Making the most of LHC data

Bayesian neural networks and SMEFT global analysis

Dissertation

Michel Ansgar Luchmann

Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics
of Heidelberg University, Germany

for the degree of

Doctor of Natural Sciences

Put forward by

Michel Ansgar Luchmann

born in Bad Wildungen, Germany

Oral examination: 23.11.2022

Making the most of LHC data

Bayesian neural networks and SMEFT global analysis

Referees: Prof. Dr. Tilman Plehn
Prof. Dr. Monica Dunford

Abstract

The LHC produces huge amounts of data in which signs of new physics can be hidden. To take full advantage of existing or future LHC data, it is worth exploring novel techniques such as deep learning methods or global analysis strategies. We first study BNNs, a deep learning method which has the benefit of providing uncertainty estimates while still performing similarly to ordinary deep neural networks. We show in different studies how BNNs can be applied to LHC physics. A BNN is trained for the task of jet calibration and we illustrate how we can disentangle and understand the predicted uncertainty types. Furthermore, we discuss different ideas of how BNNs can be used for collider event generation, introducing error bars which are necessary when replacing existing Monte Carlo simulations with deep learning methods.

The focus is then shifted to the global analysis of LHC data. To derive accurate bounds on the space of new physics, it is crucial to have an optimal understanding of all the uncertainties involved. We first discuss the results of matching a specific UV model to the SMEFT. The matching procedure introduces an additional theory uncertainty which has a significant impact on the derived bounds. We then study more generally the results of a global SMEFT analysis for different statistical approaches, namely a Bayesian one and a profile likelihood based one. Both procedures are compared and the impact of different uncertainty treatments is discussed. We encounter that it is crucial to describe correlations between measurements as accurate as possible.

Zusammenfassung

Der LHC produziert riesige Datenmengen, in denen sich Anzeichen neuer Physik verbergen können. Um die Vorteile bestehender oder zukünftiger LHC-Daten voll auszuschöpfen, lohnt es sich, neue Techniken wie Deep Learning Methoden oder globale Analysestrategien zu erforschen. Wir untersuchen zuerst BNNs, eine Deep Learning Methode, die den Vorteil hat, Unsicherheitsschätzungen zu liefern, während sie dennoch ähnlich funktioniert wie gewöhnliche tiefe neuronale Netze. Wir zeigen in verschiedenen Studien, wie BNNs auf die LHC-Physik angewendet werden können. Ein BNN wird für die Aufgabe der Jet-Kalibrierung trainiert und wir veranschaulichen, wie wir die vorhergesagten Unsicherheitstypen entwirren und verstehen können. Darüber hinaus diskutieren wir verschiedene Ideen, wie BNNs für die Generierung von Teilchenbeschleuniger-Ereignissen verwendet werden können, und führen Fehlerbalken ein, die erforderlich sind, wenn bestehende Monte Carlo Simulationen durch Deep Learning Methoden ersetzt werden.

Der Fokus wird dann auf die globale Analyse von LHC Daten verlagert. Um genaue Grenzen für den Raum neuer Physik herzuleiten, ist es entscheidend, ein optimales Verständnis aller damit verbundenen Unsicherheiten zu haben. Wir diskutieren zunächst die Ergebnisse des Matchings eines bestimmten UV-Modells an die SMEFT. Das Matching-Verfahren führt eine zusätzliche Theorieunsicherheit ein, die einen signifikanten Einfluss auf die abgeleiteten Grenzen hat. Wir untersuchen dann allgemeiner die Ergebnisse einer globalen SMEFT-Analyse für verschiedene statistische Ansätze, nämlich einen Bayesschen und einen auf Profile Likelihoods basierenden. Beide Verfahren werden verglichen und die Auswirkungen unterschiedlicher Unsicherheitsbehandlungen werden diskutiert. Wir stellen fest, dass es entscheidend ist, Korrelationen zwischen Messungen so genau wie möglich zu beschreiben.

Contents

List of Abbreviations

1LPI	one-light-particle irreducible
1PI	one-particle irreducible
BINN	Bayesian invertible neural network
BNN	Bayesian neural network
BSM	beyond the Standard Model
CR	control region
EFT	effective field theory
EW	electroweak
EWPO	electroweak precision observable
GAN	generative adversarial network
INN	invertible neural network
KL	Kullback-Leibler
LHC	Large Hadron Collider
ML	machine learning
MSE	mean squared error
ReLU	rectified linear unit
SGD	stochastic gradient descent
SM	Standard Model
SMEFT	Standard Model effective field theory
VEV	vacuum expectation value

Preface

The research presented in this thesis has been conducted between 2019 and 2022 in collaboration with other researchers. It covers the following five publications:

[1] Gregor Kasieczka, Michel Luchmann, Florian Otterpohl, Tilman Plehn,
Per-Object Systematics using Deep-Learned Calibration,
SciPost Phys. 9, 089 (2020), arXiv:2003.11099 [hep-ph]

[2] Marco Bellagente, Manuel Haußmann, Michel Luchmann, Tilman Plehn,
Understanding Event-Generation Networks via Uncertainties,
SciPost Phys. 13, 003 (2022), arXiv:2104.04543 [hep-ph]

[3] Ilaria Brivio, Sebastian Bruggisser, Emma Geoffray, Wolfgang Kilian,
Michael Krämer, Michel Luchmann, Tilman Plehn, Benjamin Summ,
From Models to SMEFT and Back?,
SciPost Phys. 12, 036 (2022), arXiv:2108.01094 [hep-ph]

[4] Simon Badger, Anja Butter, Michel Luchmann, Sebastian Pitz, Tilman Plehn,
Loop Amplitudes from Precision Networks,
to be published in SciPost, arXiv:2206.14831 [hep-ph]

[5] Ilaria Brivio, Sebastian Bruggisser, Nina Elmer, Emma Geoffray,
Michel Luchmann, Tilman Plehn,
To Profile or To Marginalize? – A SMEFT Case Study,
to be published in SciPost, arXiv:2208.08454 [hep-ph]

During this time, the author of this thesis was a doctoral student at the Institute for Theoretical Physics (ITP) of the Heidelberg University.

Chapter 1

Introduction

Since its start in 2008, the Large Hadron Collider (LHC) has been the main focus of attention in the field of particle physics. Its greatest contribution is by far the discovery of the Higgs boson [6–8] in 2012 [9–11] which was the last missing piece predicted by the Standard Model (SM) of particle physics. Since then, many new analyses of LHC data have successfully confirmed the SM and helped to constrain the space of new physics [12–14]. No significant deviations from the SM have been found yet. While being extremely successful, it is clear that the SM leaves many questions open such as the nature of dark matter [15], baryogenesis [16], the strong CP problem [17] or quantum gravity [18]. While it is within the range of possibilities that the LHC does not reach energies necessary to find answers to these questions, it is also possible that new physics is still hiding in the vast amount of current or future LHC data.

To take full advantage of the measurements provided by the LHC, it is worth exploring new tools developed in the rapidly growing field of modern machine learning [19]. These techniques which usually rely on deep neural networks can boost sensitivities of searches by optimizing tagging performance for applications such as quark-gluon discrimination [20–25], Higgs-tagging [26, 27] or top-tagging [28]. Other ideas[1] include the detection of anomalies using deep neural networks [30, 31], inverting detector effects [32, 33] or optimizing Monte Carlo simulations [34] required for almost every analysis. While classical machine learning methods such as shallow neural networks or boosted decision trees have been applied to high energy physics for decades [35], the field is slowly moving towards including more and more deep learning methods. LHC physics is ideally suited for the application of these methods. Datasets are typically large and high dimensional and simulations of collider events are accurate and well controlled, providing truth labels for fully supervised learning methods.

While these tools can be very powerful, they usually lack any type of uncertainty estimate which is a necessary ingredient for any physics application. Bayesian neural network (BNN) [36–39] address this issue by providing a framework to include uncertainties in deep learning. BNNs can be used whenever an ordinary neural network could be used. They show similar performance to deterministic networks while providing additional information about uncertainties. The first part of this thesis is dedicated to these networks. Chapter 2 introduces the theory behind Bayesian neural networks, discusses the different approximations taken and gives practical advice on how to implement such a network. Chapters 3 to 5 focus on different possible applications of BNNs to LHC physics. Chapter 3 presents the results of a study in which BNNs were applied to jet calibration, demonstrating that the predicted uncertainties are accurate and well

[1] A more complete overview can be found in Ref. [29].

calibrated. Chapters 4 and 5 show how BNNs can be used for collider event generation, introducing necessary error bars when replacing existing Monte Carlo simulations with deep learning methods.

While Machine Learning techniques can help boost the sensitivity of existing searches, the searches themselves are often highly model dependent. A vast amount of possible new physics models have been developed over the last decades [40, 41], too many to dedicate an analysis to each of them. A more model agnostic analysis strategy is to not assume a specific model but to parameterize new physics with an effective field theory (EFT) approach. The EFT that extends the SM is referred to as the Standard Model effective field theory (SMEFT) [42]. One of the few assumptions in a SMEFT analysis is that new physics – in the form of new particles – lives at a higher energy scale than the one probed by the experiment. This framework is ideally suited for a global analysis where many individual measurements are combined into a single fit, taking full advantage of the vast amount of LHC data. The second part of this thesis is dedicated to this approach. Chapter 6 discusses how to match SMEFT results to a specific model and derives parameter bounds. The difference between a SMEFT and a model-specific analysis is highlighted. Chapter 7 presents the results of a study performed purely within the SMEFT framework and focuses on different statistical approaches, highlighting the difference between a profile likelihood based approach and a Bayesian one.

Chapter 2

Bayesian neural networks and high energy physics

Since the deep learning revolution [43] in the early 2010s there has been a continuously increasing amount of research in the field of machine learning. Various new techniques and ideas have been developed in the past couple of years. While classical machine learning methods have been applied to high energy physics data in many analyses [35], and plenty of research has already been performed using deep neural networks [29], there are still many areas left within the field of high energy physics which could potentially benefit from these new techniques.

LHC physics data is well suited for the application of deep learning methods. Existing datasets are large and typically high dimensional and simulations are accurate and well controlled. The former is usually seen as a necessary ingredient because the training of deep neural networks requires large datasets and the latter provides truth labels which are a needed for fully supervised learning methods.

While deep learning methods are powerful tools, they usually lack uncertainty estimates which are crucial for any physics analysis. Bayesian neural networks (BNN) [36–39] address this issue by providing a framework to include uncertainties in deep learning. This chapter introduces the theory and concept behind BNNs. Section 2.1 provides a very brief introduction to ordinary neural networks, Sec. 2.2 discusses BNNs and Sec. 2.3 focuses on normalizing flows and their Bayesian counterpart.

2.1 Short introduction to neural networks

This section provides a very brief introduction to neural networks. A more detailed overview can for instance be found in Ref. [44] or Ref. [45].

There are various types of neural networks such as fully connected dense neural networks [46], convolutional neural networks [43, 47], graph neural networks [48, 49], transformers [50, 51] and many other. The common characteristic of these different ideas is that they consist of layers and each layer is defined by some simple mathematical operation. The prototype layer which can be found in almost all of these different architectures is the fully connected dense layer. It is defined by a linear operation and a subsequent non-linear operation:

$$\begin{aligned} z &= \hat{W}x + b \\ y &= g(z)\,, \end{aligned} \tag{2.1}$$

where x is the input vector, $\hat{W}$ is the weight matrix, b is the bias vector and $g(\cdot)$ is a non-linear function. In the context of neural networks, this non-linearity is referred to as the activation function. The weights and the entries of the bias vector are free parameters and adjusted in the training process. The most common activation function is the rectified linear unit (ReLU) [46] which is defined as:

$$g(z) = \begin{cases} z & \text{if } z > 0\,, \\ 0 & \text{otherwise} \end{cases}\,. \tag{2.2}$$

Many of such simple operations performed on top of each other is what defines a neural network. While a neural network with only one inner or hidden layer usually has very restricted expressive power, practice has shown that deep neural networks, where deep simply refers to a network with many hidden layers, can approximate almost every function. This high expressivity, the fact that gradients can be computed very efficiently due to algorithms such as back propagation [19] and the fact that the computation of a forward pass can be highly parallelized are upon the main reasons why neural networks became the default choice in the field of machine learning.

While (deep) neural networks can approximate almost any function [52], it comes with the cost of having many tunable parameters. These parameters are adjusted by minimizing a loss function on a given training dataset. The choice of the loss function depends on the task. For regression where the training objective is to learn continuous labels y for a given input x a common choice is the mean squared error (MSE):

$$\begin{aligned} \mathcal{L} = \text{MSE} &= \sum_{i=1}^{M} |y_i^{\text{pred}} - y_i^{\text{truth}}|^2\,, \\ y_i^{\text{pred}} &= f_\theta(x_i)\,, \end{aligned} \tag{2.3}$$

where y^{pred} refers to the label predicted by the neural network $f_\theta(x)$ and y^{truth} to the truth label given by the training dataset. The sum runs over examples of the given dataset. As the name suggests the MSE is a measure for the error the neural network makes. The process of adjusting iteratively the free parameters θ to find a minimal error is what is referred to as the training. While there exist various optimization algorithms to find suitable local or global minima, in modern machine learning they all rely on evaluating gradients with respect to θ. One of the most simple update rules is referred to as gradient descent [53] and given by:

$$\theta_{i+1} = \theta_i + \lambda \nabla \mathcal{L}(\theta)\,, \tag{2.4}$$

where λ is called the learning rate. Because training datasets are large these updates are usually performed with respect to a subset of the full training data and the corresponding algorithm is called stochastic gradient descent (SGD).

2.2 Bayesian neural networks

While it has been shown that deep neural networks can reach excellent performance on different tasks within high energy physics [29], they usually do not come with any type of uncertainty estimate. BNNs [36–39] address this issue by introducing a distribution over the network parameters. This is in contrast to ordinary neural networks as they

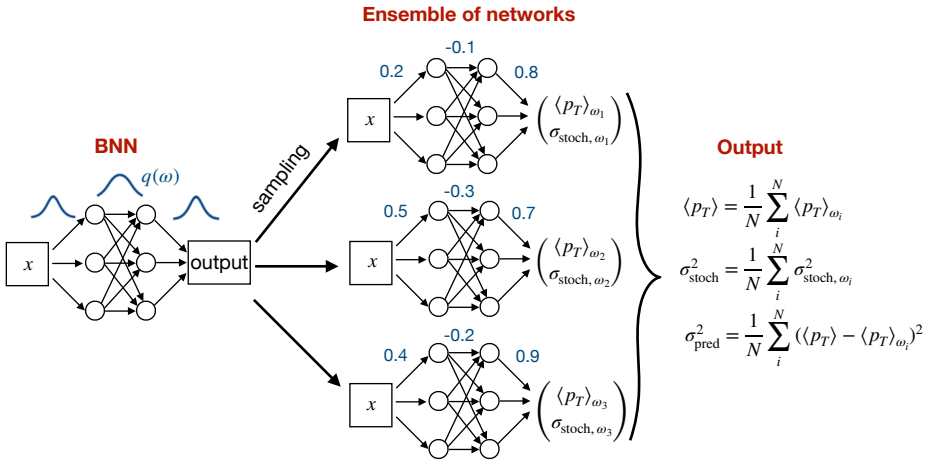


Figure 2.1: Illustration of a BNN as an ensemble of deterministic models. Taken from Ref. [4].

were discussed in Sec. 2.1. After the training procedure the network parameters are fixed to specific values $\hat{\theta}$. A BNN on the other hand treats θ as a random variable with a corresponding probability distribution. By creating samples from this distribution an ensemble of ordinary networks is created, each giving a different prediction. The spread of these predictions incorporates uncertainties related to the model and the training data. This basic intuition about BNNs is illustrated in Fig. 2.1. The two dimensional output of each network presented in the figure is specific to the case of regression and will be discussed in more detail in Sec. 2.2.2.

