THE EUROPEAN
PHYSICAL JOURNAL C

Testable likelihoods for beyond-the-standard model fits

Anja Beck[1,a] , Méril Reboud[2,b] , Danny van Dyk[2,c]

[1] Department of Physics, University of Warwick, Coventry CV4 7AL, UK
[2] Institute for Particle Physics Phenomenology and Department of Physics, Durham University, Durham DH1 3LE, UK

Abstract Studying potential BSM effects at the precision frontier requires accurate transfer of information from low-energy measurements to high-energy BSM models. We propose to use normalising flows to construct likelihood functions that achieve this transfer. Likelihood functions constructed in this way provide the means to generate additional samples and admit a "trivial" goodness-of-fit test in form of a χ^2 test statistic. Here, we study a particular form of normalising flow, apply it to a multi-modal and non-Gaussian example, and quantify the accuracy of the likelihood function and its test statistic.

1 Introduction

Contemporary experimental analyses at the Large Hadron Collider have, so far, not been able to discover particles beyond the Standard Model (BSM) at energy scales below $\simeq 1 \text{TeV}$. As a consequence, model building has increasingly turned toward using effective field theories (EFT) to describe any potential BSM effects below these scales. The Standard Model Effective Field Theory (SMEFT) [1] is one of the main choices, as is the Higgs effective field theory (HEFT) [2]. Constraining the EFT parameters is a challenge, due the large dimensionality of the parameter space. Taking the SMEFT as an example, the basis of operators in the leading mass-dimension six Lagrangian amounts to 2499 independent EFT Wilson coefficients [3]. The observed hierarchies among masses and mixing of quark flavours at low-energies has inspired systematic approaches like minimal flavour violation to reduce the number of free parameters [4].

Low-energy phenomena beside quark mixing have long been used to further our understanding of BSM physics at large scales; flavour-changing processes and anomalous elec-

tric dipole moments are excellent examples therof, which contribute substantial statistical power to constrain BSM effects [5]. Flavour-changing processes in particular are commonly interpreted in a "model-independent" fashion by inferring their relevant parameters within another EFT, the Weak Effective Theory (WET) [6–8]. However, including low-energy constraints in this way provides for a challenge: the interpretation of a majority of the low-energy constraints relies on our understanding of hadronic physics in some capacity. This leads to a proliferation of hadronic nuisance parameters, which renders a global interpretation of all WET parameters impractical if not practically impossible. Examples for this proliferation are plentiful in b-quark decay and include analyses of exclusive $b \to s\ell^+\ell^-$ processes with 113 hadronic nuisance parameters and only 2 parameters of interest [9] as well as analyses of exclusive $b \to u\ell^-\bar{\nu}$ processes with 50 hadronic nuisance parameters and only 5 parameters of interest [10]. To overcome this problem, the following strategy has been devised [11]:

- divide the free parameters of the WET into so-called "sectors", which are mutually independent to leading power in G_F;
- identify sets of observables that constrain a single sector of the WET and infer that sectors parameters;
- repeat for as many sectors as possible.

In this way, statistical constraints on the parameters in the individual WET sectors in form of posterior densities can be obtained, providing a *local picture* of the low-energy effects of BSM physics.

To gain a global picture of the BSM landscape, either in terms of the parameters of the SMEFT, the HEFT, or a UV-complete BSM model, the previously obtained local constraints for individual WET sectors can, in principle, be used as a likelihood function. On the physics side, this requires matching between the WET and the genuine parameters of

[a] e-mail: anja.beck@cern.ch (corresponding author)

[b] e-mail: merilreboud@gmail.com

[c] e-mail: danny.van.dyk@gmail.com

interest, i.e, the parameters of a UV-complete model, the SMEFT, or the HEFT. In the case of the SMEFT, a complete matching of all dimension-6 operators to the full dimension-6 wet has been achieved at the one-loop level [12].

To date, the strategy outlined above has not yet been implemented: no library of statistical WET constraints is available. The key obstacles in the implementation are not specific to the underlying physics. Instead, they are the accurate transfer of the statistical results for even a single sector; the potential to test a likelihood's goodness of fit through a suitable test statistic; and the ease of use by providing a reference code that exemplifies the approach.[1]

Here, we present a small piece to solving the overall puzzle of how to efficiently and accurately include the low-energy likelihoods in BSM fits: the construction of a likelihood function that encodes the WET constraints and that admits a test statistic. Our approach uses methods from the field of automatized learning and generational models. We illustrate the approach at the hand of a concrete example likelihood, for which we use posterior samples for the WET parameters obtained in the course of a previous analysis [10]; cf. also a direct determination of SMEFT parameters in a similar setup [14]. Our example likelihood is multi-modal, and the shape of each mode is distinctly non-Gaussian. This renders the objective of providing a testable likelihood function in terms of the BSM parameters quite challenging.

The possible applications for obtaining such a likelihood function are two-fold: first, it permits to generate additional samples that are (ideally) identically distributed as the training samples. Second, it provides a test statistic to use in a subsequent EFT fit. We illustrate how normalising flows make both applications possible, by translating our example posterior "target" density to a unimodal multivariate Gaussian "base" density. We propose a number of tests to check the quality of this translation, gauging the validity of both applications.

2 Preliminaries

2.1 Notation and objective

We begin by introducing our notation for the physics parameters and their associated statistical quantities. Let

$P_T^*(\vec{\vartheta} \mid \text{WET})$ be the "true" posterior density for the low-energy WET parameters $\vec{\vartheta} \in T \equiv \mathbb{R}^D$, also known as the WET Wilson coefficients. We refer to the vector space T as the *target* space of dimension D. In our later example, we will consider T to be the space of WET Wilson coefficients in the $ub\ell\nu$ sector as studied in Ref. [10]. We assume that we can access P_T^* numerically by means of Monte Carlo importance samples. Our intent is to construct a likelihood

$$L(\vec{\varphi}) \equiv P_T(\vec{\vartheta}(\vec{\varphi}) \mid \text{WET}) \tag{1}$$

where P_T is a model of P_T^* and $\vec{\varphi} \in F$ represents some set of BSM physics parameters within some vector space F. In an envisaged SMEFT application, F would represent the vector space of SMEFT Wilson coefficients. We stress at this point that, although $P_T^{(*)}$ is a probability density, L is in general not a probability density and is hence labelled a *likelihood* function.

