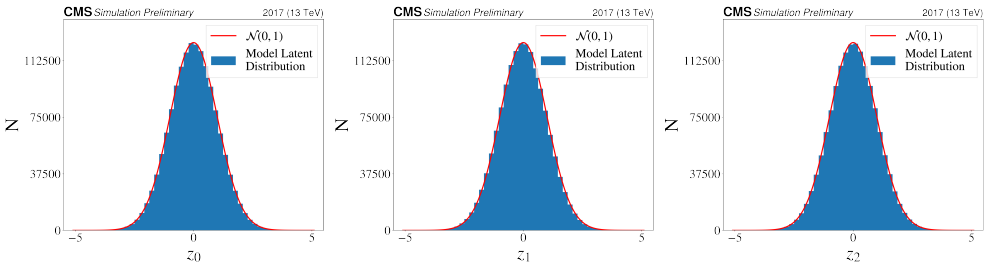

Figure 3. Latent space variable distributions for three selected z_i. The latent space distributions all closely follow the expected normal distribution in red.

4.2 Calibration Curves

The confidence of the network in the predictions can be characterized through the calibration curves. The calibration error for a given confidence interval is defined as

$$e_{\text{cal}}(q) = \frac{N_{\text{in}}}{N} - q, \tag{4}$$

where N_{in} is the total number of inferred posteriors containing the true MC value within the q quantile of the posterior. Wherever $e_{\text{cal}}(q)$ is negative is a sign of an overconfident model predicting too narrow posterior distributions; conversely, positive values of $e_{\text{cal}}(q)$ describe an under-confident model, which yields too broad posteriors. The ratio N_{in}/N as a function of q is shown in fig. 4 for four selected parameters. Most calibration errors are $e_{\text{cal}} \approx 0$, and the maximum absolute median calibration error of all parameters is $e_{\text{cal}} \approx 4\%$. For this reason, the model posteriors are expected to be scattered well around the true MC values and no strong biases are expected towards a parameter region.

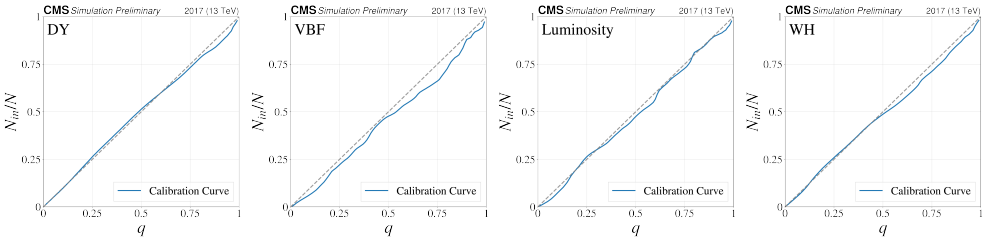

Figure 4. The calibration curves of the trained model (blue). Here, the curves of four selected input variables are shown. The fraction of histograms N_{in}/N has only minor deviations from q.

4.3 Network Predictions, Posterior Distributions and Parameter Reconstruction Quality

The network predictions are obtained by taking the mean of the obtained posteriors distributions for a given condition c. The uncertainty on these values was determined from the 68% quantiles of posteriors. The comparison of these predictions to the true values is shown in fig. 5. Depending on the reconstruction quality, the parameters can be grouped into three categories:

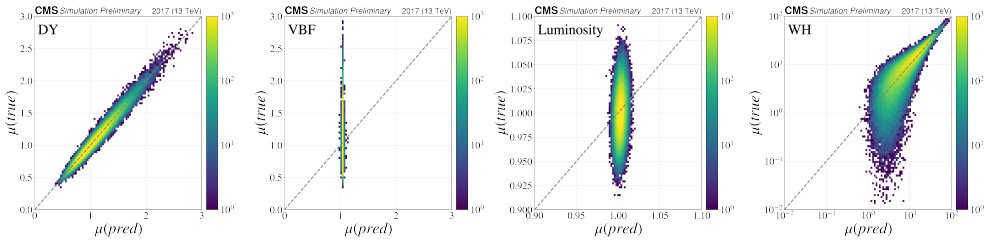

Figure 5. The network predictions and the the true MC values for three selected nuisance parameters and one selected signal strength modifier parameter.

- *Well-reconstructed parameters*: the posteriors of these parameters are narrow and centered around the true MC value. For this reason, the uncertainty on these parameters is small as well.

- *Weakly-reconstructed parameters*: for these parameters, the network has already achieved some sensitivity to the parameters. However, the posteriors are broader than those of the well-reconstructed parameters and network shows less confidence in their reconstruction.