To define the distribution over network parameters three ingredients are required: the likelihood $p(D|\theta)$ of the training dataset D, the prior $p(\theta)$ and the normalization $p(D)$. Applying Bayes theorem leads to:

$$p(\theta|D) = \frac{p(D|\theta)\, p(\theta)}{p(D)} \sim p(D|\theta)\, p(\theta)\,. \tag{2.5}$$

The problematic part of this expression is $p(D)$ which can be expressed as an integral over θ:

$$p(D) = \int d\theta\, p(D|\theta)\, p(\theta)\,. \tag{2.6}$$

For high dimensional models, it is extremely challenging to perform the integration numerically which is why different techniques have been developed to avoid its computation. One of these methods is referred to as variational inference where $p(\theta|D)$ is approximated by simple tractable distributions. These simple distributions $q_\alpha(\theta)$ are parameterized by parameters α which are then tuned to have $q_\alpha(\theta)$ as close as possible to $p(\theta|D)$. To achieve this, one possibility is to minimize the Kullback-Leibler (KL) divergence [54]:

$$\arg\min_\alpha \mathrm{KL}[q_\alpha(\theta)|p(\theta|D)]\,. \tag{2.7}$$

The KL divergence provides a measure for how similar two distributions are. It is zero if

two distributions are identical and positive otherwise. Inserting Eq. (2.5) leads to:

$$\begin{aligned}\mathrm{KL}[q_\alpha(\theta), p(\theta|D)] &= \int d\theta\ q_\alpha(\theta)\ \log \frac{q_\alpha(\theta)}{p(\theta|D)} \\ &= \int d\theta\ q_\alpha(\theta)\ \log \frac{q_\alpha(\theta)p(D)}{p(\theta)p(D|\theta)} \\ &= \mathrm{KL}[q_\alpha(\theta), p(\theta)] - \int d\theta\ q_\alpha(\theta)\ \log p(D|\theta) + \log p(D) \int d\theta\ q_\alpha(\theta)\ .\end{aligned} \tag{2.8}$$

If $q_\alpha(\theta)$ is properly normalized, the last term is a constant with respect to α:

$$\log p(D) \underbrace{\int d\theta\ q_\alpha(\theta)}_{=1} = \log p(D) \sim \text{const}\,. \tag{2.9}$$

Therefore, we can define the Bayesian loss function as:

$$\mathcal{L}_{\mathrm{BNN}} = -\int d\theta\ q_\alpha(\theta)\ \log p(D|\theta) + \mathrm{KL}[q_\alpha(\theta), p(\theta)]\ . \tag{2.10}$$

The negative of this expression is also referred to as the Evidence Lower Bound (ELBO) [55]. The step of removing the constant term containing $p(D)$ is crucial because it is the reason why $\mathcal{L}_{\mathrm{BNN}}$ is tractable.

Eq. (2.10) consists of two terms. The first term contains the negative log-likelihood with an integral over all possible weight configurations and the second term is the KL-divergence between the variational distribution and the prior. While the former is data dependent, the latter depends only on the prior and is, therefore, a pure regularization term. The integral over all possible weight configurations can be seen as an $q(\omega)$-weighted average of the log-likelihood:

$$\mathbb{E}_\theta \Big[-\log p(D|\theta) \Big] = -\int d\theta\ q_\alpha(\theta)\ \log p(D|\theta)\,. \tag{2.11}$$

Minimizing the first term dictates the network to describe the training data well, while the second term forces the variational distribution to not deviate too much from the prior distribution. This can be seen more explicitly by giving an example: Setting the variational distribution to a Gaussian with a central value of μ_q and a width of σ_q and the prior to a Gaussian with a central value of zero and a width of σ_p the KL-divergence becomes:

$$\mathrm{KL}[q_{\mu,\sigma}(\theta), p_{\mu,\sigma}(\theta)] = \frac{\sigma_q^2 - \sigma_p^2}{2\sigma_p^2} + \frac{\mu_q^2}{2\sigma_p^2} + \log \frac{\sigma_p}{\sigma_q}\ . \tag{2.12}$$

It contains the term $\frac{\mu_q^2}{2\sigma_p^2}$ which is usually referred to as L2-regularization or weight decay [56]. It is a penalty term for large weights and a common regularization technique in machine learning. The other two terms constrain the width of the variational distribution. They become zero for $\sigma_p = \sigma_q$. For simplicity only the result for a one-dimensional θ is shown. In the usual case of a higher dimensional θ the KL-divergence factorizes into a sum of the individual contributions, if independent distributions for each dimension are assumed.

2.2.1 Mean-field approximation

In all of the studies presented in this thesis and in many studies performed in the literature [38, 57] $q_\alpha(\theta)$ is set to a product of Gaussian distributions:

$$q_\alpha(\theta) = \prod_i^K N(\theta_i|\mu_i, \sigma_i)\,, \tag{2.13}$$

where $\alpha = \{\mu_i, \sigma_i\}$ and K is the number of parameters of the neural network. This assumption which is usually referred to as the mean-field approximation [55] introduces two parameters per weight: A mean value and a width. The number of free parameters is twice the amount of the corresponding deterministic network. Anything beyond the mean-field approximation usually does not scale well with the number of network weights. For instance, one could go one step further by introducing correlation terms. This can be achieved by setting $q_\alpha(\theta)$ to a multivariate normal distribution with a covariance matrix Σ:

$$q_\alpha(\theta) = N(\vec{\theta}|\vec{\mu}, \Sigma) \tag{2.14}$$

where the vector symbols were introduced to emphasize that $\vec{\mu}$ is a K-dimensional vector. The mean-field approximation is recovered by setting all off-diagonal values of Σ to zero. While this could lead to an improved approximation of the posterior distribution, it introduces many new trainable parameters. A symmetric matrix with $K \times K$ entries has $K(K+1)/2 \sim K^2$ degrees of freedom which increases the number of trainable parameters significantly compared to the deterministic case. Therefore, this construction is only feasible for shallow neural networks with a small number of weights. While Eq. (2.13) is a strong simplification, the experiments presented in this thesis (compare Chaps. 3 to 5) demonstrate that the uncertainty estimates of a mean-field BNN can still be meaningful and accurate. In the machine learning literature on the other hand it is an open question how well the mean-field approximation works. To give one argument in favor of this approach: Ref. [57] argues that the "deepness" of neural networks can compensate for the lack of correlations directly described by $q_\alpha(\theta)$. The predictive distribution of a deeper neural network with the simple mean-field approximation can be as powerful as the predictive distribution of a shallower network where the posterior is approximated more precisely.

Another important aspect worth mentioning, is the "direction" of the KL-divergence. The KL-divergence is asymmetric in its arguments. While $\mathrm{KL}[p(\theta|D)|q_\alpha(\theta)]$ is referred to as the forward direction, $\mathrm{KL}[q_\alpha(\theta)|p(\theta|D)]$ is referred to as the reverse direction [58]. The forward direction is usually associated to a mode-averaging approximation while the other direction is a uni-mode seeking approximation which means that only one of the modes is approximated well while the other modes are neglected. Equation (2.10) represents the reverse direction. Therefore, $q_\alpha(\theta)$ should be seen as a description of the posterior around only one of its modes. The posterior of a deep neural network is very high dimensional which makes it difficult to visualize. However, one can easily argue that it must possess many modes. Training deterministic deep neural networks several times, starting from different initial weight configurations, results into very different final weight configurations. Even if the performance of the trained networks is very similar the individual weights can have completely different values. The same is usually true for variational BNNs which means that each training converged in a different mode. While this could in principle be problematic for the reliability of the mean-field approximation,

one possible workaround is to simply train a BNN several times starting from different initial weight configurations. If the posterior distribution has indeed many modes, the different trainings may converge to different modes. If the resulting uncertainty estimates are significantly different for the individual trainings, one should be careful about the reliability, otherwise one can proceed with the single (mean-field) BNN. Simple checks like this have always been performed for the studies presented in this thesis.

2.2.2 Bayesian regression

Regression refers to the task of finding a suitable function which approximately describes the relation between a label A and some given input x. While x can be arbitrarily high dimensional, A is usually only a one-dimensional number. To give an example relevant for high energy physics: x could be the detector response of a given hadronic jet and the task is to reconstruct the transverse momentum of the jet. In this case, the label would be given by $A = p_{T,\text{jet}}$ [1].

The loss function of a BNN is given by Eq. (2.10). The problem specific part of this function is the log-likelihood. While it would be possible to insert an arbitrarily complicated expression for the log-likelihood, a common assumption is to simply have some Gaussian noise in the dataset. In this scenario, the likelihood can be modeled as a Gaussian distribution. If each data point of the training dataset is seen as an independent observation, the log-likelihood is then given by:

$$-\log p(D|\theta) = \sum_j^N \left[\frac{\left|\overline{A}(x_j,\theta) - A_j^{(\text{truth})}\right|^2}{2\sigma^2_{\text{stoch}}(x_j)} + \frac{1}{2}\log\sigma^2_{\text{stoch}}(x_j) + \frac{1}{2}\log 2\pi \right] , \tag{2.15}$$

where the sum goes over all points of the training dataset. "truth" indicates that the label was provided by the training dataset. In the p_T-calibration context mentioned above the truth label could be provided by simulations and refers to the p_T before the detector simulation. $\overline{A}(x_j,\theta)$ is the prediction of the model and, therefore, depends on the observation x_j and the neural network parameters θ. The standard deviation of the Gaussian noise is given by $\sigma_{\text{stoch}}(x_j)$. It is in general x-dependent and unknown. Rather than assuming some constant noise or trying to estimate it in some complicated way, the neural network can be constructed to directly predict it. This can be achieved by making the output of the last layer two dimensional:

$$\text{BNN} : x, \theta \ \rightarrow \begin{pmatrix} \overline{A}(x,\theta) \\ \sigma_{\text{stoch}}(x,\theta) \end{pmatrix} . \tag{2.16}$$

In this way σ_{stoch} becomes θ dependent. Omitting constant terms the full loss function becomes:

$$\begin{aligned} \mathcal{L}_{\text{BNN}} = \int d\theta \; q_{\mu,\sigma}(\theta) \sum_j^N \left[\frac{\left|\overline{A}(x_j,\theta) - A_j^{(\text{truth})}\right|^2}{2\sigma^2_{\text{stoch}}(x_j,\theta)} + \log\sigma^2_{\text{stoch}}(x_j,\theta) \right] \\ + \text{KL}[q_\alpha(\theta)|p(\theta)] \, . \end{aligned} \tag{2.17}$$

For the default assumption of uncorrelated Gaussian variational distributions and a Gaussian prior, a closed form solution for the KL-divergence can be given (compare

Eq. (2.12)). Special care should be taken when working with mini-batch optimization. In this case, the KL-term has to be rescaled accordingly (see discussion in Sec. 2.2.4).

Once a BNN is trained by minimizing Eq. (2.17), the prediction of a BNN can be extracted from the predictive distribution which is defined as:

$$p(A|x) = \int d\theta \, q_\alpha(\theta) \, p(A|x,\theta) \,, \tag{2.18}$$

where x is a data point typically given by some statistically independent test dataset. By construction the likelihood $p(A|x,\theta)$ is a Gaussian with its mean value and width given as outputs of the BNN (compare Eq. (2.16)). The prediction can be defined as the mean value of the predictive distribution:

$$\begin{aligned}
\langle A \rangle &= \int dA \, p(A|x) \, A \\
&= \int dA \int d\theta \, q_\alpha(\theta) \, p(A|x,\theta) \, A \\
&= \int d\theta q_\alpha(\theta) \underbrace{\int dA \, p(A|x,\theta) \, A}_{=\overline{A}(\theta)} \\
&= \int d\theta \, q_\alpha(\theta) \overline{A}(x,\theta) \,.
\end{aligned} \tag{2.19}$$

By exchanging integrals, $\langle A \rangle$ can be expressed as an q-weighted average of $\overline{A}(x,\theta)$. The uncertainty on this prediction is given by the variance of the predictive distribution:

$$\begin{aligned}
\sigma^2_{\text{tot}} &= \langle (A - \langle A \rangle)^2 \rangle \\
&= \int dA \, (A - \langle A \rangle)^2 \, p(A|x) \\
&= \int dA \int d\theta \, (A - \langle A \rangle)^2 \; q_\alpha(\theta) \, p(A|x,\theta) \\
&= \int d\theta \; q_\alpha(\theta) \left[\int dA \; A^2 \; p(A|x,\theta) - 2\langle A \rangle \int dA \; A \; p(A|x,\theta) + \langle A \rangle^2 \int dA \; p(A|x,\theta) \right] \\
&= \int d\theta \; q_\alpha(\theta) \left[\overline{A^2}(x,\theta) - \overline{A}(x,\theta)^2 + \left(\overline{A}(x,\theta) - \langle A \rangle \right)^2 \right] \equiv \sigma^2_{\text{stoch}} + \sigma^2_{\text{pred}} \,.
\end{aligned} \tag{2.20}$$

In the last line two separate sources of uncertainty were identified:

$$\sigma^2_{\text{pred}} = \int d\theta \; q_\alpha(\theta) \left[\overline{A}(x,\theta) - \langle A \rangle \right]^2 \,, \tag{2.21}$$

and

$$\begin{aligned}
\sigma^2_{\text{stoch}} &= \int d\theta \; q_\alpha(\theta) \left[\overline{A^2}(x,\theta) - \overline{A}(x,\theta)^2 \right] \\
&= \int d\theta \; q_\alpha(\theta) \; \sigma^2_{\text{model}}(x,\theta) \\
&= \langle \sigma^2_{\text{stoch}}(x,\theta) \rangle \,.
\end{aligned} \tag{2.22}$$

To understand the difference between these two uncertainties, it is helpful to consider the case of large training statistics. From general arguments about parameter estimation, the width of the posterior distribution $p(\theta|D)$ should get smaller if the number of training examples is increased. The same should hold for $q_\alpha(\theta)$ because it is an approximation

of $p(\theta|D)$. Therefore, in the limit $N \to \infty$ where N is the training size $q_\alpha(\theta)$ becomes a delta distribution:

$$q_\alpha(\theta) \to \delta(\theta - \theta_0) \quad \text{for} \quad N \to \infty \,. \tag{2.23}$$

Evaluating all the integrals for this limiting case results into:

$$\begin{aligned}
\langle A \rangle &\to \int d\theta \, \delta(\theta - \theta_0) \, \overline{A}(x, \theta) = \overline{A}(x, \theta_0) \\
\sigma^2_{\text{pred}} &\to \int d\theta \; \delta(\theta - \theta_0) \; \left[\overline{A}(x, \theta) - \langle A \rangle \right]^2 \; = 0 \\
\sigma^2_{\text{stoch}} &\to \int d\theta \, \delta(\theta - \theta_0) \, \sigma_{\text{stoch}}(x, \theta)^2 = \sigma^2_{\text{stoch}}(x, \theta_0) \\
\sigma^2_{\text{tot}} &\to \sigma^2_{\text{stoch}}(\theta_0) \,.
\end{aligned} \tag{2.24}$$

This illustrates the key difference between both uncertainties: While σ_{pred} goes to zero for infinite training statistics, σ_{stoch} approaches a constant. σ_{pred} describes the part of the total uncertainty which is there due to limited training statistics. It incorporates uncertainties from the weight distributions. σ_{stoch} on the other hand is the remaining part. It should be roughly constant as a function of training size and capture inherent noise of the training dataset and/or limited expressive power of the model itself. These key differences will be further discussed in Chaps. 3 and 5. In the machine learning literature the two sources of uncertainties are referred to as epistemic and aleatoric uncertainty with the former being captured by σ_{pred} and the latter by σ_{stoch} [59]. These terms were not adopted in the work presented in this thesis because they are not commonly used in the field of high energy physics.

The arguments from above hold even if $p(\theta|D)$ is multi-modal. If $p(\theta|D)$ is multi-modal with degenerate modes, the width around each of these modes should get smaller when observing more data points and $q_\alpha(\theta)$ can be seen as an expansion around one of these modes (compare discussion of Sec. 2.2.1).

All of the expressions above can be computed via Monte Carlo integration by sampling from the distribution $q_\alpha(\theta)$:

$$\begin{aligned}
\langle A \rangle \approx \frac{1}{L} \sum_i^L \overline{A}(\theta_i) \,, \qquad \sigma^2_{\text{stoch}} \approx \frac{1}{L} \sum_i^L \sigma^2_{\text{stoch}}(\theta_i) \,, \qquad \sigma^2_{\text{pred}} \approx \sum_i^L (\overline{A}(\theta_i) - \langle A \rangle)^2 \,, \\
\text{with } \theta_i \sim q_\alpha(\theta) \,.
\end{aligned} \tag{2.25}$$

Thus, by simply generating a set of random numbers from $q_\alpha(\theta)$ which is in our case a product of Gaussian distributions, evaluating a forward pass through the neural network, and repeating this L times the prediction and all of the uncertainties can be computed efficiently. In the research presented in this thesis L was typically set to 50.

2.2.3 Bayesian classification

Classification refers to the task of identifying to which category a certain object x belongs to. To give an example in the context of high energy physics, x could be a calorimeter image of a jet and possible classes could be top jets and QCD jets [28]. A neural network is then trained to predict to which of these two categories the jet image belongs to.