Next, we introduce a bijective mapping f between our target space T and some *base* space B

$$B \ni \vec{\beta} = f(\vec{\vartheta}). \tag{2}$$

Our objective is to find a mapping f and its associated Jacobian $J_f \equiv \partial f / \partial \vec{\vartheta}$

$$\vec{\beta} \sim P_B(\vec{\beta}) = J_f(P_T \circ f^{-1})(\vec{\beta}) \tag{3}$$

that ensures that $P_B(\vec{\beta})$ is multivariate standard Gaussian density: $P_B(\vec{\beta}) = \mathcal{N}^{(D)}(\vec{\beta} \mid \vec{\mu} = \vec{0}, \Sigma = \mathbb{1})$. As a consequence, one finds immediately that the two-norm in base space follows a χ^2 distribution:

$$||\vec{\beta}||_2 \equiv ||f(\vec{\vartheta})||_2 \sim \chi^2(\nu = D). \tag{4}$$

The likelihood function is thus fully defined by the mapping f. A natural test statistic is provided by the χ^2 statistic for the squared 2-norm in base space.

2.2 A normalising flow for a real-valued non-volume preserving model

The framework of normalising flows [15] has been developed with the explicit intent to transform an existing probability density to a (standard)normal density, i.e., to "normalise" the density. To this end, f is constructed as a composition of K individual bijective *mapping layers* $f^{(k)}$,

$$f = f^{(1)} \circ f^{(2)} \circ \cdots \circ f^{(K)}. \tag{5}$$

Here, we restrict ourselves to a particular class of mapping layer, following the real non-volume preserving (RealNVP) model [16]: the affine coupling layers. This class of layers are

---

[1] An alternative to this strategy is to use the most constraining observables for each WET sector as low-energy likelihoods and use the latter directly within a global (SMEFT) likelihood. Despite the large number of nuisance parameters, this approach has been shown to be very useful and has been implemented as part of the `smelli` software [13]. However, it currently comes with the drawback of repeating the low-energy analyses for every point in the global (here: SMEFT) parameter space, leading to a waste of computational resources compared to the strategy proposed above. Improvements to `smelli` overcoming this issue are presently under development.

chosen here for their proven ability to map multi-modal densities of real-valued parameters to a standard normal through non-volume-preserving transformations [16]. Each of these layers has the structure

$$f^{(k)} : \vec{x} \in \mathbb{R}^D \mapsto f^{(k)}(\vec{x}) \in \mathbb{R}^D \tag{6}$$

$$f^{(k)}(\vec{x}) = (p \circ a^{(k)})(\vec{x}). \tag{7}$$

In the above, $\vec{x} = (x_1, x_2, \ldots, x_D)^T$, and p is a permutation operation

$$p(\vec{x}) = (x_2, \ldots, x_D, x_1)^T \tag{8}$$

with a trivial Jacobian $|J_p| = 1$. The power of the RealNVP model lays in the use of *affine coupling layers* $a^{(k)}$ [16]. Each of these layers splits its input vector into two parts $\vec{x} = (\vec{x}_{\text{lo}}^T, \vec{x}_{\text{hi}}^T)^T$, where $d_{\text{lo}} \equiv \dim \vec{x}_{\text{lo}} = \lceil D/2 \rceil$ and $d_{\text{hi}} \equiv \dim \vec{x}_{\text{hi}} = \lfloor D/2 \rfloor$. The affine coupling layers then map their inputs to

$$a^{(k)}(\vec{x}) = \begin{pmatrix} \vec{x}_{\text{lo}} \\ \vec{x}_{\text{hi}} \odot \exp(s^{(k)}(\vec{x}_{\text{lo}})) \oplus t^{(k)}(\vec{x}_{\text{lo}}) \end{pmatrix}, \tag{9}$$

where $s^{(k)}$ and $t^{(k)}$ are real-valued functions $\mathbb{R}^{d_{\text{lo}}} \mapsto \mathbb{R}^{d_{\text{hi}}}$, and $\odot$ and $\oplus$ indicates element-wise multiplication and addition. Since both $s^{(k)}$ and $t^{(k)}$ are independent of $\vec{x}_{\text{hi}}$, the Jacobian for each affine coupling layer $a^{(k)}$ takes the simple form:

$$J_{a^{(k)}} \equiv \exp\left[\sum_i s_i^{(k)}(\vec{x}_{\text{lo}}) \right]. \tag{10}$$

There are no further restrictions on the functions $s^{(k)}$ and $t^{(k)}$ and they are commonly learned from data using a neural network [17].