- *Unrecognized parameters*: for these parameters, the network does not recognize the effect of the parameters on the conditions and predicts the average of the prior distribution exclusively. If the physics parameter x is weakly statistically dependent on c, then the priors $p(x)$ and the posteriors $p(x|c)$ coincide:

$$p(x|c) \approx p(x) \tag{5}$$

For the dominant background processes such as DY, the nuisance parameters are well-reconstructed and the network has achieved the highest sensitivity to them. In the 2D histograms of fig. 5, these prediction of the parameters scatter closely around the dashed gray perfect prediction line and the histogram is symmetric to this line. In fig. 6, the narrow posterior distribution can be clearly seen, which greatly differs from the orange prior distribution. For the weakly-reconstructed parameters, such as the luminosity nuisance parameter, the network develops some sensitivity from the mean of the prior distribution but fails to predict the true values over the complete prior region. For these parameters, a close similarity between the priors and posteriors can be observed as it can be seen in fig. 6. For the unrecognized parameters such as VBF, a narrow distribution in fig. 5 can be observed as the model returns the mean of the prior as prediction. As a consequence, the posterior and prior distribution in fig. 6 coincide.

For the signal processes, clear deviations from the true values in the low-signal $\mu \lesssim 10$ region was observed. Since the latent space distributions and calibration curves show no signs of biases in the trained network model, this effect can be explained by the network losing sensitivity in this region; hence the predictions are mapped towards the mean of the prior, resulting in the tail in fig. 5. This effect can be most effectively reduced by improving the analysis sensitivity further. For the higher-than-expected signal region, the network is able to reconstruct these parameters with good confidence. In fig. 6, the similarity between the posteriors and priors can still be observed; the width of the predicted posterior distribution speaks of a low-confidence reconstruction quality.

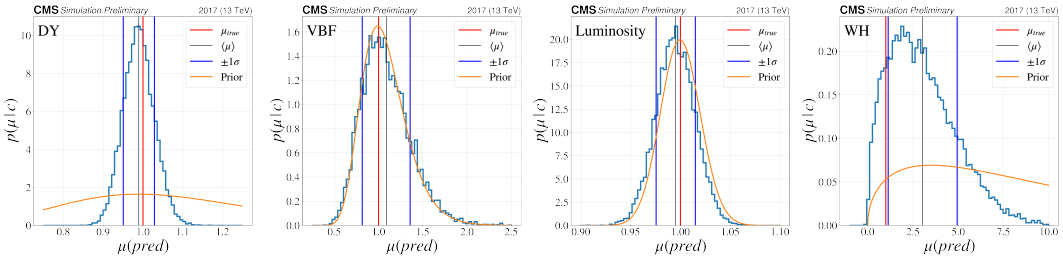

Figure 6. The obtained posterior distributions for the standard model expectation (blue). The priors are shown in orange. The location of the MC-truth and that of the predictions are shown in red and black, respectively. The blue horizontal bars represent the 68% quantile edges of the obtained posteriors.

5 Conclusion

In this work, the capability of the cINNs in the reconstruction of signal strength modifier parameters has been shown and the performance of the trained network has been described. cINNs are versatile networks which encode a continuous and continuously differentiable model, which makes them applicable in differentiable analysis workflows. These networks can infer these parameters several orders of magnitude faster than many-parameter likelihood fits and are able to characterize the reconstruction quality reliably. The trained model's output matches the expected latent distribution's shape, shows no strong biases and is neither under- or over-confident. The trained network is able to reconstruct the value of μ in regions where the analysis sensitivity is sufficient. The resolution of the network can be characterized by the widths of the posteriors: the better the resolution the narrower the posterior.

References

[1] L. Ardizzone, C. Lüth, J. Kruse, C. Rother, U. Köthe, CoRR (2019), 1907.02392

[2] S.T. Radev, U.K. Mertens, A. Voss, L. Ardizzone, U. Köthe, IEEE Transactions on Neural Networks and Learning Systems **33**, 1452 (2022)

[3] V.F. Ksoll, L. Ardizzone, R. Klessen, U. Koethe, E. Sabbi, M. Robberto, D. Gouliermis, C. Rother, P. Zeidler, M. Gennaro, Monthly Notices of the Royal Astronomical Society **499**, 5447 (2020)

[4] T. Bister, M. Erdmann, U. Köthe, J. Schulte, The European Physical Journal C **82** (2022)

[5] M. Bellagente, A. Butter, G. Kasieczka, T. Plehn, A. Rousselot, R. Winterhalder, L. Ardizzone, U. Köthe, SciPost Phys. **9**, 074 (2020)

[6] S. Gieseke et al. (2011), 1102.1672

[7] T. Sjöstrand, S. Mrenna, P. Skands, Computer Physics Communications **178**, 852 (2008)

[8] D.P. Kingma, P. Dhariwal, *Glow: Generative flow with invertible 1x1 convolutions* (2018)

[9] M. Baak, S. Gadatsch, R. Harrington, W. Verkerke, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **771**, 39 (2015)

[10] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga et al., in *Advances in Neural Information Processing Systems 32* (Curran Associates, Inc., 2019), pp. 8024–8035

[11] D.P. Kingma, J. Ba (2017), 1412.6980