To be able to apply a BNN to this problem an expression for the log-likelihood has to be constructed. Let $p(c|x)$ be the probability of x being a member of class c. In high energy physics there are usually only two classes: Something which is considered interesting, signal, and something which is considered less interesting, background. They are exclusive classes. For this binary case the probability that x belongs to the opposing class is $p(\neg c|x) = 1 - p(c|x)$. If each data point of the training dataset is seen as an independent observation, the log-likelihood is given by:

$$\begin{aligned} -\log p(D|\theta) &= \sum_i^N \log p(c_i^{\text{truth}}|x_i, \theta) \\ &= \sum_i^N \log \left(p(c|x_i,\theta)^{y_i^{\text{truth}}} \left(1 - p(c|x_i,\theta)\right)^{1-y_i^{\text{truth}}} \right) \\ &= \sum_i^N y_i^{\text{truth}} \log p(c|x_i,\theta) + (1 - y_i^{\text{truth}}) \log(1 - p(c|x_i,\theta)) \end{aligned} \tag{2.26}$$

where c stands for one of the two classes and y_i^{truth} refers to a binary label which is 1 if c is the correct class and 0 otherwise. Eq. (2.26) is called the binary cross entropy and is the most common loss function for binary classification. As for the case of regression (compare Sec. 2.2.2) plugging this expression into Eq. (2.10) defines the Bayesian loss function. However, unlike in the regression case there is no additional parameter appearing to describe noise. The loss function is fully defined by $p(c|x_i, \theta)$ which is simply given by the output of the (Bayesian) neural network:

$$\text{BNN} : x, \theta \to f(x,\theta) \equiv p(c|x,\theta)\,. \tag{2.27}$$

In this case the activation function of the last layer is usually set to the sigmoid function to constrain the output between zero and one which is required to give it its probabilistic interpretation.

Similar to the regression case the prediction of a BNN can be extracted from the predictive distribution which is defined as:

$$p(c|x) = \int d\theta\, q_\alpha(\theta)\, p(c|x,\theta)\,. \tag{2.28}$$

By selecting a threshold t one can then assign a class to each x. A data point fulfilling $p(c|x) > t$ is given the label c and everything below the threshold the opposing class. The predictive uncertainty can be defined as:

$$\sigma^2_{\text{pred}} = \int d\theta\, q_\alpha(\theta)\, \left(p(c|x,\theta) - p(c|x)\right)^2\,. \tag{2.29}$$

Experiments in high energy physics regarding this uncertainty were, for instance, performed in Ref. [60]. For further information about how to construct an uncertainty similar to Eq. (2.22) see for instance Ref. [61].

2.2.4 Reparameterization trick and how to implement a BNN

The loss function of a BNN (compare Eq. (2.10)) consists of two terms: an integral over the log-likelihood and the KL-divergence of the variational distribution and the prior. For all studies presented in this thesis the mean field approximation (see Sec. 2.2.1) was

used with each individual distribution set to a Gaussian. In addition, the prior were set to Gaussian distributions as well. For these particular choices a closed form solution for the second term can be given (see Eq. (2.12)) which makes the implementation trivial. In contrast, the log-likelihood part is more challenging to compute because it involves an integration over all possible weight configurations. Solving the integral via Monte Carlo by sampling from the distribution $q_{\mu,\sigma}(\theta)$, leads to:

$$\begin{aligned}\mathcal{L}_{\text{BNN}} &\supset \int d\theta\, q_{\mu,\sigma}(\theta)\, \log p(D|\theta) \\ &\approx \frac{1}{L}\sum_i^L \log p(D|\theta_i) \quad \text{with } \theta_i \sim q_{\mu,\sigma}(\theta)\,,\end{aligned} \tag{2.30}$$

where L is the number of points sampled. However, for the training of a BNN the gradients with respect to σ and μ are required and the dependency on these parameters seems to be lost in Eq. (2.30). This issue can be solved by taking advantage of the fact that a random sample, θ, of a Gaussian distribution can always be written as:

$$\theta = \mu + \sigma\epsilon \quad \text{with } \epsilon \sim N(0,1)\,, \tag{2.31}$$

where $N(0,1)$ is a standard normal distribution. In this way the randomness is separated from the parameters of interest. This idea is referred to as the reparameterization trick [38, 62]. The gradients can then simply be computed via the chain rule:

$$\begin{aligned}\nabla_{\mu/\sigma}\frac{1}{L}\sum_i^L \log p(D|\theta_i) \quad &\text{with } \theta_i = \mu + \sigma\epsilon,\ \epsilon \sim N(0,1) \\ &= \frac{1}{L}\sum_i^L \Big(\nabla_{\theta_i}\log p(D|\theta_i)\Big)\nabla_{\mu/\sigma}\,\theta_i(\mu,\sigma) \\ &= \frac{1}{L}\sum_i^L \nabla_{\theta_i}\log p(D|\theta_i)\begin{cases}1 \text{ for } \mu \\ \epsilon \text{ for } \sigma\end{cases}\,.\end{aligned} \tag{2.32}$$

L is usually set to 1 which provides only a very rough approximation of the actual loss function and the corresponding gradients. However, because the gradients are only computed on mini-batches they are stochastic in nature and the Monte Carlo error only adds additional noise to an already noisy estimate. As typically argued in the context of mini-batch optimization the stochasticity of the gradients does not need to be a disadvantage, in many cases it can actually help moving away from local minima. For the studies discussed in this thesis a slight variation of this algorithm has been used which is referred to as the local reparameterization trick. The basic idea is the same but the advantage is that networks usually converge faster. For further information see Ref. [63].

A simply way to implement a BNN with a library such as `Pytorch` [64] is by replacing the forward pass of a layer, for instance of a fully connected dense layer, with a modified forward pass involving the (local) reparameterization trick. The mathematical operations defining the layer should remain the same, only an additional sampling step has to be added (see Eq. (2.31)). By making μ and σ trainable parameters everything else will be taken care of by `Pytorch` via `Pytorch`'s automatic differentiation package [65] and backpropagation [19]. Special caution should be taken when defining the full loss function. If performing mini-batch optimization, it is common to compute the mean over the

training examples instead of the sum:

$$\log p(D|\theta) = \sum_{i=1}^{M} p(x_i|\theta) \;\rightarrow\; \frac{1}{M}\sum_{i=1}^{M} p(x_i|\theta)\,, \tag{2.33}$$

where x_i represents one point of the training dataset. To derive the correct prefactors of the two contributions, the loss function is first divided by the total number of training examples N and then approximated with a mini-batch of data:

$$\begin{aligned}\tilde{\mathcal{L}}_{\text{BNN}} \coloneqq \frac{\mathcal{L}_{\text{BNN}}}{N} &= \underbrace{\frac{1}{N}\sum_{i=1}^{N}\log p(D|\theta_k)}_{\approx\frac{1}{M}\sum_{i=1}^{M}\log p(D|\theta_k)} + \frac{1}{N}\text{KL}[q(\theta)|p(\theta)] \\ &\approx \frac{1}{M}\sum_{i=1}^{M}\log p(D|\theta_k) + \frac{1}{N}\text{KL}[q(\theta)|p(\theta)]\,, \\ &\qquad \text{with } \theta_k = \mu + \sigma\epsilon_k,\ \epsilon_k \sim N(0,1)\end{aligned} \tag{2.34}$$

where M is the mini-batch size. The correct prefactor of the KL-divergence is $\frac{1}{N}$ and not $\frac{1}{M}$ as one would expect naively. As a final remark: Instead of making σ a trainable parameter, usually the expression $\log\sigma^2$ is used. The expression $\log\sigma^2$ has the advantage that it is, unlike σ, not strictly positive. The gradient updates during training could push σ to negative numbers. Therefore, it is numerical much more stable to learn $\log\sigma^2$ instead. The same parameterization is usually used for the output of a BNN in the case of regression (compare Eq. (2.16)) for the same reasons.

2.2.5 Maximum a posteriori and deterministic networks

While for a BNN θ is treated as a random variable and described by the distribution $p(\theta|D)$, for an ordinary neural network the network parameters are fixed to specific values $\hat{\theta}$. These values are found in the training process and can correspond to some local or global minimum of the loss function. To get more intuition about the relation between $p(\theta|D)$ and the training of an ordinary neural network, it is worth to consider the special case of a Gaussian likelihood and a Gaussian prior. The negative log-posterior can be written as:

$$\begin{aligned}L(\theta) &= -\log p(\theta|D) \\ &= -\log p(D|\theta) - \log p(\theta)\end{aligned} \tag{2.35}$$

Treating each training example as an independent observation the log-likelihood becomes:

$$-\log p(D|\theta) = \sum_{i}^{N}\left(\frac{1}{2}\log 2\pi\sigma_i^2 + \frac{1}{2\sigma_i^2}|y_i^{\text{pred}} - y_i^{\text{tuth}}|^2\right)\,, \tag{2.36}$$

where N is the number of training samples. With the additional assumption of having a constant width ($\sigma_i = \sigma$) the normalization term is a constant and can be neglected:

$$-\log p(D|\theta) \sim \sum_{i}^{N}\frac{1}{2\sigma^2}|y_i^{\text{pred}} - y_i^{\text{tuth}}|^2\,. \tag{2.37}$$

Repeating the same for a prior with a width of σ_p and a central value of 0 leads to:

$$-\log p(\theta) \sim \frac{1}{2\sigma_p} \sum_j^K \theta_k^2 \tag{2.38}$$

where the sum runs over all network parameters. Combining both and neglecting all constant terms leads to:

$$\begin{aligned} L(\theta) &\sim \sum_i^N \frac{1}{2\sigma^2} |y_i^{\text{pred}} - y_i^{\text{tuth}}|^2 + \frac{1}{2\sigma_p} \sum_j^K \theta_k^2 \\ &\sim \sum_i^N |y_i^{\text{pred}} - y_i^{\text{tuth}}|^2 + \underbrace{\frac{\sigma}{\sigma_p}}_{=:\lambda} \sum_j^K \theta_k^2 \end{aligned} \tag{2.39}$$

which is the MSE loss introduced in Sec. 2.1 with an additional regularization term. This additional regularization term is called L2-regularization or weight decay [56] and is a common technique used to reduce overfitting. λ is a free parameter which dictates the relative importance of this regularization term. The derivation shows that a loss function typically used in regression can be recovered from the posterior used for BNNs. While a BNN is constructed by sampling from the distribution $p(\theta|D)$ or an approximation of it, a deterministic network is defined by the maximum of it. However, one should note that the MSE-loss was only recovered after making simplifying assumptions such as having input independent noise.

2.3 Normalizing flow networks

Normalizing flow networks [66–69], (variational) Autoencoders [62, 70], Generative Adverserial Networks [71, 72] and diffusion models [73, 74] are some of the many different ideas which have been developed for the task of generating samples of complicated and high dimensional probability distributions. For instance, in the field of computer vision one is interested in generating realistic looking images. The naive dimensionality of such a space of images is the number of pixels. An image with only 28x28 pixels as it is the case for the MNIST dataset [75] is already 729 dimensional. In the field of high energy physics one application of generative networks is the generation of collider events. While there are very powerful and accurate simulations available, they are typically computationally very expensive. A generative model can solve this problem because once its trained the generation of new examples is typically much faster than a full simulation based on first principles. While all of the approaches mentioned above can be used for this particular task, this chapter will focus on introducing normalizing flows. Normalizing flows were the tool of choice in the studies presented in Ref. [2] which is discussed in Chap. 4. The advantage of a normalizing flow network compared to the other approaches is that it is a likelihood based model. This makes it easy to formulate the loss function of the corresponding Bayesian version which provides a framework to introduce uncertainty estimates for the learned probability densities.

The idea of a normalizing flow network is to transform a random number z drawn from a simple probability distribution (typically a Gaussian) into a sample x following the distribution of interest. By choosing a bijective transformation one can make use of the

change of variable formula:

$$p_x(x) = p_z\left(f^{-1}(x)\right)\left|\det\left(\frac{\partial f^{-1}(x)}{\partial x}\right)\right|, \tag{2.40}$$

where $x = f(z)$ is the map from z to x; $p_z(z)$ is the prior of z which is called the latent space in this context, $p_x(x)$ is the distribution of interest and $\det(\cdot)$ is the determinant of the Jacobian. The function $f(x)$ is in general unknown. A normalizing flow tries to approximate $f(x)$ by a bijective transformations which depend on tunable parameters θ. The change of variable formula becomes:

$$p_x(x|\theta) = p_z\left(f^{-1}(x;\theta)\right)\left|\det\left(\frac{\partial f^{-1}(x;\theta)}{\partial x}\right)\right|, \tag{2.41}$$

where $f(z;\theta)$ is the transformation given by the normalizing flow. The distribution $p_x(x|\theta)$ is an approximation of the true distribution $p_x(x)$. By having a training dataset D with data points drawn from $p_x(x)$ the loss function can be defined via the negative log likelihood as:

$$\begin{aligned}\mathcal{L}_{\mathrm{NF}} &= -\log p(D|\theta) = -\sum_i^N \log p_x(x_i|\theta) \\ &= -\sum_i^N \log p_z\left(f^{-1}\left(x_i;\theta\right)\right) - \log\left|\det\frac{\partial f^{-1}(x_i;\theta)}{\partial x}\right|,\end{aligned} \tag{2.42}$$

where the sum runs over points of the training dataset. Because the samples of the training dataset follow the true distribution, minimizing this object should push $p_x(x|\theta)$ as close as possible to the true distribution $p_x(x)$. By setting p_z to a multivariate Gaussian distribution centred around 0 with unit covariance matrix the loss function becomes:

$$\begin{aligned}\mathcal{L}_{\mathrm{NF}} &= \sum_i^N \frac{\left(f^{-1}\left(x_i;\theta\right)\right)^2}{2} + \frac{n}{2}\log 2\pi - \log\left|\det\frac{\partial f^{-1}(x_i;\theta)}{\partial x}\right| \\ &= \sum_i^N \frac{\left(f^{-1}\left(x_i;\theta\right)\right)^2}{2} - \log\left|\det\frac{\partial f^{-1}(x_i;\theta)}{\partial x}\right| + \text{const}\,,\end{aligned} \tag{2.43}$$

where n is the number of dimensions of x. After the normalizing flow has been trained, new samples of $p_x(x|\theta)$ can be created by simply drawing samples from $p_z(z)$ and then computing $f(z;\theta_{\mathrm{trained}})$. An alternative interpretation of the loss function is provided by the fact that minimizing $\mathcal{L}_{\mathrm{INN}}$ can also be seen as minimizing the KL-divergence between the true distribution and the learnt distribution.

By construction $f(x;\theta)$ needs to be a bijective transformation which introduces several requirements for a normalizing flow. First, the dimensionality of the latent space has to be the same as the dimensionality of the target space. Second, $f(x;\theta)$ needs to be invertible. While the first requirement is easy to fulfill and also necessary for the invertibility, the second requirement restricts the type of transformations strongly. For instance, ordinary dense neural networks are typically not invertible, thus, not suited to be used out of the box for $f(x;\theta)$. Furthermore, the determinant of the Jacobian has to be efficiently computable. The Jacobian directly enters the loss function which has to be computed for every gradient update during the training process. If there is no easy analytic expression for the Jacobian of this transformation the training would be intractable. To fulfill these

requirements flow networks are constructed as a series of simple transformations each individually fulfilling the conditions. One possible type of flow networks are build from so called affine coupling blocks [67]. Within an affine coupling block an input vector x is split up into two halves x_1 and x_2. One of the two parts is kept unchanged while for the other one a translation and rescaling is applied:

$$\begin{aligned} y_1 &= x_1 \\ y_2 &= x_2 \odot \exp\left(s(x_1)\right) + t(x_1) \end{aligned} \tag{2.44}$$

where t and s are arbitrarily complicated functions typically given by a series of fully connected dense layer. $\odot$ refers to the element-wise product. The affine coupling block fulfills the desired properties by being easily invertible and the determinant of the Jacobian is efficiently computable because it is triangular:

$$\frac{\partial y}{\partial x} = \begin{pmatrix} \mathbb{I} & 0 \\ \frac{\partial y_2}{\partial x_1} & \exp\left(s(x_1)\right) \end{pmatrix} . \tag{2.45}$$

Because a part of the input is never actually changed several of such coupling blocks have to be applied after each other, each time permuting the output dimensions such that no input dimension stays unchanged. Stacking many of such coupling blocks on top of each other can lead to very expressive models.

2.3.1 Bayesian normalizing flows

The derivation of the loss function of a Bayesian normalizing flow is completely analogous to the one presented in Sec. 2.2. The negative log-likelihood is given by Eq. (2.42) and can be plugged into Eq. (2.17). The resulting loss function is:

$$\begin{aligned} \mathcal{L}_{\text{BINN}} = \sum_i^N \int d\theta q_\alpha(\theta) \left(\frac{f^{-1}\left(x_i;\theta\right)^2}{2} - \log \left| \det \frac{\partial f^{-1}(x_i;\theta)}{\partial x} \right| \right) \\ + \text{KL}[q_\alpha(\theta), p(\theta)] \end{aligned} \tag{2.46}$$

where the additional integration and regularization term are computed as discussed in Sec. 2.2.4.