3 Physics case: semileptonic *B* decays

To illustrate the viability of our approach, we carry out a proof-of-concept (POC) study using existing posterior samples [18] from a previous BSM analysis [10]. This analysis investigates the BSM reach in exclusive semileptonic $b \to u\ell\bar{\nu}$ processes within the framework of the $ub\ell\nu$ sector of the WET:

$$\mathcal{H}^{ub\ell\nu} = -\frac{4G_F}{\sqrt{2}} \tilde{V}_{ub} \sum_i \mathcal{C}_i^\ell \mathcal{O}_i^\ell + \text{h.c.}. \tag{11}$$

Here $\mathcal{C}_i^\ell$ represents a WET Wilson coefficient (i.e., a free parameter in the fit to data) while $\mathcal{O}_i^\ell$ represents a local dimension-six effective field operator (i.e., the source of the 50 hadronic nuisance parameters in the original analysis [10]). The assumptions inherent to that analysis lead to a basis composed of five WET operators:

$$\begin{aligned} \mathcal{O}_{V,L}^\ell &= [\bar{u}\gamma^\mu P_L b][\bar{\ell}\gamma_\mu P_L \nu], & \mathcal{O}_{V,R}^\ell &= [\bar{u}\gamma^\mu P_R b][\bar{\ell}\gamma_\mu P_L \nu], \\ \mathcal{O}_{S,L}^\ell &= [\bar{u} P_L b][\bar{\ell} P_L \nu], & \mathcal{O}_{S,R}^\ell &= [\bar{u} P_R b][\bar{\ell} P_L \nu], \\ \mathcal{O}_T^\ell &= [\bar{u}\sigma^{\mu\nu} b][\bar{\ell}\sigma_{\mu\nu} P_L \nu]. \end{aligned} \tag{12}$$

For the purpose of a BSM study, only the marginal posterior of the WET parameters is of interest. One readily obtains this marginal posterior by discarding the sample columns corresponding to any (hadronic) nuisance parameters. A corner plot of this marginal posterior originally published in Ref. [10] is shown in Fig. 1. As illustrated, the 5D marginal posterior is very obviously non-Gaussian with a substantial number of isolated modes. The samples obtained in this analysis are overlaid with a kernel-density estimate (KDE) of the marginalised posterior. Although estimating the density is one of our main goals, the use of a KDE is neither sufficient nor advisable for the purpose we pursue. First, KDEs are notorious for their computational costs. Second, KDEs are sensitive to misestimation of densities close to the boundaries of the parameter space. Last but not least, KDEs do not provide a test statistic.

The objective of the POC study is now to determine whether a generative and testable likelihood can be constructed from the available posterior samples by employing normalising flows. To achieve this objective, we investigate a two-dimensional subset of our posterior samples before we progress to the full five-dimensional problem. For both cases, we use normalising flows f of the form shown in Eq. (5), i.e., a sequence of K individual mapping layers Eq. (6), each consisting of a composition of an affine coupling layer [16] and a permutation. This class of mapping layer is implemented as part of the `normflows` software package [17], which we use to carry out our analysis. The `normflows` software is based on `pytorch` [19], which is used in the training of the underlying neutral networks.

We train the normalising flow on a total of 144k posterior samples using a total of $K = 32$ mapping layers. Each mapping layer is associated with a 4-layer perceptron, consisting of the input layer with $\lceil D/2 \rceil$ nodes, two hidden layers with 64 nodes each, and the output layer with $2 \times \lfloor D/2 \rfloor$ nodes. The outputs correspond to the vector-valued functions $s^{(k)}$ and $t^{(k)}$ as used in Eq. (6). We use the default choice of loss function provided by `normflows`, the "forward" Kullback–Leibler divergence [20]

$$\begin{aligned} \text{KL}(\vec{\alpha}) &= -\mathbb{E}_{P_T^*(\vec{\vartheta})}\left[\log P_T(\vec{\vartheta} \mid \vec{\alpha}) \right] + \text{const} \\ &= -\mathbb{E}_{P_T^*(\vec{\vartheta})}\left[\log P_B(f(\vec{\vartheta} \mid \vec{\alpha}) + \log \left| \det J_f(\vec{\vartheta} \mid \vec{\alpha}) \right| \right] \\ &\quad + \text{const}. \end{aligned} \tag{13}$$
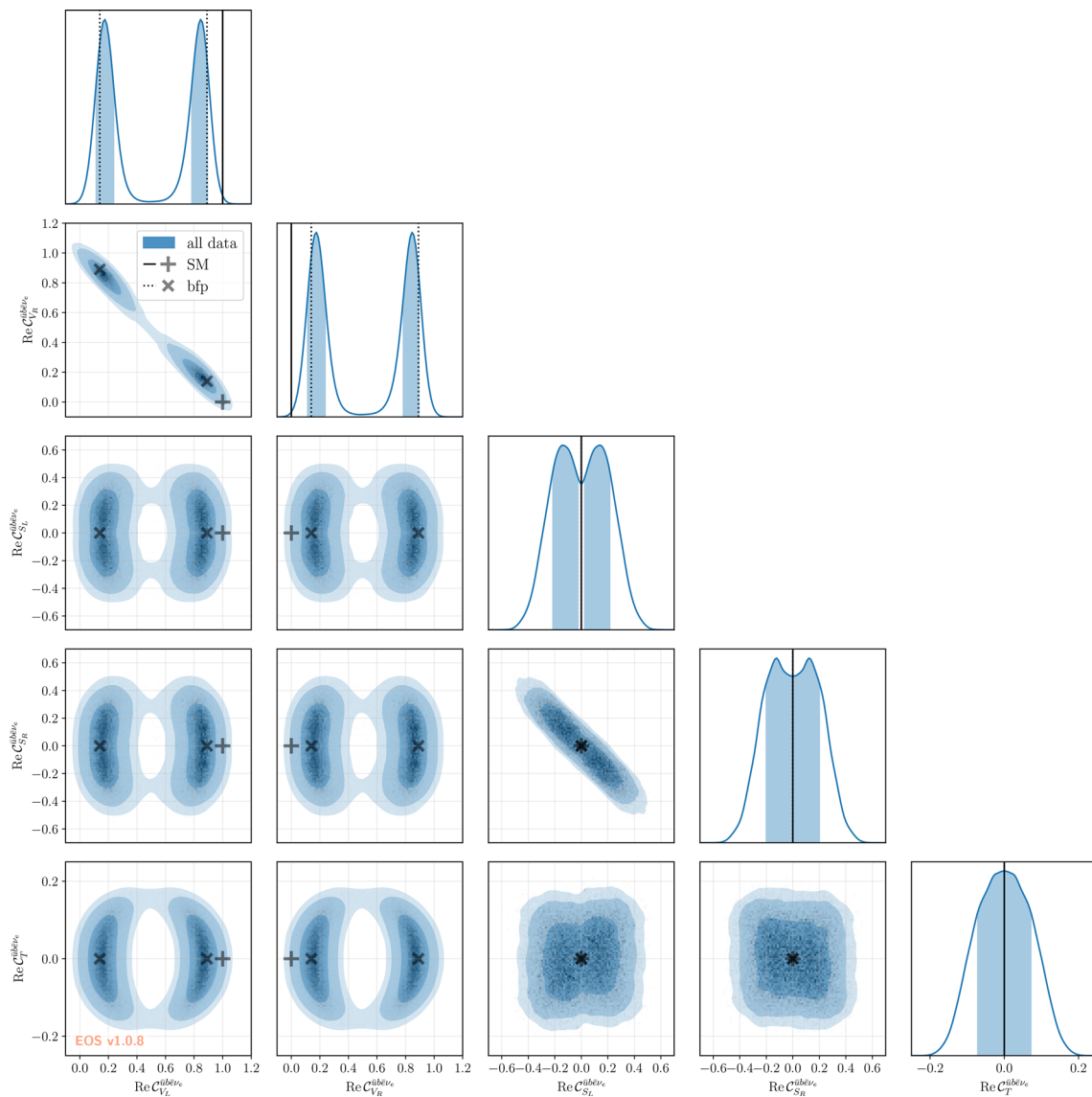
Fig. 1 Marginalized 5D posterior for the parameters of interest $\vec{\vartheta}$ in the $ub\ell\nu$ example. The black dots indicate the distribution of the posterior samples in form of a scatter plot. The blue-tinted areas correspond to the 68, 95, and 99% probability regions as obtained from a smooth histogram based on kernel-density estimation

The loss function is minimised with respect to the neural network parameters $\vec{\alpha}$ using the "Adam" algorithm [21] as implemented in `pytorch`. Although we use a total of $10k$ optimisation iterations, we find that a close-to-optimal solution is found around the $2k$ iterations mark in both cases. We show the evolution of the loss functions in terms of the number of optimisation steps in Fig. 2.

3.1 Application to a two-dimensional subset of samples

In a first step, we perform a POC study at the hand of a subset of WET Wilson coefficients. The target space only consists of the two parameters:

$$\vec{\vartheta} = (\vartheta_1, \vartheta_2)^T \equiv (\text{Re } \mathcal{C}_{V_L}^\ell, \text{Re } \mathcal{C}_{S_L}^\ell)^T . \tag{14}$$

Our choice of 2D example exhibits two difficult features in target space as illustrated in Fig. 1. First, the 2D marginal posterior is multi-modal. Second, each mode is distinctly non-Gaussian and its isoprobability contours resemble a bean-like shape. In the following discussion, we will frequently refer to these shapes simply as "beans". The central objective of this 2D POC study is to train a normalising flow f such that the base space variables

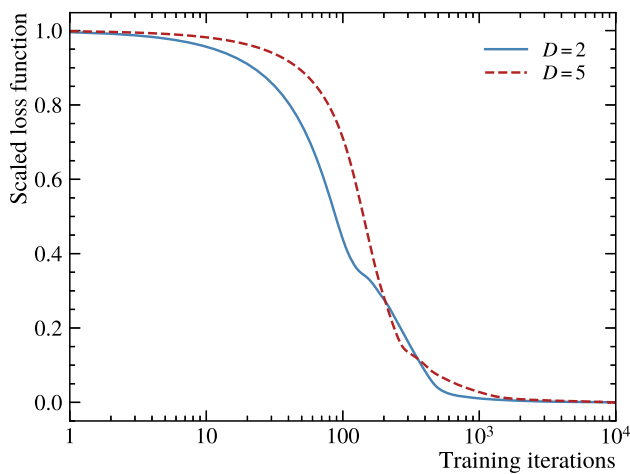$$\vec{\beta} \equiv f(\vec{\vartheta}) . \tag{15}$$

Fig. 2 Evolution of the loss function for the two- and five-dimensional case as the blue solid and red dashed lines, respectively. The loss values have been scaled so that all values are confined to the interval [0, 1]

have a bivariate standard normal distribution: $\vec{\beta} \sim \mathcal{N}_2(\vec{0}, \mathbb{1})$.

3.1.1 Distribution of the parameters in target and base space

In Fig. 3 we show the result of the training in the target and base spaces. The top plots contain the model in the target space after training and the base distribution used as starting point for the transformation. The bottom plots show the distribution of the posterior samples in the target space, used in the training, and after transformation to the base space. Visually, the trained model in the target space (top left) captures the main features of the training sample (bottom left). Nevertheless, the model contains a faint filament connecting the left and right beans (top left), which is not present in the true posterior samples (bottom left). Although this additional structure becomes less prominent as the value of the loss function decreases, it never fully disappears using our analysis setup. This additional feature translates to a less populated line across the sample distribution in the base space (bottom right). This line however becomes invisible after a sufficient number of training iterations as is shown in Fig. 4, where we juxtapose two transformation of the samples of P_T^* from T space to B space using the normalising flows trained with 2000 and trained with 10,000 iterations. We conclude that the presence of the filament requires careful diagnosis if the RealNVP model can be used to achieve our stated objective: to obtain a generative and testable likelihood function.