A Bayesian normalizing flow comes with two sampling steps: sampling from the latent space distribution $p_z(z)$ and sampling from the learned weight distribution $q_\alpha(\theta)$. The additional Bayesian specific sampling step can be used to derive uncertainty bands on the joint target density $p_x(x)$. A normalizing flow provides an analytic expression for the target density via:

$$p_x(x|\theta) = p_z(z|\theta) \left| \det \left(\frac{\partial f^{-1}(x)}{\partial x} \right) \right| \tag{2.47}$$

For a Bayesian normalizing flow the joint density can be constructed by marginalizing over θ:

$$\hat{p}_x(x) = \mathbb{E}_q[p_x(x|\theta)] = \int d\theta q_\alpha(\theta)\, p_x(x|\theta)\,, \tag{2.48}$$

where the notation $\hat{p}$ was introduced to not mistake the marginal distribution with the

true distribution $p_x(x)$. Similar to the definition of the predictive variance in the previous sections (see Eq. (2.21)) an uncertainty on the joint distribution can be defined as:

$$\sigma^2_{\text{pred}} = \text{Var}_q[p_x(x|\theta)] = \int d\theta q_\alpha(\theta) \; (p_x(x|\theta) - \hat{p}_x(x))^2 \; . \tag{2.49}$$

While this definition is straight forward and computationally efficient, in a typical application of normalizing flows to high energy physics one is usually interested in the marginalized distributions where most of the dimensions of x are integrated out. Using a normalizing flow to generate collider events the interest usually lies in, for instance, the p_T of one of the final state particles. Unfortunately, the normalizing flow only provides an analytic estimate for the joint distribution. To get an estimate for the marginal distribution, samples have to be created. This is done by first generating samples from $p_z(x)$, then computing $f(z|\theta)$ and as a last step creating a histogram in the dimension of interest. By repeating this procedure several times for different θ where each θ was sampled from $q_\alpha(\theta)$ an uncertainty can be defined similar to Eq. (2.49). Each sample $\theta \sim q_\alpha(\theta)$ produces a different set of samples $x \sim p_x(x|\theta)$ and correspondingly a different histogram height. The standard deviation of these histogram heights provides an estimate of the uncertainty. Formally one can write:

$$\hat{p}_x(x_1) = \mathbb{E}_q[p_x(x_1|\theta)] = \int d\theta \, q_\alpha(\theta) p(x_1|\theta) \, , \tag{2.50}$$

and

$$\sigma^2_{\text{pred}} = \text{Var}_q[p_{x_1}(x|\theta)] = \int d\theta \, q_\alpha(\theta) \; (p_x(x_1|\theta) - \hat{p}_x(x_1))^2 \; , \tag{2.51}$$

with

$$p_x(x_1|\theta) = \int d^{n-1}x \, p_x(x|\theta) \, . \tag{2.52}$$

The interpretation and meaning of this uncertainty estimate is discussed in more detail in Chap. 4 by focusing on experiments performed on toy datasets and a high energy physics dataset.

Chapter 3

Jet calibration with Bayesian neural networks

The research presented in this chapter has been published in Ref. [1]. Most of the figures, tables and text are completely identical to the content of this publication.

3.1 Introduction

Modern methods of machine learning are becoming a crucial tool in experimental and theoretical particle physics. An especially active field in this direction is subjet physics and jet tagging [76], where multi-variate analyses of high-level observables are being replaced with deep neural networks working on low-level inputs. Early applications of deep learning techniques in LHC physics rely on image recognition of jet images [77, 78]. Their main challenge is to combine calorimeter and tracking information, motivating graph convolutional networks and point clouds [79]. Established benchmarks processes for these methods include quark-gluon discrimination [20–25], flavor tagging [80], W-tagging [81–84], Higgs-tagging [26, 27], or top-tagging [47, 83–91]. By now we can consider top jet classification at the level of tagging performance as essentially solved [28, 92]. This gives us room to consider question beyond the performance, for instance what the networks are learning, how they can be visualized, how robust they are, how we can control the uncertainties, and how machine learning methods affect typical LHC analyses structurally.

One open question is driven by particle physics' obsession with error bars: how do we quantify the different uncertainties in analyses using neural networks [60, 93–95]? This question is related to visualization [96], understanding the relevant physics features [97–101], and weakly supervised learning approaches [30, 31, 102–107] — all combined under the general theme of explainable AI. In LHC physics we have the advantage of excellent Monte Carlo simulations and full control of the experimental setup. This allows us to define and control different sources uncertainties very precisely. If we accept that a neural network is just a function relating training data to an output there exist (at least) two main kinds of uncertainties:

1. first, labeled training data comes with statistical and systematic uncertainties, where we define the former as uncertainties which vanish with more training data. The systematic uncertainties can be Gaussian or include shifts, depending on their sources. Unstable network training also belongs to this category of training-induced uncertainties [60];

2. second, on the test data or analysis side we also encounter statistical and systematic uncertainties. When we include an inference or any kind of analysis we also encounter model or theory uncertainties [93]. For these uncertainties it is crucial that we ensure our analysis outcome is conservative.

In a previous paper [60] we have shown how Bayesian classification networks can track uncertainties and provide jet-by-jet error bars for the tagging output. Such a Bayesian network can supplement a probabilistic classification output of '60% signal' with an error estimate of the kind '(60 ± 10)% signal' for a given jet. This kind of jet-by-jet information exceeds what is available from standard LHC classification tools. In principle, this approach covers both, statistical errors from the size of the training sample and systematic uncertainties for instance from the calibration of the training sample. However, our quantitative analysis of Bayesian top taggers encountered practical limitations, for instance that the jet energy scale simultaneously affects the central value and the error bar of the probabilistic output. A similar study of uncertainties just appeared for a matrix element regression task [108].

In this follow-up study we look at this problem from a slightly different angle, now defining the *regression task* of extracting the energy of a tagged top quark inside a fat jet. Again, we translate statistical and systematic uncertainties from the training sample to the test output. The Bayesian network, introduces in Sec. 3.2, allows us to construct a per-jet probability distribution function over possible top momenta, or $p(p_t|\text{fat jet})$. The main advantage of using the regression task as example is that it does not enforce a closed interval for the network output and hence removes the correlation between central value and error estimate in the network output. We use this advantage to cleanly separate effects from the finite size of the training sample and from the stochastic nature of the training sample in Sec. 3.4.

In Sec. 3.5 the stochastic uncertainty leads us to a discussion of systematics in the sense of training-related uncertainties which do not shrink with more training data. Our regression task naturally leads us to developing a framework to calibrate deep network taggers and account for uncertainties in the training sample. We find that a straightforward treatment should be based on smearing the momentum labels in the training sample. It directly accounts for the uncertainties in the underlying measurements of the calibration sample and treats them as an additional systematic effect on the top momentum measurement. As before, the Bayesian network allows us to cleanly separate all different sources of uncertainty.

Our simple application serves as an example how we can use Bayesian networks to define statistical and systematic uncertainties coming from the training sample and affecting the network output. These error bars are defined jet by jet, or event by event, giving us more control than standard methods do. Training on smeared labels allows us to implement energy calibration in a straightforward and automized manner. While our modeling of uncertainties on the reference measurements for calibration is simplified, our approach can be extended in a straightforward manner. For instance, the effect of different jet algorithms or different Monte Carlo simulations can be implemented as a non-Gaussian contribution to the label smearing. The key observation is that Bayesian networks allow us to quote uncertainties from all kinds of statistical and systematic limitations of the labeled training data.

3.2 Bayesian regression for jet calibration

For this study, we are interested in extracting the transverse momentum p_T of a hadronically decaying top quark from a fat top jet. If we define $p(p_T|j)$ as the probability over possible p_T values for a given top jet, j, we can extract the mean value as:

$$\langle p_T \rangle = \int dp_T \, p_T \, p(p_T|j) \, . \tag{3.1}$$

As discussed in Sec. 2.2.2 this distribution is generated by integrating over the weight distributions:

$$p(p_T|j) = \int d\omega \, p(p_T|\omega, j) \, p(\omega|M) \approx \int d\omega \, p(p_T|\omega, j) \, q(\omega) \, , \tag{3.2}$$

where we introduced the variational distributions $q(\omega)$ to approximate the integration. Continuing from this expression the derivation of $\mathcal{L}_{\text{BNN}}$ and the two types of uncertainties is completely analogous to the one presented int Sec. 2.2.2. In this chapter we adopt the notation σ_{stoch} instead of σ_{model} because the limiting factor in this study is the stochastic noise in the dataset or in other words the energy resolution of each jet and not the expressive power of the model.

While we will introduce a more sophisticated likelihood in Sec. 3.5, the simplest construction is to set the likelihood $p(p_T|\omega j)$ to a Gaussian. This is again in complete analogy to Sec. 2.2.2. A Gaussian comes with two parameters: a width and a mean. By making the network output 2 dimensional

$$\text{NN}(\omega) = \begin{pmatrix} \langle p_T \rangle_\omega \\ \sigma_{\text{stoch},\omega} \end{pmatrix} \, , \tag{3.3}$$

we get an analytic expression for the likelihood for each input-jet and weight sample ω. To extract the per-jet probability distribution $p(p_T|j)$ following Eq. (3.2), we usually rely on Monte Carlo integration by sampling weights from the weight distributions $q(\omega)$. For large training statistics the distribution $q(\omega)$ should become narrow. The effect of a finite width of $q(\omega)$ can be tracked by σ_{pred}, so in the limit $\sigma_{\text{pred}} \ll \sigma_{\text{stoch}}$ we can approximate $p(p_T|j)$ as a Gaussian with weight-independent mean $\langle p_T \rangle$ and width σ_{stoch}.

3.3 Data set and network

The correct and precise reconstruction of the momentum of tagged top quarks is important for instance in top resonance searches and has influenced the design of many top taggers [109]. Our data set is therefore similar to standard top tagging references, with some modifications which simplify our regression task. We generate a sample of $R = 1.2$ top jets in the range $p_{T,t}^{\text{truth}} = 400 \, ... \, 1000$ GeV with Pythia [110] at 14 TeV collider energy and the standard ATLAS card for Delphes [111]. We always neglect multi-parton interactions and always include final state radiation. Given initial state radiation we work with two event samples, one with ISR switched on and one with ISR switched off. We require the jets to be central $|\eta_j| < 2$ and truth-matched in the sense that each fat jet has to have a top quark within the jet area. These settings essentially correspond to the public top tagging data set from Refs. [86] and [28]. The difference to the standard tagging reference sets is that we flatten our data set in $p_{T,t}^{\text{truth}}$, such that even

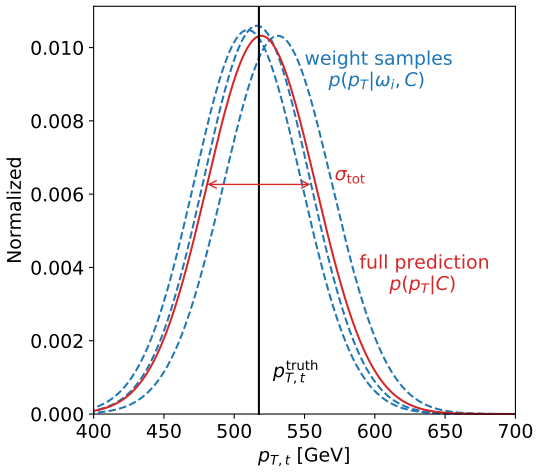


Figure 3.1: Illustration of the predicted distribution from our Bayesian setup for a single top jet. We show the individual predictions from sampling the weights (petrol) as well as the aggregate prediction (red) and the corresponding per-jet uncertainty σ_{tot}.

accounting for bin migration effects we can safely assume that in the fat jet momentum the sample is flat for $p_{T,j} = 500 \ldots 800$ GeV.

The final result of our Bayesian network will be a probability distribution over possible $p_{T,t}$ values for a given jet. For our labeled data we know the corresponding $p_{T,t}^{\text{truth}}$. However, the fact that we will modify this truth label as part of the calibration training makes it the less attractive option to organize our samples. The closest alternative observable is the momentum of the fat jet, so we can think of $p_{T,j}$ as representing the complete fat jet input to the network. So unless explicitly mentioned we train our networks on a large data set defined in terms of the fat jet momentum,

$$p_{T,j} = 400 \ldots 1000 \text{ GeV} \qquad \text{(training sample)} \,, \tag{3.4}$$

Whenever we need a homogeneous sample without boundary effects we choose a narrow test sample with

$$p_{T,j} = 600 \ldots 620 \text{ GeV} \qquad \text{(narrow test sample)} \tag{3.5}$$

The data format for the fat jet information is a p_T-ordered list of up to 200 constituent 4-vectors ($\vec{p}$ and E) with ISR and 100 constituents without. Our total sample size is 2.2M jets without ISR, of which we use 400k jet for validation and testing, each. The training size is varied throughout our analysis.

Our regression network is a simple 5-layer fully connected dense network. Its first two layers each consist of 100 units, the next two 50 units, followed by a 2-unit output layer, unless mentioned otherwise. For the prior we choose a Gaussian around zero and with width 0.1. We have confirmed that our results are width-independent over a wide range [60]. The typical sizes and widths of the weights depends on the input data. The input is a flattened set of 4-vectors where we re-scale the p_T values by a factor 1000 to end up between zero and one. The activation function is ReLU, except for the output layer. That one predicts the mean value $\langle p_T \rangle$ without any need for an activation function and the SoftPlus function for the error to have a smooth function which guarantees positive

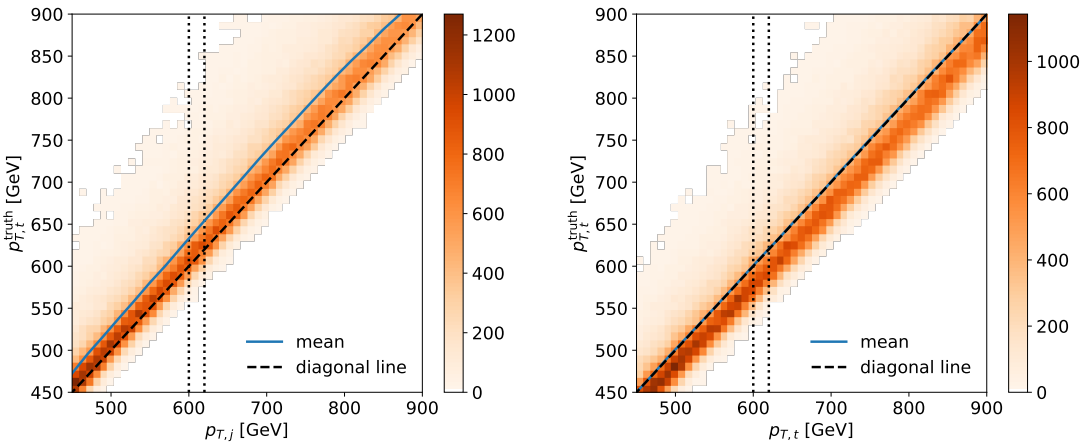


Figure 3.2: Correlation between the fat jet's $p_{T,j}$ and the truth label $p_{T,t}^{\text{truth}}$ (left) and between the extracted $p_{T,t}$ and the truth label $p_{T,t}^{\text{truth}}$ (right). Both correlations are shown with initial state radiation in the training and test samples switched off.

values for the error. We have checked that this setup with these hyper-parameters is not fine-tuned.

For the Bayesian network features we rely on Tensorflow Probability [112] with Flipout Dense layers [113] replacing the dense layer of the deterministic network. All networks are trained with the Adam optimizer [114] and a learning rate of 10^{-4}, determined by early stopping when the loss function evaluated on the training dataset does not improve for a certain number of epochs. This patience was set to 10 for a training size of 1M jets and to larger values for smaller training sizes because the loss function is more fluctuating. For the Bayesian network with a training batch size of 100 we observe no over-fitting.

3.4 Momentum determination and statistics

As a first part of our Bayesian regression analysis we need to show how well the networks reconstructs the top momentum and what the limiting factors are. We then have to separate the statistical and systematic uncertainties. In analogy to Ref. [60] we first study how the size of the training sample affects the regression output, i.e. how well the Bayesian network keeps track of the statistical uncertainty.

To illustrate the output of our Bayesian network for a single jet we show an example in Fig. 3.1. Sampling from the weight distributions, $q(\omega)$, provides us with a Gaussian per sampled set of weights, shown in petrol. The combination of these distributions is shown in red. The width of the combined distribution is the predicted per-jet uncertainty σ_{tot}, defined in Eq. (2.20). For illustration purposes we pick a top jet where $p_{T,t}^{\text{truth}}$ coincides with the peak of the predicted distribution.