3.1.2 Comparison of the modelled and true distributions

We investigate the quality of our trained normalising flows with a comparison of the modelled and true distributions. The comparison is first performed using histograms of the

distributions in both the base and the target space. The value n_i of the true distribution in bin x_i is obtained by dividing the number of samples in the bin by the total number of samples N. The value of the modelled distribution is represented by the value of the modelled density at the bin centre, $P_X(x_i)$, multiplied with the area of the bin, $A_i \equiv A(x_i)$. Our measure for the comparison is the deviation defined as

$$\text{deviation} = \frac{n_i - P_X(x_i)A_i}{\sigma_i}, \tag{16}$$

where the uncertainty on the true distribution in bin x_i is assumed to be

$$\sigma_i = \frac{\sqrt{\max(1, Nn_i)}}{N} \tag{17}$$

The deviation is shown in Fig. 5. As expected, the deviation is close to zero in highly populated regions and the connecting filament between the beans in the target space is clearly visible. The larger deviations in sparsely populated regions are a result of the limited size of the data sample, where the red bins contain a single data point and the blue regions are empty bins.

3.1.3 Sample classification with a BDT

The next step is to quantify the quality of the normalising flows in regard to the two primary purposes, generating samples from the posterior and determining the test statistic for the actual distribution. To this end, we train classifiers to distinguish samples generated from the true and modelled posterior distributions on the sample position in the base space, the position in target space, and the density in target space together with $\|\vec{\beta}\|_2$. In all three cases, the classification is carried out with a Boosted Decision Tree (BDT) [22] with the default settings of the `XGBClassifier` class provided by the python library `xgboost` [23].

The data generated from the true posterior is split into two equal parts, which are then only used for training and testing respectively. A sample of the same statistical power is generated from the modelled distribution and also split into two equal parts. The receiver-operator statistic is the dependence of the true positive rate on the false positive rate on the testing sample. The area under its curve (AUC), for a set of features x used in the classification, is a measure for the quality of the classification. Perfect classification results in $\text{AUC}(x) = 1$, while random classification results in $\text{AUC}(x) = 0.5$.

Fig. 3 Modelled (top) and empirical (bottom) distributions in the target (left) and base (right) space. Top: model in the target space after training (left) and base space (right). Bottom: distributions of the samples of the true posterior used for training the normalising flows in the target space (left) and after transformation to the base space using the trained flow (right). The circles in the base space serve only to guide the eye and show the contours $||\vec{\beta}||_2 = 1$ and $||\vec{\beta}||_2 = 9$
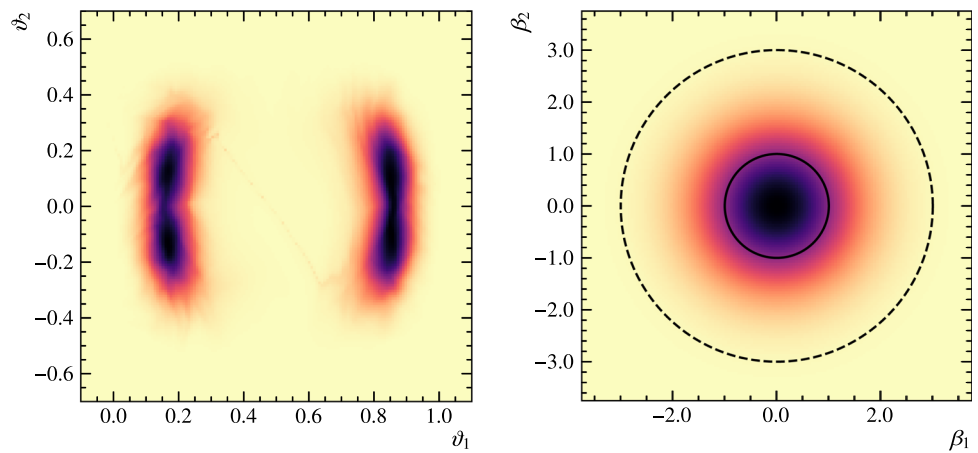
Fig. 4 Data distributions in the base space after 2000 (left) and 10,000 (right) training iterations. The circles in the base space serve only to guide the eye and show the contours $||\vec{\beta}||_2 = 1$ and $||\vec{\beta}||_2 = 9$
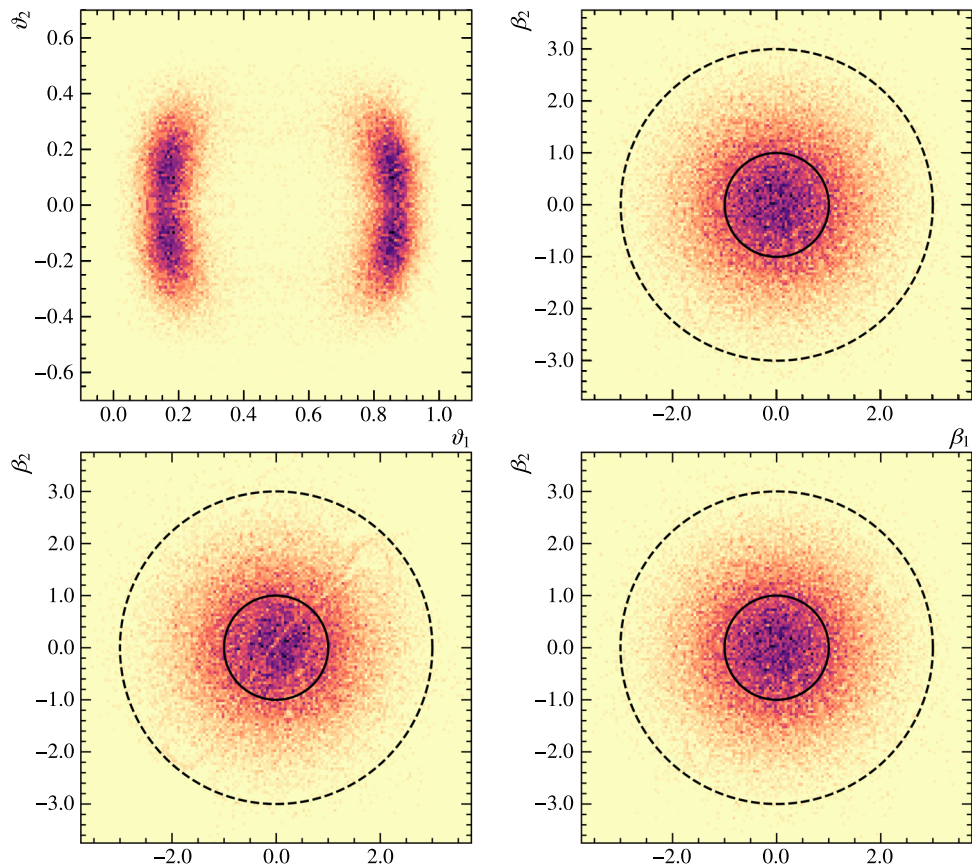
Fig. 5 Selected diagnostic plots for the proof of concept with the 2D subset of samples in the target (left) and base (right) space. Deviation of the model from the sample distribution in units of the uncertainty due to the finite size of the data sample. The circles in the base space serve only to guide the eye and show the contours $||\vec{\beta}||_2 = 1$ and $||\vec{\beta}||_2 = 9$
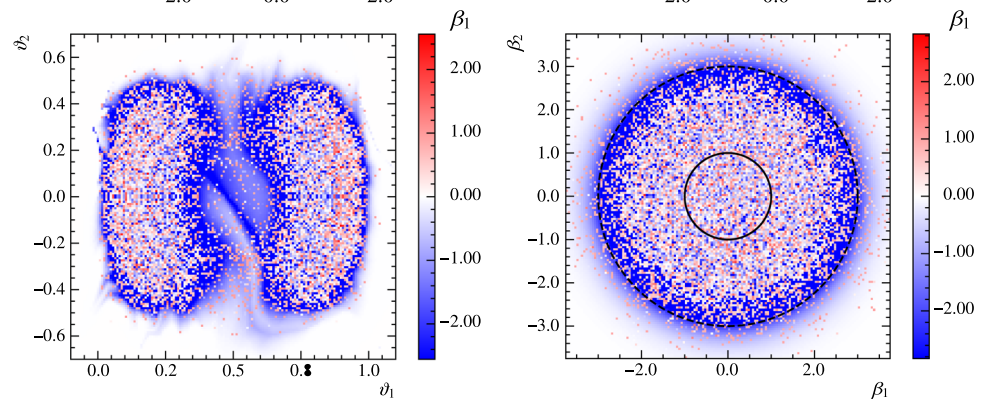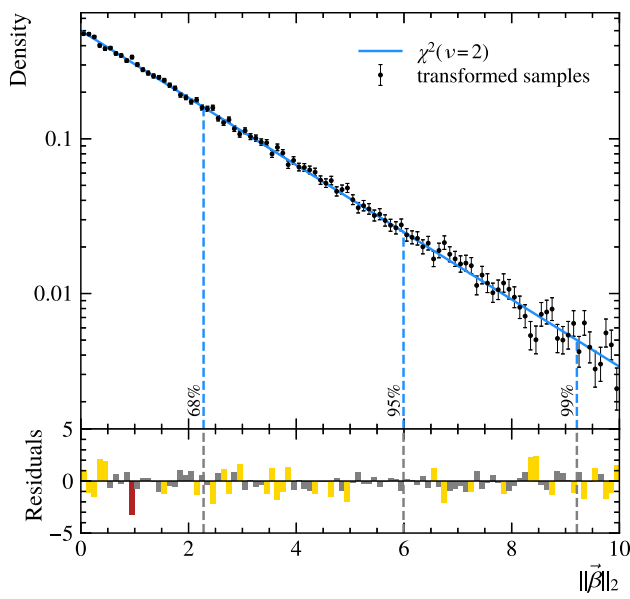
Fig. 6 Empirical distribution of $||\vec{\beta}||_2$ for the $ub\ell\nu$ example using only two dimensions, see Eq. (15). The colour of the bars representing the residuals, r, correspond to the magnitude, where grey, yellow, and red represent a residual with absolute value $|r| < 1$, $1 < |r| < 3$, and $3 < |r|$ respectively. The dashed vertical lines indicate the positions where the cumulative distribution function of $\chi^2(\nu = 2)$ equals 68, 95, and 99%

3.1.4 Differences between the modelled and the true posterior density

Training and testing the BDT on the position in target and base space, results in

$$\text{AUC}(\vec{\vartheta}) = 0.577 \pm 0.002 \quad \text{and} \quad \text{AUC}(\vec{\beta}) = 0.549 \pm 0.003, \tag{18}$$

respectively. These low AUC values indicate that the true and modelled density are very similar. However, they are significantly larger than 0.5 showing that the BDT can find exploitable differences. The smaller AUC score in the base space compared to the target space is in line with the observation of a small connecting artefact in target space and the resulting diagonal gap in the base space, as discussed in relation to Figs. 4 and 5. As a sanity check, the BDT is also trained on two samples which are both generated from the model resulting in AUC values compatible with 0.5 as expected for two indistinguishable samples.

3.1.5 Testing the modelled test statistic

Finally, Fig. 6 shows the distribution of the 2-norm of the data samples in the base space. They appear to follow a two-dimensional χ^2 distribution as anticipated in Eq. (4).

We determine the reliability of the test statistic by investigating the ability of a BDT to distinguish the true and modelled posterior densities. This BDT is trained on two variables: the posterior density in the target space and the 2-norm in the base space, $||\vec{\beta}||_2$, which resembles a χ^2 distribution.

The posterior density is not available as a function. We are therefore forced to use an empirical estimation. Because the value of the density relies on the data itself, the AUC value for classification of two samples generated from the same model is larger than 0.5 and depends on the estimation method used; we investigate KDEs with different kernels as well as nearest neighbours density estimation.