Regression performance

To begin, we show in the left panel of Fig. 3.2 the correlation between the measurable $p_{T,j}$ and the MC label $p_{T,t}^{\text{truth}}$. We see that over the entire range the two values are aligned well. This allows us to use $p_{T,j}$ as a proxy to the truth information, keeping in mind that

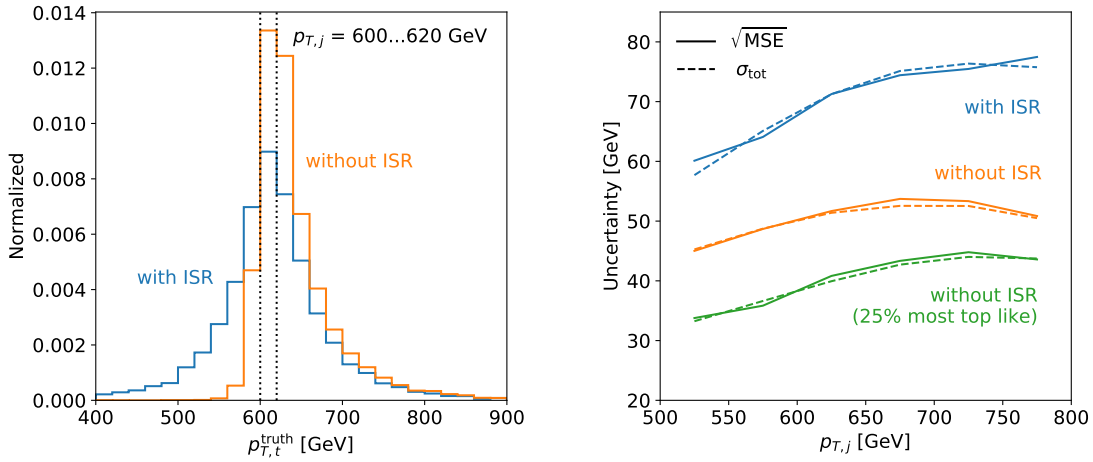


Figure 3.3: Left: distribution of the truth label $p_{T,t}^{\text{truth}}$ for jets with $p_{T,j} = 600 \ldots 620$ GeV, without and with initial state radiation. Right: regression uncertainty as a function of $p_{T,j}$ (solid), compared with the average σ_{stoch} as the network output (dashed). The most top-like events are defined with a simple LoLa tagger [86].

we will eventually smear the truth label to describe the jet calibration. In the right panel of Fig. 3.2 we show the correlation between the central extracted $p_{T,t}$ value, which in Sec. 3.2 is properly denoted as the expectation value $\langle p_T \rangle$, and the label $p_{T,t}^{\text{truth}}$.

In the left panel of Fig. 3.3 we show the $p_{T,t}^{\text{truth}}$ distribution for the narrow slice $p_{T,j} = 600 \ldots 620$ GeV. In the absence of initial state radiation the distribution is asymmetric. The simple reason is that the jet clustering can only miss top decay constituents, so we are more likely to observe $p_{T,j} < p_{T,t}^{\text{truth}}$. Aside from that we see a clear peak, suggesting that we can indeed represent $p_{T,t}^{\text{truth}}$ with $p_{T,j}$. Because the peak is washed out by ISR, we switch off ISR to make it easier to understand the physics behind our network task. In practice, this could be done through a pre-processing and grooming step.

Whenever we have access to MC truth, we can measure the performance of the regression network for each top jet as $(p_{T,t} - p_{T,t}^{\text{truth}})^2$. The squared difference measure only uses the mean or central value reported by a Bayesian or deterministic network, not the additional uncertainty information from the Bayesian network. For a given test sample with N top

$p_{T,j} = 600 \ldots 620$ GeV	$\sqrt{\text{MSE}}$	$\sqrt{\text{MSE}}/p_{T,j}$	$\sqrt{\text{MSE}}$	$\sqrt{\text{MSE}}/p_{T,j}$
	With ISR		Without ISR	
All jets	69.7 ± 0.2	$(11.43 \pm 0.03)\%$	50.6 ± 0.1	$(8.30 \pm 0.02)\%$
75% most top-like	67.8 ± 0.2	$(11.11 \pm 0.01)\%$	45.5 ± 0.1	$(7.47 \pm 0.02)\%$
50% most top-like	66.5 ± 0.1	$(10.89 \pm 0.01)\%$	41.8 ± 0.1	$(6.85 \pm 0.01)\%$
25% most top-like	66.5 ± 0.1	$(10.89 \pm 0.02)\%$	40.4 ± 0.1	$(6.63 \pm 0.02)\%$

Table 3.1: Performance of $p_{T,t}$ regression, uncertainty representing the standard deviation of 5 trainings. The narrow $p_{T,j}$ range refers to the 5k test jets, not the 500k training jets.

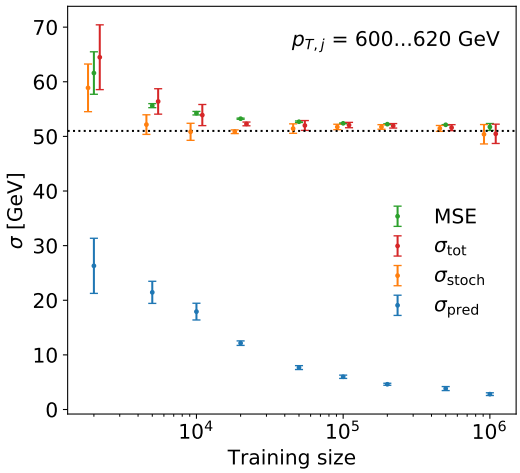


Figure 3.4: Uncertainty contributions σ_{pred} and σ_{stoch} as a function of the size of the training sample. The error bar represents the standard deviation of five different trainings. In addition we include $\sqrt{\text{MSE}}$ as defined in Eq. (3.6).

jets t_i we construct the mean quadratic error as

$$\sqrt{\text{MSE}} = \left[\frac{1}{N} \sum_{\text{jets } i} \left(p_{T,t_i} - p_{T,t_i}^{\text{truth}} \right)^2 \right]^{1/2} \tag{3.6}$$

We evaluate it over homogeneous samples, for example our usual slice in $p_{T,j}$. In Tab. 3.1 we contrast results with and without ISR and show what happens if we limit ourselves to the most top-like jets based on a standard LoLa tagger [86], trained on events with ISR. To estimate the effect of different trainings we also give an error bar based on five independent trainings and the resulting standard deviation. Expectedly, the p_T-measurement benefits from more top-like events, but the effect is not as significant as in the HEPTopTagger analysis [109]. One of the reasons is that we are using relatively large $R = 1.2$ jets for the high transverse momentum range. Similarly, we confirm that additional ISR jets have the potential to affect the top momentum measurement whenever hard extra jets enter the fat jet area.

In the right panel of Fig. 3.3 we show $\sqrt{\text{MSE}}$ as a function of $p_{T,j}$ for a bin width of 40 GeV. While the absolute error increases, the relative error on the extracted $p_{T,t}$ shrinks for more boosted jets. If we assume that an improved jet pre-selection can efficiently remove ISR contributions our regression network can measure the top momentum to roughly 4%. This result is only a rough benchmark to confirm that the regression network performs in a meaningful manner. It would surely be possible to improve the network performance, but we deliberately keep the network simple, to understand the way it processes information and the related uncertainties. From the right panel of Fig. 3.3 we know that boundary effects will appear already around 200 GeV away from the actual boundaries. Indeed, around $p_{T,j}$ we see such effects indicating the phase space boundary of $p_{T,j} < 1$ TeV in our training sample.

In the same Fig. 3.3 we also show this uncertainty estimate of the Bayesian network, σ_{tot} as defined in Eq. (2.20). It follows the $\sqrt{\text{MSE}}$ estimate of the network error, indicating that the Bayesian output captures the same physics as the frequentist-defined spread of the central values.

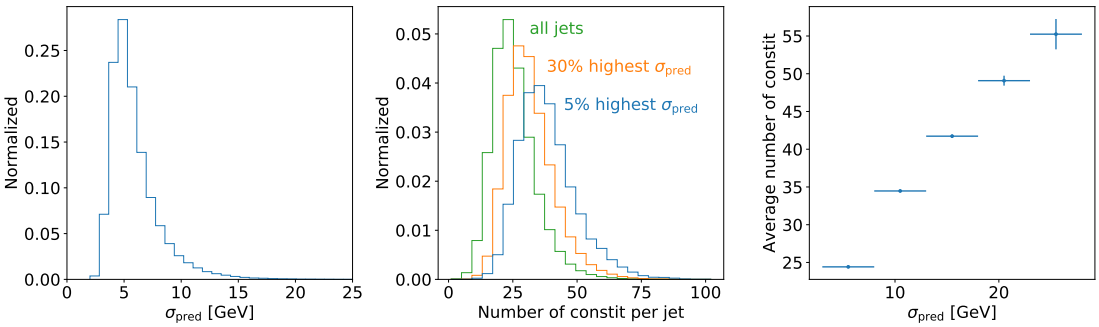


Figure 3.5: Left: distribution of the statistical uncertainty σ_{pred} for 400k jets. Center: number of constituents per jet for different σ_{pred}. Right: average number of constituents per jet as a function of the extracted statistical uncertainty.

Training sample size and σ_{pred}

As discussed in Sec. 3.2 the contribution σ_{pred} to the uncertainty reported by the network can be identified as a statistical uncertainty in the sense that it should vanish in the limit of infinitely many training jets. In complete analogy to the classification task described in Ref. [60] we confirm this by training Bayesian networks on 2k, 5k, 10k, 15k, 20k, 30k, 50k, 100k, 200k, 500k, and 1M jets. We test these networks on the narrow range $p_{T,j} = 600 \ldots 620$ GeV, similar to the results shown in Tab. 3.1. The uncertainties quoted by the Bayesian network are shown in Fig. 3.4. In the lower part of the figure we first see that the statistical error σ_{pred} indeed asymptotically approaches zero for 1M training jets. The error bars on the extracted uncertainty are given by the standard deviation of five independent trainings. As expected, they grow for smaller training samples, where the Bayesian networks also give fluctuating results.

In the same figure we also show the systematic σ_{stoch} and the combined σ_{tot}, defined in Eq. (2.20). We confirm that the extracted σ_{stoch} hardly depends on the size of the training sample. Once we have a reasonably number of training events it reaches a plateau of around 50 GeV or 8%, while for less than 10000 training events the network simply fails to capture the full information. We can compare the plateau value for σ_{stoch} to the $\sqrt{\text{MSE}}$ value and find again that the two values agree.This allows us to conclude that σ_{stoch} describes a systematic uncertainty and that it is related to the truth-based $\sqrt{\text{MSE}}$ estimate. We will discuss it in more detail in Sec. 3.5.

After observing the average effect of the training sample size on σ_{pred} the obvious question is if we can understand this behavior. In the left panel of Fig. 3.5 we show the distribution of σ_{pred} values for a sample of 400k jets. The network is trained on 100k jets with an extended range $p_{T,j} = 500 \ldots 900$ GeV. We see a clear maximum around $\sigma_{\text{pred}} \approx 5$ GeV, with a large tail towards large uncertainties. It is induced by the constraint that no network should quote an uncertainty close to zero.

The jet property we can relate to the σ_{pred} behavior is the number of particle-flow constituents. As mentioned before, we cover up to 100 constituents for jets without ISR. Their effect on top tagging is discussed for instance in Ref. [86]. The center panel of Fig. 3.5 shows how the number of constituents in the test sample jets peaks at around 25, but with a tail extending to 60. Jets with a larger quoted uncertainty have significantly more constituents. The same information is shown in the right panel, where we see the average number of jets increases with the range of quoted statistical uncertainties. The

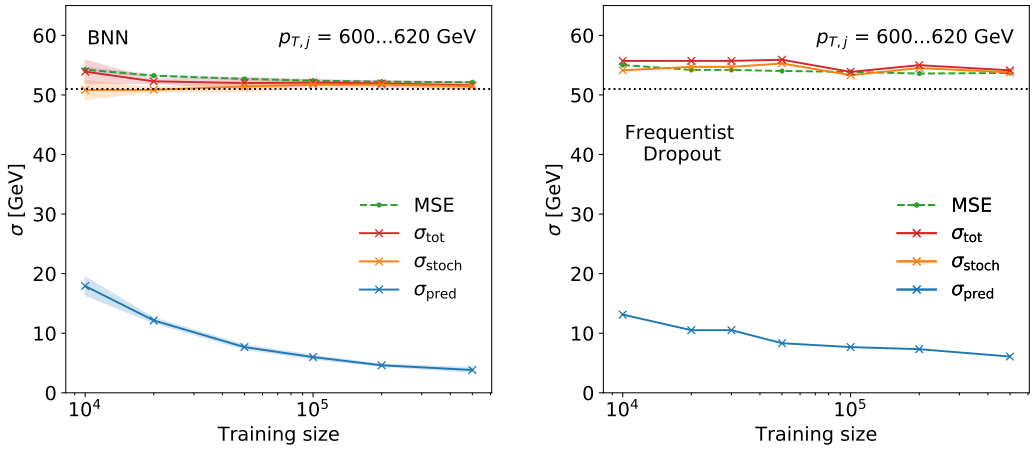


Figure 3.6: All uncertainties as a function of the training size, comparing the Bayesian network (left) with a (frequentist) set of deterministic networks (right). The left panel corresponds to Fig. 3.4, and the ranges indicate the standard deviation for five trainings.

reason for this pattern is that also within the training sample the number of constituents will peak around 25, limiting the number of training jets with higher constituent numbers. We note that we could use the same argument using the jet mass.

Frequentist approach

From a practical point of view it is crucial to validate the Bayesian network using a frequentist approach. We do this by showing that predictions from many trainings of a deterministic network reproduce our Bayesian network results for the statistical uncertainty σ_{pred}.

For the deterministic networks we use the same architecture as for the Bayesian network. The loss function of the deterministic networks is the negative log-likelihood given in Eq. (2.15), and we fix the L2-regularization to match the Bayesian network in Eq. (2.12),

$$\lambda_{\text{L2}} = \frac{1}{2\sigma_{\text{prior}} N} \,, \tag{3.7}$$

where N is the total training size and $\sigma_{\text{prior}} = 0.1$ is our prior width. We then train 40 deterministic networks on statistically independent samples, which we sample from the total of 2.2M training jets. Each set of deterministic network then predicts a mean and a standard deviation, in analogy to Eq. (3.3). The difference between the Bayesian evaluation and the frequentist networks is that we replace the integral over weights with a sum over independent networks.

For deterministic networks we need to avoid over-training. An over-trained set of networks will underestimate σ_{stoch}, while the spread represented by σ_{pred} increases. However, it is not guaranteed that these two effects compensate each other for finite training time. This is why we introduce dropout for each inner layer with a rate of 0.1. This value is a compromise between network performance and over-training. Unlike in our earlier study [60] we do not use a MAP modification of the Bayesian network.

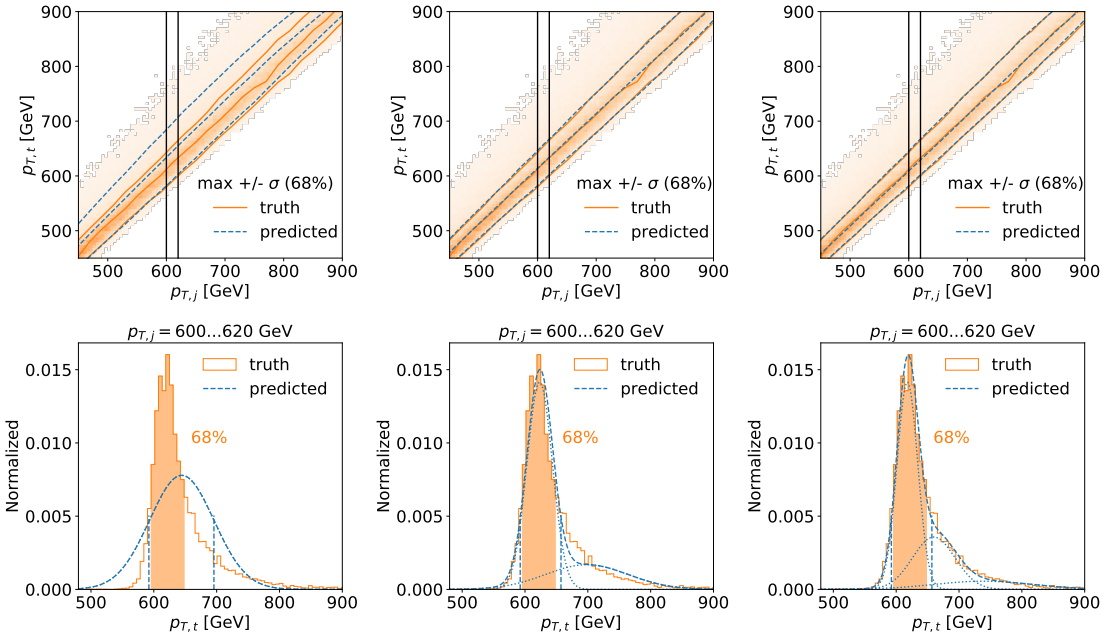


Figure 3.7: Upper: 2-dimensional distribution of $p_{T,t}^{\text{truth}}$ vs $p_{T,j}$ including its 68% CL around the maximum. In blue we show the BNN results. Lower: $p_{T,t}^{\text{truth}}$-distribution for a narrow slice in $p_{T,j}$. From left to right we approximate $p_{T,t}^{\text{truth}}$ with one, two, and three Gaussians.

In Fig. 3.6 we compare the Bayesian and frequentist uncertainties for different training sample size. While the results agree well for properly trained networks or large training samples, the frequentist approach slightly underestimates the uncertainty for small training samples. The plateau value of σ_{stoch} depends on the chosen dropout value. Accounting for this effect we see that the training-size-dependent σ_{pred} and the plateau value of σ_{stoch}, agree well between the Bayesian network and the frequentist sanity check.

3.5 Systematics and calibration

In our original paper [60] we have shown that the Bayesian setup propagates uncertainties from statistical and systematic limitations of the training data through a neural network. In addition to the usual output the Bayesian network provides event-by-event error bars. A limitation we encounter in Ref. [60] is that forcing the network output onto a closed interval, like a probability $p \in [0, 1]$, strongly correlates the the central value and the error bars in the network output. This makes it difficult to track systematic uncertainties.

We circumvent this problem by extracting the transverse momentum, which does not live on a closed interval. In the previous section this allowed us to decompose σ_{tot} into a statistical component, σ_{pred}, and a systematic component, σ_{stoch}. What we still need to study is the actual output distribution of the Bayesian network, $p(p_T|M)$, and how it compared to the truth information from the test data.

	$\alpha^{(i)}$	$\langle p_T\rangle^{(i)}$	$\sigma^{(i)}_{\text{stoch}}$	σ_{stoch}	$\sqrt{\text{MSE}}$	$\langle p_{T,t}\rangle$	$\langle p^{\text{truth}}_{T,t}\rangle$	Max	68%CL	68%CL (truth)
1	1	644.4	51.43	51.4		644.4		644.4	593.0...695.9	
2	0.72	623.4	20.4	51.1		644.1		623.4	592.4...657.3	
	0.28	698.3	65.6							
3	0.59	617.8	16.6	51.5	52.2	643.8	643.8	619.1	592.4...656.8	590.0...654.0
	0.30	659.8	33.7							
	0.11	738.6	78.6							

Table 3.2: Parameters used in Fig. 3.7, specifically $p_{T,j} = 600...620$ GeV.

Variance of training data and σ_{stoch}

In the upper left panel of Fig. 3.7 we show the correlation of $p^{\text{truth}}_{T,t}$ and $p_{T,j}$. The orange curves represent the maximum and the 68% CL interval in 20 GeV bin. The corresponding maximum and 68% CL interval of the BNN output are illustrated in blue. Both confidence intervals are constructed by requiring equal functional values at both ends. In the lower left panel we see why the two sets of curves agree very poorly: for the narrow $p_{T,j}$ slide the $p^{\text{truth}}_{T,t}$ distribution is all but Gaussian, while the Bayesian output in our naive approach is forced to be Gaussian.

From Sec. 3.2 we know that it is not necessary to assume that the Bayesian network output is Gaussian. As a simple generalization we can replace the two-parameter Gaussian form of $p(p_T|\omega, j)$ (compare $p(D|\omega)$ in Eq. (2.15)) with a mixture of Gaussians,

$$p(p_T|\omega, j) = \sum_i \alpha_{i,\omega}\, G(\langle p_T\rangle^{(i)}_\omega, \sigma^{(i)}_{\text{stoch},\omega})\,, \tag{3.8}$$

with $\sum_i \alpha_{i,\omega} = 1$. The network output from Eq. (3.3) then becomes

$$\text{NN}(\omega) = \begin{pmatrix} \alpha_{1,\omega} & \alpha_{2,\omega} & \cdots \\ \langle p_T\rangle^{(1)}_\omega & \langle p_T\rangle^{(1)}_\omega & \cdots \\ \sigma^{(1)}_{\text{stoch},\omega} & \sigma^{(2)}_{\text{stoch},\omega} & \cdots \end{pmatrix} \tag{3.9}$$

To guarantee $\sum_i \alpha_{i,\omega} = 1$ we use SoftMax as an activation function for $\alpha_{i,\omega}$ and the SoftPlus function for $\sigma^{(i)}_{\text{stoch},\omega}$ to ensure positive values. In the center and right sets of panels in Fig. 3.7 we see what happens if we use two or three Gaussians, specifically with the parameters averaged over weights and jets in a bin. For three Gaussians the BNN output and the $p^{\text{truth}}_{T,t}$ distribution agree perfectly. The corresponding parameters are shown in Tab. 3.2.

Technically, we follow Sec. 2.2.2 in extracting σ_{stoch} and σ_{pred} independently of the form of the underlying assumption. Two aspects render this computation slightly expensive: the integration over all weights and, if required, the combination of different predictions in one $p_{T,j}$ bin. On the other hand we know that $\sigma_{\text{pred}} \ll \sigma_{\text{stoch}}$ and we can always use narrow bin sizes. This means that in both cases we can replace the integrals by simply averaging over the parameters of the Gaussian mixture model. This implementation is computationally less expensive and gives us simple analytic expressions from which we extract the maximum and 68% CL interval.

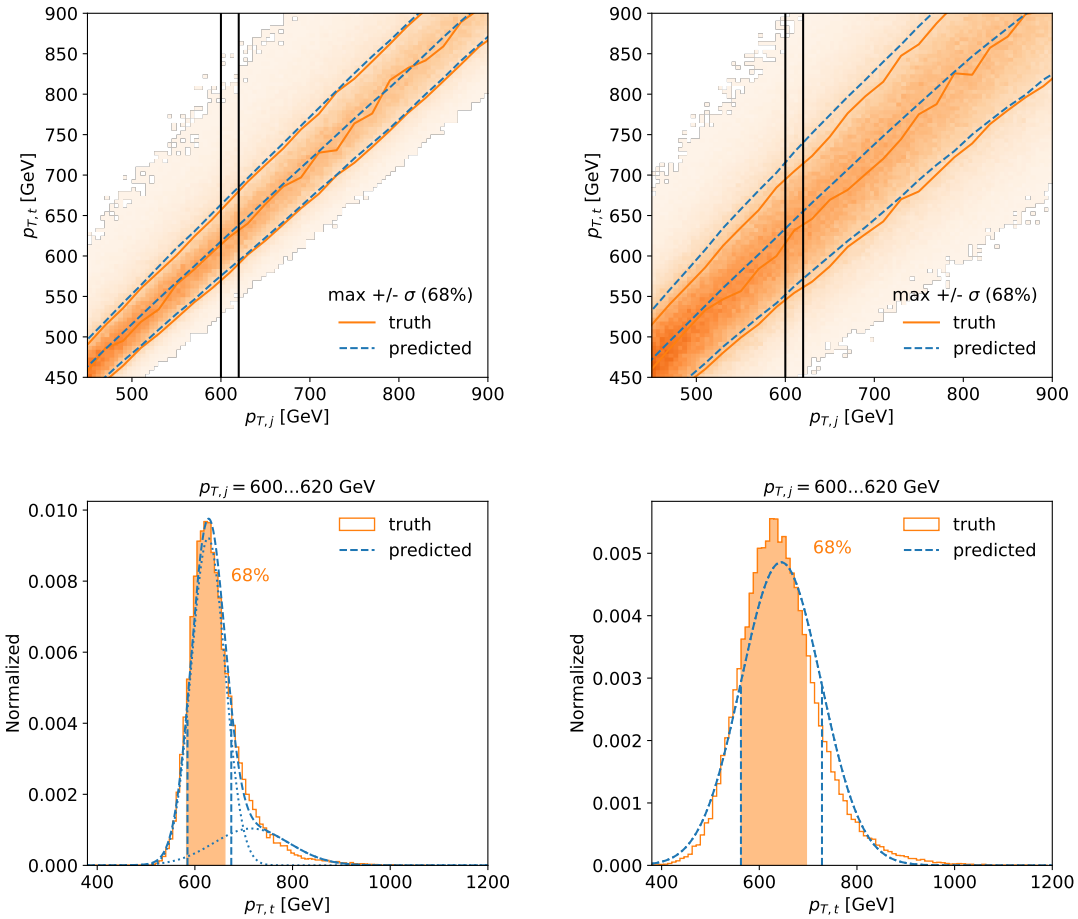


Figure 3.8: Upper: 2-dimensional distribution of $p_{T,t}^{\text{truth}}$ vs $p_{T,j}$ including its 68% CL around the maximum, after adding 4% (left) and 10% (right) Gaussian noise on the top momentum label. In blue we also show the BNN error estimate. Lower: corresponding $p_{T,t}^{\text{truth}}$-distribution for a narrow slice in $p_{T,j}$.

Noisy labels

A crucial question in experimental physics is how we include a systematic uncertainty for instance on the jet energy scale in the training procedure. We can understand such an energy calibration when we remind ourselves that the jets in the calibration sample come with a measured reference value for their energies and the corresponding error bar; and that the calibration sample in our case is the training sample. There are two ways we can include the error on the calibration measurements in our analysis:

1A. fix the label or "true energy" and smear the jets in the training sample;

1B. fix the jets and smear the continuous label in the training sample;

2. train the Bayesian network on the smeared label-jet combination;

3. extract a systematics error bar for each jet in the test sample.

In Ref. [60] we have followed the option 1A and encountered some practical/numerical problems when tracing the corresponding systematics to the network output. For this approach we refer to Appendix A.1. In this study we shift to the less standard and

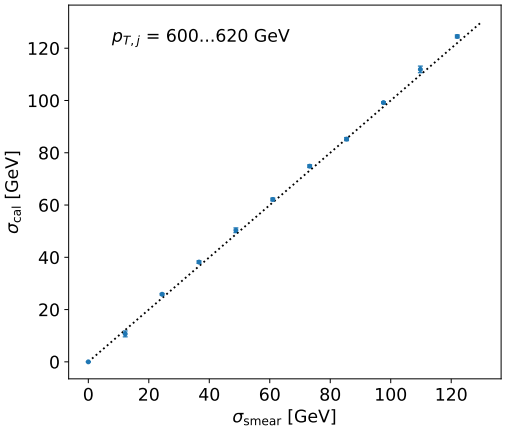


Figure 3.9: Correlation between σ_{stoch}, as given by the Bayesian network and the smearing σ_{smear} applied to the label in the training data. The baseline $\sigma_{\text{stoch},0}$ is defined as σ_{stoch} in the limit of no smearing. The error bars indicate the standard deviation from five independent trainings.

yet straightforward option 1B. We assume that jet calibration incorporates external information on the training sample, be it another measurement or a theory requirement (one-shell Z-decays) or a MC prediction. This information defines a label together with a corresponding error bar. This means we train our network on a fixed sample of jets with a smeared label representing the full reference measurement. In this approach we can trivially include additional uncertainties from pre-processing the training data, like running a jet algorithm of the Z-sample, removing underlying event and pile-up, etc. As a side effect our setup also allows us to capture possible transfer uncertainties, whenever our test sample cannot easily be linked to the training sample. In the ML literature such uncertainties are referred to as out-of-sample error.

To illustrate and test our setup we smear $p_{T,t}^{\text{truth}}$, the label in the training data, according to Gaussians with widths of

$$\sigma_{\text{smear}} = (4 \ldots 10)\% \times p_{T,t}^{\text{truth}} \,. \tag{3.10}$$

In Fig. 3.8. we see that for a small amount of smearing the non-Gaussian shape of Fig. 3.7 remains, so we use two Gaussians in the BNN. For sizable Gaussian smearing we see that the resulting distributions all assume a Gaussian shape and we can stick to the single-Gauss standard BNN. In both cases the distribution of the BNN output and the (smeared) label $p_{T,t}^{\text{truth}}$ agree almost perfectly.

From the previous sections we know that the reported uncertainty by the BNN includes a statistical uncertainty vanishing with an increasing amount of training data and a systematic uncertainty representing the stochastic nature of the training data. When we introduce another uncertainty induced by smeared labels we expand Eq. (2.20) to

$$\begin{aligned} \sigma_{\text{tot}}^2 &= \sigma_{\text{stoch}}^2 + \sigma_{\text{pred}}^2 \\ &= \sigma_{\text{stoch},0}^2 + \sigma_{\text{cal}}^2 + \sigma_{\text{pred}}^2 \quad \Leftrightarrow \quad \sigma_{\text{cal}}^2 = \sigma_{\text{stoch}}^2 - \sigma_{\text{stoch},0}^2 \,, \end{aligned} \tag{3.11}$$

added in quadrature because of the central limit theorem. The baseline value $\sigma_{\text{stoch},0}$ is defined as σ_{stoch} in the limit of no smearing. In Fig. 3.9 we show how σ_{cal} correlates with the input σ_{smear} over a wide range of scale uncertainties. As usually, the error bar

represents the standard deviation from five independent trainings. This correlation shows that our network picks up the systematic uncertainties from smeared training labels perfectly. We note that, as before, this analysis does not require a Gaussian shape of the network output.

3.6 Outlook

We have shown that Bayesian networks keep track of statistical and systematic uncertainties in the training data and translate them into a jet-by-jet error budget for instance in a momentum measurement. Outside particle physics it is not unusual to treat uncertainties as a smearing of labels, whereas in particle physics we usually model them by smearing the input data. We show that smearing labels is a natural, feasible, and self-consistent strategy in combination with deep learning. An advantage of this approach is that the treatment of uncertainties is moved from the evaluation time to the training time and so-trained networks accurately report predictions of the central value as well as systematic uncertainties.

We have shown that the corresponding Bayesian networks allow us to cleanly separate statistical and systematic uncertainties. In addition, the smeared labels are ideally suited to translate uncertainties from reference or calibration data to the network output.

Technically, we have modified the Bayesian network approach of Ref. [60] to include non-Gaussian behavior. This step is crucial for modeling systematic uncertainties in general.

We emphasize that before this approach can be generally adapted, open questions such as multiple correlated uncertainties and the translation between input-uncertainties and label-uncertainties need to be answered. However, our first results show great promise for smeared labels describing uncertainties in particle physics applications of deep learning.

Chapter 4

Event generation with Bayesian neural networks

The research presented in this chapter has been published in Ref. [2]. Most of the figures, tables and text are completely identical to the content of this publication.

4.1 Introduction

The role of first-principle simulations in our understanding of large data sets makes LHC physics stand out in comparison to many other areas of science. Three aspects define the application of modern big data methods in this field:

- ATLAS and CMS deliver proper big data with excellent control over uncertainties;
- perturbative quantum field theory provides consistent precision predictions;
- fast and reliable precision simulations generate events from first principles.

The fact that experiments, field theory calculations, and simulations control their uncertainties implies that we can work with a complete uncertainty budget, including statistical, systematic, and theory uncertainties. To sustain this approach at the upcoming HL-LHC, with a data set more than 25 times the current Run 2 data set, the theory challenge is to provide faster simulations and keep full control of the uncertainties at the per-cent level and better.

In recent years it has been shown that modern machine learning can improve LHC event simulations in many ways [34]. Promising techniques include generative adversarial networks (GANs) [71, 72, 115], variational autoencoders [62, 70], and normalizing flows [66, 69, 116, 117], including invertible neural networks (INNs) [67, 68, 118]. They can improve phase space integration [119, 120], phase space sampling [121–123], and amplitude computations [108, 124]. Further developments are fully NN-based event generation [125–129], event subtraction [130], event unweighting [131, 132], detector simulation [133–142], or parton showering [143, 144, 144–146]. Generative models will also improve searches for physics beyond the Standard Model [147], anomaly detection [148, 149], detector resolution [150, 151], and inference [152–154]. Finally, conditional GANs and INNs allow us to invert the simulation chain to unfold detector effects [33, 155] and extract the hard scattering process at parton level [156]. The problem with these applications is that we know little about

1. how these generative networks work, and

2. what the uncertainty on the generative network output is.

As we will see in this paper, these two questions are closely related.

In general, we can track statistical and systematic uncertainties in neural network outputs with Bayesian networks [36,37,39,59]. Such networks have been used in particle physics for a long time [157–159]. For the LHC we have proposed to use them to extract uncertainties in jet classification [60] and jet calibration [1]. They can cover essentially all uncertainties related to statistical, systematic, and structural limitations of the training sample [93]. Similar ideas can be used as part of ensemble techniques [160]. We propose to use a Bayesian invertible neural network (BINN) to extract uncertainties on a generated event sample induced by the network training.

Because Bayesian networks learn the density and uncertainty maps in one pass, their relation offers us fundamental insight into the way an INN learns a distribution. While Bayesian classification [60] and regression networks [1] highlight the statistical and systematic nature of uncertainties, our Bayesian generative network exhibits a very different structure. We will discuss the learning pattern of the Bayesian INN in details for a set of simple toy processes in Sec. 4.3, before we apply the network to a semi-realistic LHC example in Sec. 4.4.