The AUC scores for classification of samples generated from the true and modelled posterior density are always larger than the corresponding value for classification of two model-based samples. The AUC value, obtained from estimating the density from the number of points with a distance below 0.015 around a given point, is AUC(density) $= 0.614 \pm 0.006$ which is an increase of 0.039 with respect to the baseline value calculated when classifying two model-based samples.

3.1.6 Conclusion of the two-dimensional study

Our conclusion for this 2D POC study is that our approach is a viable one and provides both an evaluatable likelihood and a reliable test statistic. Publishing 2D WET posteriors in form of a normalising flow therefore clearly beats current best practices such as Gaussian approximations or Gaussian mixtures models (as for example used in Ref. [10]) in terms of both usefulness and reliability. We find that the presence of the filament does not adversely affect achieving our stated objectives at the current level of precision. We leave an investigation as to other choices of mapping layers, in particular autoregressive rational quadratic splines as used in Ref. [24], to future work. Instead, we move on to apply our approach to the full five-dimensional POC.

3.2 Application to the full five-dimensional samples

We carry out the same analysis on the full five-dimensional posterior of Fig. 1. This is a much more complicated case in terms of possible artefacts between beans. Since this POC study is only illustrative, we do not try to optimise the training of the normalising flows and keep the same setup as for the two-dimensional case. As visual comparisons of the distributions suffer from loss of information due to the projection from five to two dimension, we rely on the BDT tests that we introduced for the two-dimensional case.

Figure 7 shows the distribution of the 2-norm of the samples transformed to the five-dimensional base space. The distribution is overall consistent with a χ^2 distribution with five dimensions. the residuals reveal small systematic differences between the shape of the distribution of the data points and
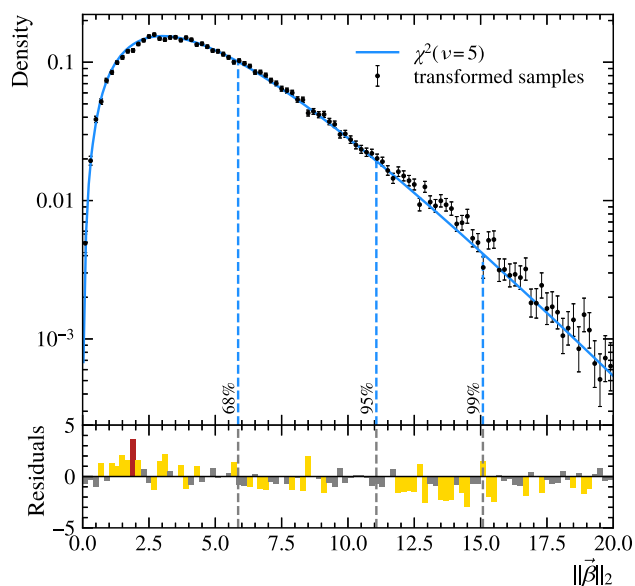
Fig. 7 Empirical distribution of $||\vec{\beta}||_2$ for the $ub\ell\nu$ example using all five dimensions. The colour of the bars representing the residuals, r, correspond to the magnitude, where grey, yellow, and red represent a residual with absolute value $|r| < 1$, $1 < |r| < 3$, and $3 < |r|$ respectively. The dashed vertical lines indicate the positions where the cumulative distribution function of $\chi^2(\nu = 5)$ equals 68, 95, and 99%

the expected PDF. In particular, the residuals tend to aggregate positive values at small $\chi^2 < 5$ and negative values at large $10 < \chi^2 < 20$. This conclusions are confirmed by the BDT classification in the base space which reaches an AUC score of

$$\mathrm{AUC}(\vec{\beta}) = 0.5708 \pm 0.0005 .\tag{19}$$

This value is slightly larger than the value obtained in Sect. 3.1 but corroborates that the data sampled from the true posterior distribution resembles a five-dimensional Gaussian distribution after transformation with the trained normalising flow.

On the other hand, the closeness of the true and modelled distributions in the target space is much less convincing. The BDT classification in the target space reaches AUC values close to one. Similarly, training and testing the BDT on the value of the posterior density in target space and the 2-norm in base space gives $\mathrm{AUC}(\text{density}) = 0.9150 \pm 0.0005$ which confirms a poor modelling of the true distribution.

We conclude that the full five-dimensional variable space presents a more complex challenge compared to the two-dimensional POC. The basic normalising flow training as pursued in this work does not provide a satisfactory model for the multi-modal distribution in target space. This is not surprising since we use the same setup for the two- and five-dimensional tests. The small number of iterations needed to train the five-dimensional normalising flow, visible in Fig. 2,

indicates that more complicated models can be trained to achieve better performances. The tuning of the model and the training procedure are however beyond the scope of this paper.

4 Summary

We show that normalising flows can overcome two major hurdles to the full exploitation of statistical constraints on EFT parameters stemming from phenomenological analyses of low-energy processes. On the one hand, they enable sampling from a previous analysis' posterior distribution by converting it into a simple multivariate Gaussian density. On the other hand, they provide a simple test statistics for the distribution.

We investigate this procedure by training a RealNVP normalising flow to a physics case that exhibits multi-modal posterior densities in two and five dimensions. The compatibility of the modelled and the empirical posterior distributions are examined using different statistical tools.

Our conclusions are as follows. When facing multi-modal distributions, the modelling of the true posterior densities leads to the presence of artefacts as observed here. Nevertheless, these artefacts do not prevent the successful usage of normalising flows as proposed here. Instead, they require careful (supervised) learning of the features of the true posterior density. In the two-dimensional case, we have illustrated that both of our objectives, the sampling and the presence of a test statistics, are achieved. However, applying our approach to the five-dimensional case without modification leads to less satisfactory performances. Improvements to performance through detailed modelling, at the cost of more computationally-intensive training, are left for future work.

We are currently modifying the EOS software [25] to provide users with an interface that transparently makes use of our proposed method of constructing low-energy likelihoods for existing and future analyses of flavour-changing processes.