4.2 Generative networks with uncertainties

We start by reminding ourselves that we often assume that a generative model has learned a phase space density perfectly, so the only remaining source of uncertainty is the statistics of the generated sample binned in phase space. However, we know that such an assumption is not realistic [1, 60], and we need to estimate the effect of statistical or systematic limitations of the training data. The problem with such a statistical limitation is that it is turned into a systematic shortcoming of the generative model [128] — once we generate a new sample, the information on the training data is lost, and the only way we might recover it is by training many networks and comparing their outcome. For most applications this is not a realistic or economic option, so we will show how an alternative solution could look.

4.2.1 Uncertainties on event samples

Uncertainties on a simulated kinematic or phase space distribution are crucial for any LHC analysis. For instance, we need to know to what degree we can trust a simulated p_T-distribution in mono-jet searches for dark matter. We denote the complete phase space weight for a given phase space point as $p(x)$, such that we can illustrate a total cross section as

$$\sigma_\text{tot} = \int_0^1 dx\, p(x) \quad \text{with} \quad p(x) > 0 \,. \tag{4.1}$$

In this simplified notation x stands for a generally multi-dimensional phase space. For each phase space position, we can also define an uncertainty $\sigma(x)$.

Two contributions to the error budget are theory and systematic uncertainties, $\sigma_\text{th/sys}(x)$. The former reflects our ignorance of aspects of the training data, which do not decrease when we increase the amount of training data. The latter captures the degree to which

we trust our prediction, for instance based on self-consistency arguments. For example, we can account for possible large, momentum-dependent logarithms as a simple function of phase space. If we use a numerical variation of the factorization and renormalization scales to estimate a theory uncertainty, we typically re-weight events with the scales. Another uncertainty arises from the statistical limitations of the training data, $\sigma_{\text{stat}}(x)$. For instance in mono-jet production, the tails of the predicted p_T-distribution for the Standard Model will at some point be statistics limited. In the Gaussian limit, a statistical uncertainty can be defined by binning the phase space and in that limit we expect a scaling like $\sigma_{\text{stat}}(x) \sim \sqrt{p(x)}$, and we will test that hypothesis in detail in Sec. 4.3.

Once we know the uncertainties as a function of the phase space position, we can account for them as additional entries in unweighted or weighted events. For instance, relative uncertainties can be easily added to unweighted events,

$$\text{ev}_i = \begin{pmatrix} \sigma_{\text{stat}}/p \\ \sigma_{\text{syst}}/p \\ \sigma_{\text{th}}/p \\ \{x_{\mu,j}\} \\ \{p_{\mu,j}\} \end{pmatrix}, \qquad \text{with } \mu = 0 \ldots 3 \text{ for each particle } j. \tag{4.2}$$

The entries σ or σ/p are smooth functions of phase space. The challenge in working with this definition is how to extract σ_{stat} without binning. We will show how Bayesian networks give us access to limited information in the training data. Specific theory and systematics counterparts can be either computed directly or extracted by appropriately modifying the training data [1, 60].

4.2.2 Invertible Neural Networks

To model complex densities such as LHC phase space distributions, we can employ normalizing flows [66–69]. As discussed in Sec. 2.3 their advantage is that they are likelihood-based generative models which allow an easy formulation in terms of Bayesian neural networks.

The construction of normalizing flows relies on the property that the composition of a chain of simple invertible nonlinear maps gives us a complex map. There exists a broad literature of different transformations, each with different strengths and weaknesses [69]. In this work, we rely on the real non-volume preserving flow [67] in the invertible neural network (INN) formulation [118].

The invertible neural net provides us with a powerful generative model of the underlying data distribution. However, it lacks a mechanism to account for our uncertainty in the transformation parameters θ themselves. To model it, we switch from deterministic transformations to probabilistic transformations, replacing the deterministic sub-networks in each of the coupling layers with Bayesian neural nets. The corresponding Bayesian loss function is discussed in more detail in Sec. 2.3.1.

4.3 Toy events with uncertainties

Before we tackle a semi-realistic LHC setup, we first study the behavior of BINNs for a set of toy examples, namely distributions over the minimally allowed two-dimensional parameter space where in one dimension the density is flat. Aside from the fact that these toy examples illustrate that the BINN actually constructs a meaningful uncertainty distribution, we will use the combination of density and uncertainty maps to analyze how an INN actually learns a density distributions. We will see that the INN describes the density map in the sense of a few-parameter fit, rather than numerically encoding patches over the parameter space independently.

The default architecture for our toy models is a network with 32 units per layer, three layers per coupling block, and a total of 20 coupling blocks. It's implemented in PyTorch [64]. More details are given in Tab. 4.1. The most relevant hyperparameter is the number of coupling blocks in that more blocks provide a more stable performance with respect to several trainings of the same architecture. Generally, moderate changes for instance of the number of units per layer do not have a visible impact on the performance. For each of the trainings we use a sample of 300k events. The widths of the Gaussian priors is set to one. We check that variations of this over several orders of magnitude did not have a significant impact on the performance.

4.3.1 Wedge ramp

Our first toy example is a two-dimensional ramp distribution, linear in one direction and flat in the other,

$$p(x,y) = \text{Linear}(x \in [0,1]) \times \text{Const}(y \in [0,1]) = x \times 2 \,. \tag{4.3}$$

The second term ensures that the distribution $p(x,y)$ is normalized to one, and the network output is shown in Fig. 4.1. The network output consists of unweighted events in the two-dimensional parameters space, (x,y). We show one-dimensional distributions

Parameter	Flow
Hidden layers (per block)	3
Units per hidden layer	32
Batch size	512
Epochs	300
Trainable weights	75k
Optimizer	Adam
(α, β_1, β_2)	(1×10^{-3}, 0.9, 0.999)
Coupling layers	20
Training size	300k
Prior width	1

Table 4.1: Hyper-parameters for all toy models, implemented in `Pytorch` (v1.4.0) [64].

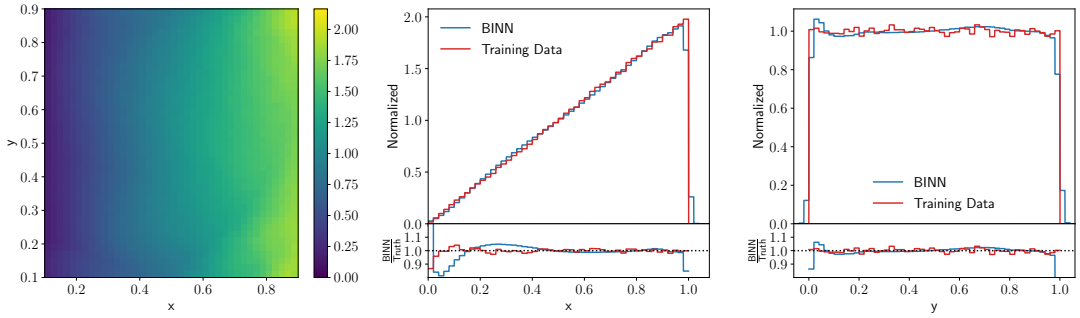


Figure 4.1: Two-dimensional and marginal densities for the linear wedge ramp.

after marginalizing over the unobserved direction and find that the network reproduces Eq. (4.3) well.

In Fig. 4.2 we include the predictive uncertainty given by the BINN. For this purpose we train a network on the two-dimensional parameter space and evaluate it for a set of points with $x \in [0, 1]$ and a constant y-value. In the left panel we indicate the predictive uncertainty as an error bar around the density estimate. Throughout the paper we always remove the phase space boundaries, because we know that the network is unstable there, and the uncertainties explode just like we expect. For this example, this is taken into account by restricting $x, y \in [0.1, 0.9]$. The relative uncertainty grows for small values of x and hence small values of $p(x, y)$, and it covers the deviation of the extracted density from the true density well. These features are common to all our network trainings. In the central and right panel of Fig. 4.2 we show the relative and absolute predictive uncertainties. The error bar indicates how much σ_{pred} varies for different choices of y. We compute it as the standard deviation of different values of σ_{pred}, after confirming that the central values agree within this range. As expected, the relative uncertainty decreases towards larger x. However, the absolute uncertainty shows a distinctive minimum in σ_{pred} around $x \approx 0.45$. This minimum is a common feature in all our trainings, so we need to explain it.

To understand this non-trivial uncertainty distribution $\sigma_{\text{pred}}(x)$ we focus on the non-trivial x-coordinate and its linear behavior

$$p(x) = ax + b \qquad \text{with} \qquad x \in [0, 1] \; . \tag{4.4}$$

Because the network learns a density, we can remove b by fixing the normalization,

$$p(x) = a \left(x - \frac{1}{2} \right) + 1 \; . \tag{4.5}$$

If we now assume that a network acts like a fit of a, we can relate the uncertainty Δa to an uncertainty in the density,

$$\sigma_{\text{pred}} \equiv \Delta p \approx \left| x - \frac{1}{2} \right| \; \Delta a \; . \tag{4.6}$$

The absolute value appears because the uncertainties are defined to be positive, as encoded in the usual quadratic error propagation. The uncertainty distribution has a minimum at $x = 1/2$, close to the observed value in Fig. 4.2.

The differences between the simple prediction in Eq. (4.6) and our numerical findings in

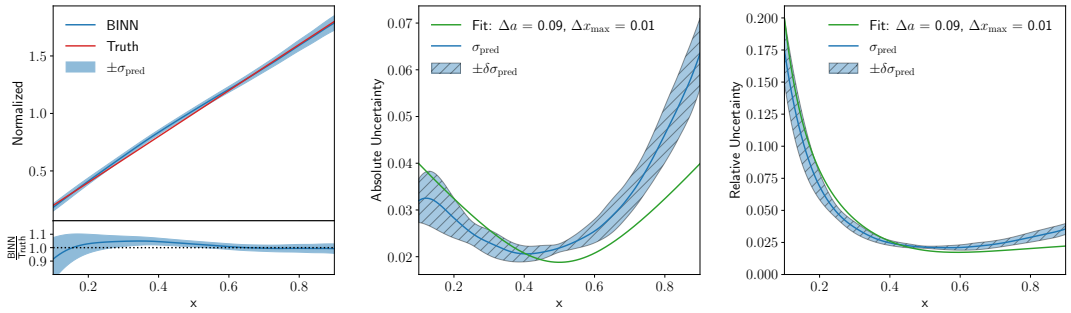


Figure 4.2: Density and predictive uncertainty distribution for the wedge ramp. In the left panel the density and uncertainty are averaged over several lines with constant y. In the central and right panels, the uncertainty band on σ_{pred} is given by their variation. The green curve represents a two-parameter fit to Eq. (4.9).

Fig. 4.2 is that the predictive uncertainty is not symmetric and does not reach zero. To account for these sub-leading effects we can expand our very simple ansatz to

$$p(x) = ax + b \qquad \text{with} \qquad x \in [x_{\text{min}}, x_{\text{max}}] \,. \tag{4.7}$$

Using the normalization condition we again remove b and find

$$p(x) = ax + \frac{1 - \frac{a}{2}(x_{\text{max}}^2 - x_{\text{min}}^2)}{x_{\text{max}} - x_{\text{min}}} \,. \tag{4.8}$$

Again assuming a fit-like behavior of the flow network we expect for the predictive uncertainty

$$\sigma_{\text{pred}}^2 \equiv (\Delta p)^2 = \left(x - \frac{1}{2}\right)^2 (\Delta a)^2 + \left(1 + \frac{a}{2}\right)^2 (\Delta x_{\text{max}})^2 + \left(1 - \frac{a}{2}\right)^2 (\Delta x_{\text{min}})^2 \,. \tag{4.9}$$

Adding x_{min} or x_{max} leads to an x-independent offset and does not change the x-dependence of the predictive uncertainty. The slight shift of the minimum and the asymmetry between the lower and upper boundaries in x are not explained by this argument. We ascribe them to boundary effects, specifically the challenge for the network to describe the correct approach towards $p(x) \to 0$.

The green line in Fig. 4.2 gives a two-parameter fit of Δa and Δx_{max} to the σ_{pred} distribution from the BINN. It indicates that there is a hierarchy in the way the network extracts the x-independent term with high precision, whereas the uncertainty on the slope a is around 4%.

4.3.2 Kicker ramp

We can test our findings from the linear wedge ramp using the slightly more complex quadratic or kicker ramp,

$$p(x, y) = \text{Quadr}(x \in [0, 1]) \times \text{Const}(y \in [0, 1]) = x^2 \times 3 \,. \tag{4.10}$$

We show the results from the network training for the density in Fig. 4.3 and find that the network describes the density well, limited largely by the flat, low-statistics approach

towards the lower boundary with $p(x) \to 0$.

In complete analogy to Fig. 4.2 we show the complete BINN output with the density $p(x,y)$ and the predictive uncertainty $\sigma_\text{pred}(x,y)$ in Fig. 4.4. As for the linear case, the BINN reproduces the density well, deviations from the truth being within the predictive uncertainty in all points of phase space. We remove the phase space boundaries restricting $x, y \in [0.05, 0.95]$, as the network becomes unstable and the predictive uncertainties grows correspondingly. The indicated error bar on $\sigma_\text{pred}(x,y)$ is given by the variation of the predictions for different y-values, after ensuring that their central values agree. The relative uncertainty at the lower boundary $x = 0$ is large, reflecting the statistical limitation of this phase space region. An interesting feature appears again in the absolute uncertainty, namely a maximum-minimum combination as a function of x.

Again in analogy to Eq. (4.7) for the wedge ramp, we start with the parametrization of the density

$$p(x) = a\,(x - x_0)^2 \qquad \text{with} \qquad x \in [x_0, x_\text{max}] \,, \tag{4.11}$$

where we assume that the lower boundary coincides with the minimum and there is no constant offset. We choose to describe this density through the minimum position x_0, coinciding the the lower end of the x-range, and x_max as the second parameter. The parameter a can be eliminated through the normalization condition and we find

$$p(x) = 3 \frac{(x - x_0)^2}{(x_\text{max} - x_0)^3} \,. \tag{4.12}$$

If we vary x_0 and x_max we can trace two contributions to the uncertainty in the density,

$$\begin{aligned} \sigma_\text{pred} \equiv \Delta p &\supset \frac{9}{(x_\text{max} - x_0)^4} \left| (x - x_0) \left(x - \frac{x_0}{3} - \frac{2x_\text{max}}{3} \right) \right| \Delta x_0 \\ \text{and} \qquad \sigma_\text{pred} \equiv \Delta p &\supset \frac{9}{(x_\text{max} - x_0)^4} \, (x - x_0)^2 \, \Delta x_\text{max} \,, \end{aligned} \tag{4.13}$$

one from the variation of x_0 and one from the variation of x_max. In analogy to Eq. (4.9) they need to be added in quadrature. If the uncertainty on Δx_0 dominates, the uncertainty has a trivial minimum at $x = 0$ and a non-trivial minimum at $x = 2/3$. From Δx_max we get another contribution which scales like $\Delta p \propto p(x)$. In Fig. 4.4 we clearly observe both contributions, and the green line is given by the corresponding 2-parameter fit to the σ_pred distribution from the BINN.

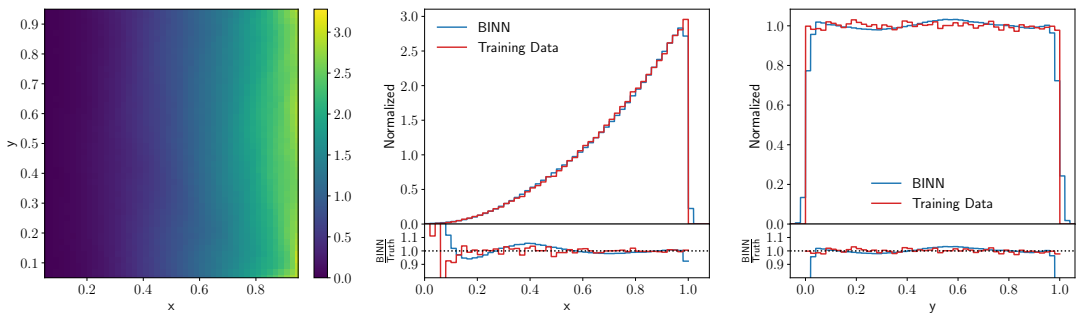


Figure 4.3: Two-dimensional and marginal densities for the quadratic kicker ramp.