References

1. W. Buchmuller, D. Wyler, Effective Lagrangian analysis of new interactions and flavor conservation. Nucl. Phys. B **268**, 621–653 (1986). https://doi.org/10.1016/0550-3213(86)90262-2

2. F. Feruglio, The Chiral approach to the electroweak interactions. Int. J. Mod. Phys. A **8**, 4937–4972 (1993). https://doi.org/10.1142/S0217751X93001946. arXiv:hep-ph/9301281

3. B. Grzadkowski, M. Iskrzynski, M. Misiak, J. Rosiek, Dimension-six terms in the standard model Lagrangian. JHEP **10**, 085 (2010). https://doi.org/10.1007/JHEP10(2010)085. arXiv:1008.4884 [hep-ph]

4. G. D'Ambrosio, G.F. Giudice, G. Isidori, A. Strumia, Minimal flavor violation: an effective field theory approach. Nucl. Phys. B **645**, 155–187 (2002). https://doi.org/10.1016/S0550-3213(02)00836-2. arXiv:hep-ph/0207036

5. M. Artuso et al., Report of the frontier for rare processes and precision measurements. (2022). arXiv:2210.04765 [hep-ex]

6. J. Aebischer, M. Fael, C. Greub, J. Virto, B physics beyond the standard model at one loop: complete renormalization group evolution below the electroweak scale. JHEP **09**, 158 (2017). https://doi.org/10.1007/JHEP09(2017)158. arXiv:1704.06639 [hep-ph]

7. E.E. Jenkins, A.V. Manohar, P. Stoffer, Low-energy effective field theory below the electroweak scale: operators and matching. JHEP **03**, 016 (2018). https://doi.org/10.1007/JHEP03(2018)016. arXiv:1709.04486 [hep-ph]

8. E.E. Jenkins, A.V. Manohar, P. Stoffer, Low-energy effective field theory below the electroweak scale: anomalous dimensions. JHEP **01**, 084 (2018). https://doi.org/10.1007/JHEP01(2018)084. arXiv:1711.05270 [hep-ph]

9. N. Gubernari, M. Reboud, D. van Dyk, J. Virto, Improved theory predictions and global analysis of exclusive $b \to s\mu^+\mu^-$ processes. JHEP **09**, 133 (2022). https://doi.org/10.1007/JHEP09(2022)133. arXiv:2206.03797 [hep-ph]

10. D. Leljak, B. Melić, F. Novak, M. Reboud, D. van Dyk, Toward a complete description of $b \to u\ell^-\bar{\nu}$ decays within the weak effective theory. JHEP **08**, 063 (2023). https://doi.org/10.1007/JHEP08(2023)063. arXiv:2302.05268 [hep-ph]

11. J. Aebischer et al., WCxf: an exchange format for Wilson coefficients beyond the standard model. Comput. Phys. Commun. **232**, 71–83 (2018). https://doi.org/10.1016/j.cpc.2018.05.022. arXiv:1712.05298 [hep-ph]

12. W. Dekens, P. Stoffer, Low-energy effective field theory below the electroweak scale: matching at one loop. JHEP **10**, 197 (2019). https://doi.org/10.1007/JHEP10(2019)197. arXiv:1908.05295 [hep-ph]. [Erratum: JHEP 11, 148 (2022)]

13. J. Aebischer, J. Kumar, P. Stangl, D.M. Straub, A global likelihood for precision constraints and flavour anomalies. Eur. Phys. J. C **79**, 509 (2019). https://doi.org/10.1140/epjc/s10052-019-6977-z. arXiv:1810.07698 [hep-ph]

14. A. Greljo, J. Salko, A. Smolkovič, P. Stangl, SMEFT restrictions on exclusive $b \to u\ell\nu$ decays. (2023). arXiv:2306.09401 [hep-ph]

15. E.G. Tabak, C.V. Turner, A family of nonparametric density estimation algorithms. Commun. Pure Appl. Math. **66**, 145–164 (2013). https://doi.org/10.1002/cpa.21423

16. L. Dinh, J. Sohl-Dickstein, S. Bengio, Density estimation using real NVP. in International Conference on Learning Representations (2017). https://openreview.net/forum?id=HkpbnH9lx

17. V. Stimper, D. Liu, A. Campbell, V. Berenz, L. Ryll, B. Schölkopf, J.M. Hernández-Lobato, normflows: a pytorch package for normalizing flows. J. Open Source Softw. **8**, 5361 (2023). https://doi.org/10.21105/joss.05361

18. D. Leljak, B. Melić, F. Novak, M. Reboud, D. van Dyk, EOS/DATA-2023-01v2: supplementary material for EOS/ANALYSIS-2022-05. (2023b). https://doi.org/10.5281/zenodo.8027015

19. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: an imperative style, high-performance deep learning library. in *Advances in Neural Information Processing Systems*, vol. 32 (Curran Associates, Inc., 2019), pp. 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

20. G. Papamakarios, E. Nalisnick, D.J. Rezende, S. Mohamed, B. Lakshminarayanan, Normalizing flows for probabilistic modeling and inference. J. Mach. Learn. Res. **22** (2021)

21. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization. (2017). arXiv:1412.6980 [cs.LG]

22. L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and Regression Trees* (Wadsworth International Group, Belmont, 1984)

23. T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 (ACM, New York, 2016), pp. 785–794. https://doi.org/10.1145/2939672.2939785

24. H. Reyes-Gonzalez, R. Torre, The NFLikelihood: an unsupervised DNNLikelihood from normalizing flows. (2023). arXiv:2309.09743 [hep-ph]

25. D. van Dyk et al. (EOS Authors), EOS: a software for flavor physics phenomenology. Eur. Phys. J. C **82**, 569 (2022). https://doi.org/10.1140/epjc/s10052-022-10177-4. arXiv:2111.15428 [hep-ph]