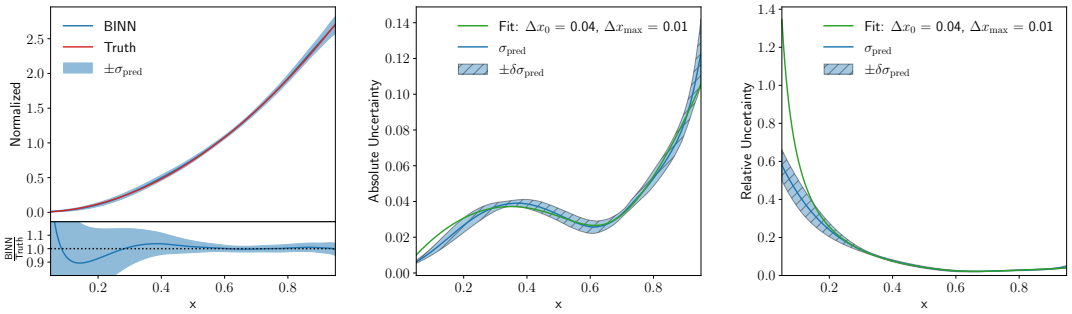


Figure 4.4: Density and predictive uncertainty distribution for the kicker ramp. In the left panel the density and uncertainty are averaged over several lines with constant y. In the central and right panels, the uncertainty band on σ_{pred} is given by their variation. The green curve represents a two-parameter fit to Eq. (4.13).

4.3.3 Gaussian ring

Our third example is a two dimensional Gaussian ring, which in terms of polar coordinates reads

$$p(r,\phi) = \text{Gauss}(r > 0; \mu = 4, w = 1) \times \text{Const}(\phi \in [0,\pi]) \,, \tag{4.14}$$

We define the Gaussian density as the usual

$$\text{Gauss}(r) = \frac{1}{\sqrt{2\pi}\, w} \exp\left[-\frac{1}{2w^2}(r-\mu)^2\right] \tag{4.15}$$

The density defined in Eq. (4.14) can be translated into Cartesian coordinates as

$$p(x,y) = \text{Gauss}(r(x,y); \mu = 4, w = 1) \times \text{Const}(\phi(x,y) \in [0,\pi]) \times \frac{1}{r(x,y)} \tag{4.16}$$

where the additional factor $1/r$ comes from the Jacobian. We train the BINN on Cartesian coordinates, just like in the two examples before, and limit ourselves to $y > 0$ to avoid problems induced by learning a non-trivial topology in mapping the latent and phase spaces. In Fig. 4.5 we once again see that our network describes the true two-dimensional density well.

In Fig. 4.6 we show the Cartesian density but evaluated on a line of constant angle.

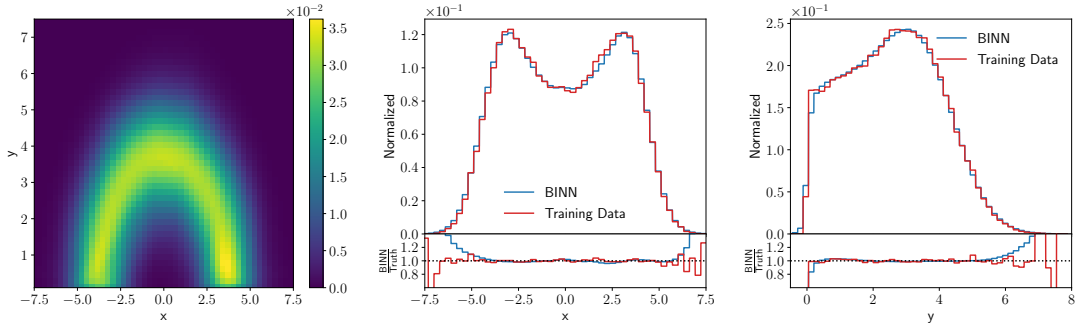


Figure 4.5: Two-dimensional and marginal densities for the Gaussian (half-)ring.

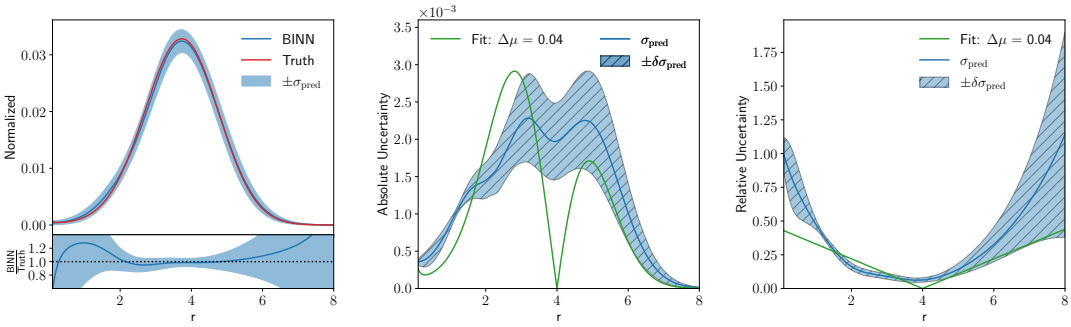


Figure 4.6: Cartesian density and predictive uncertainty distribution for the Gaussian ring. In the left panel the density and uncertainty are averaged over several lines with constant ϕ. In the central and right panels, the uncertainty band on σ_{pred} is given by their variation. The green curve represents a two-parameter fit to Eq. (4.17).

This form includes the Jacobian and has the expected, slightly shifted peak position at $r_{\text{max}} = 2 + \sqrt{3} = 3.73$. The BINN returns a predictive uncertainty, which grows towards both boundaries. The error band easily covers the deviation of the density learned by the BINN and the true density. While the relative predictive uncertainty appears to have a simple minimum around the peak of the density, we again see that the absolute uncertainty has a distinct structure with a local minimum right at the peak. The question is what we can learn about the INN from this pattern in the BINN.

As before, we describe our distribution in the relevant direction in terms of convenient fit parameters. For the Gaussian radial density these are the mean μ and the width w used in Eq. (4.14). The contributions driven by the extraction of the mean in Cartesian coordinates reads

$$\begin{aligned} \sigma_{\text{pred}} \equiv \Delta p &\supset \left| \frac{G(r)}{r} \frac{\mu - r}{w^2} \right| \Delta\mu \\ \text{and} \qquad \sigma_{\text{pred}} \equiv \Delta p &\supset \left| \frac{(r-\mu)^2}{w^3} - \frac{1}{w} \right| \Delta w \,. \end{aligned} \tag{4.17}$$

In analogy to Eq. (4.9) the two contributions need to be added in quadrature for the full, fit-like uncertainty. The contribution from the the mean has a minimum at $r = \mu = 4$ and is otherwise dominated by the exponential behavior of the Gaussian, just as we observe in the BINN result. In the central and right panels we show a one-parameter fit of the BINN output and find that the network determined the mean of the Gaussian as $\mu = 4 \pm 0.037$. We observe that including Δw doesn't improve the goodness of the fit.

4.3.4 Errors vs training statistics

Even though it is clear from the above discussion that we cannot expect the predictive uncertainties to have a simple scaling pattern, like for the regression [1] and classification [60] networks, there still remains the question how the BINN uncertainties change with the size of the training sample.

In Fig. 4.7 we show how the BINN predictions for the density and uncertainty change if we vary the training sample size from 10k events to 1M training events. Note that for all toy models, including the kicker ramp in Sec. 4.3.2, we use 300k training events. For

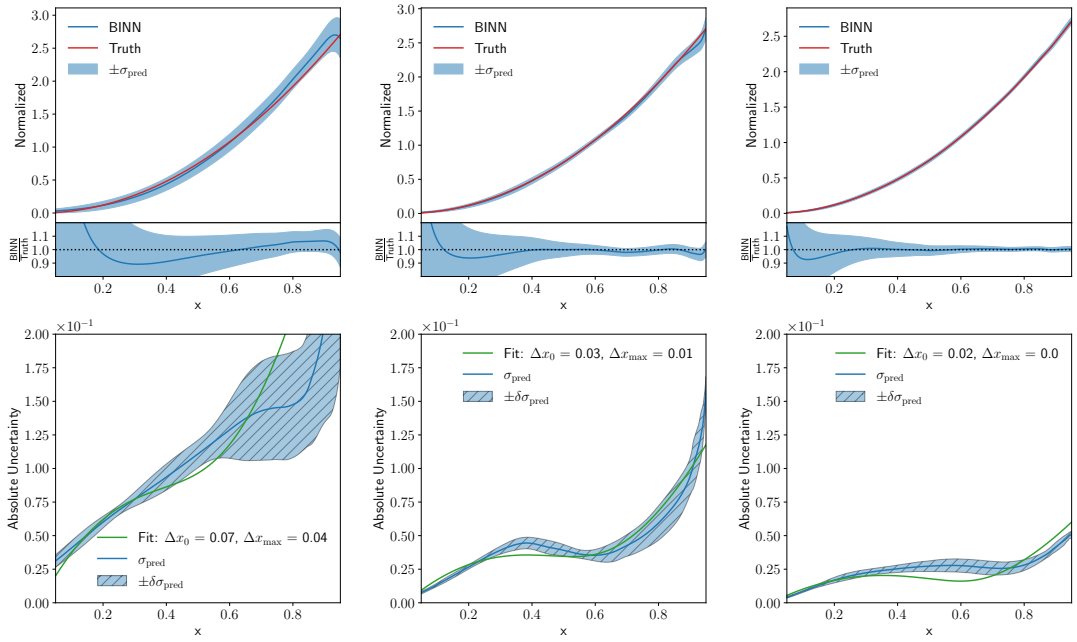


Figure 4.7: Dependence of the density (upper) and absolute uncertainty (lower) on the training statistics for the kicker ramp. We illustrate BINNs trained on 10k, 100k, and 1M events (left to right), to be compared to 300k events used for Fig. 4.4. Our training routine ensures that all models receive the same number of weights updates, regardless of the training set size.

the small 10k training sample, we see that the instability of the BINN density becomes visible even for our reduced x-range. The peak-dip pattern of the absolute uncertainty, characteristic for the kicker ramp, is also hardly visible, indicating that the network has not learned the density well enough to determine its shape. Finally, the variation of the predictive density explodes for $x > 0.4$, confirming the picture of a poorly trained BINN. As a rough estimate, the absolute uncertainty at $x = 0.5$ with a density value $p(x, y) = 0.75$ ranges around $\sigma_{\text{pred}} = 0.11 \ldots 0.15$.

For 100k training events we see that the patterns discussed in Sec. 4.3.2 begin to form. The density and uncertainty encoded in the network are stable, and the peak-dip with a minimum around $x = 2/3$ becomes visible. As a rough estimate we can read off $\sigma_{\text{pred}}(0.5) \approx 0.06 \pm 0.03$. For 1M training events the picture improves even more and the network extracts a stable uncertainty of $\sigma_{\text{pred}}(0.5) \approx 0.03 \pm 0.01$. Crucially, the dip around $x \approx 2/3$ remains, and even compared to Fig. 4.4 with its 300k training events the density and uncertainty at the upper phase space boundary are much better controlled.

Finally, we briefly comment on a frequentist interpretation of the BINN output. We know from simpler Bayesian networks [1,60] that it is possible to reproduce the predictive uncertainty using an ensemble of deterministic networks with the same architecture. However, from those studies we also know that our class of Bayesian networks has a very efficient built-in regularization, so this kind of comparison is not trivial. For the BINN results shown in this paper we find that the detailed patterns in the absolute uncertainties are extracted by the Bayesian network much more efficiently than they would be for ensembles of deterministic INNs. For naive implementations with a similar network size and no fine-tuned regularization these patterns are somewhat harder to extract. On the other hand, in stable regions without distinctive patterns the spread of

ensembles of deterministic networks reproduces the predictive uncertainty reported by the BINN.

4.3.5 Marginalizing phase space

Before we move to a more LHC-related problem, we need to study how the BINN provides uncertainties for marginalized kinematic distribution. In all three toy examples the two-dimensional phase space consists of one physical and one trivial direction. For instance, the kicker ramp in Sec. 4.3.2 has a quadratic physical direction, and in a typical phase space problem we would integrate out the trivial, constant direction and show a one-dimensional kinematic distribution. From our effectively one-dimensional uncertainty extraction, we know that the absolute uncertainty has a characteristic maximum-minimum combination, as seen in the central panel of Fig. 4.4.

To compute the uncertainty for a properly marginalized phase space direction, we remind ourselves how the BINN computes the density and the predictive uncertainty by sampling over the weights,

$$\begin{aligned} p(x,y) &= \int d\theta\, q(\theta)\, p(x,y|\theta) \\ \sigma^2_{\text{pred}}(x,y) &= \int d\theta\, q(\theta)\, [p(x,y|\theta) - p(x,y)]^2 \; . \end{aligned} \tag{4.18}$$

If we integrate over the y-direction, the marginalized density is defined as

$$\begin{aligned} p(x) = \int dy\, p(x,y) &= \int dy d\theta\, q(\theta)\, p(x,y|\theta) \\ &= \int d\theta\, q(\theta) \int dy\, p(x,y|\theta) \equiv \int d\theta\, q(\theta)\, p(x|\theta) \; , \end{aligned} \tag{4.19}$$

which implicitly defines $p(x|\theta)$ in the last step, notably without providing us with a way to extract it in a closed form. The key step in this definition is that we exchange the order of the y and θ integrations. Nevertheless, with this definition at hand, we can *define* the uncertainty on the marginalized distribution as

$$\sigma^2_{\text{pred}}(x) = \int d\theta\, q(\theta)\, [p(x|\theta) - p(x)]^2 \; . \tag{4.20}$$

We illustrate this construction with a trivial $p(x,y) = p(x,y_0)$, where we can replace the trivial y-dependence by a fixed choice $y = y_0$ just like for the wedge and kicker ramps. Here we find, modulo a normalization constant in the y-integration

$$\begin{aligned} \sigma^2_{\text{pred}}(x) &= \int d\theta\, q(\theta)\, [p(x|\theta) - p(x)]^2 \\ &= \int d\theta\, q(\theta) \int dy\; [p(x,y_0|\theta) - p(x,y_0)]^2 \\ &= \int dy d\theta\, q(\theta)\; [p(x,y_0|\theta) - p(x,y_0)]^2 = \int dy\, \sigma^2_{\text{pred}}(x,y_0) = \sigma^2_{\text{pred}}(x,y_0) \; . \end{aligned} \tag{4.21}$$

Adding a trivial y-direction does not affect the predictive uncertainty in the physical x-direction.

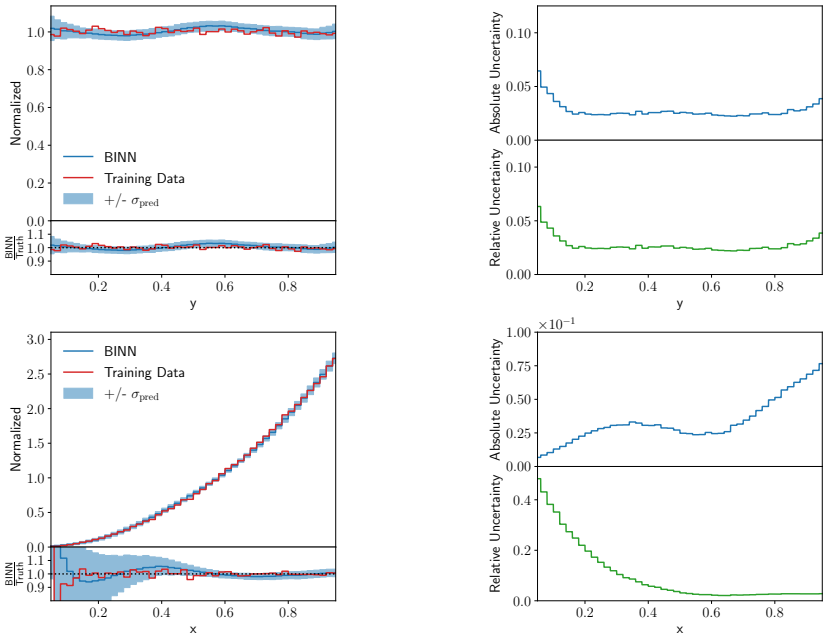


Figure 4.8: Marginalized densities and predictive uncertainties for the kicker ramp. Instead of the true distribution we now show the training data as a reference, to illustrate possible limitations. We use 10M phase space point to guarantee a stable prediction.

As mentioned above, unlike for the joint density $p(x, y|\theta)$, we do not know the closed form of the marginal distributions $p(x)$ or $p(x|\theta)$. Instead, we can approximate the marginalized uncertainties through a combined sampling in y and θ. We start with one set of weights θ_i from the weight distribution, based on one random number per INN weight. We now sample N points in the latent space, z_j, and compute N phase space points x_j using the BINN configuration θ_i. We then bin the wanted phase space direction x and approximate $p(x|\theta_i)$ by a histogram. We repeat this procedure $i = 1 \dots M$ times to extract M histograms with identical binning. This allows us to compute a mean and a standard deviation from M histograms to approximate $p(x)$ and $\sigma_{\text{pred}}(x)$. The approximation of σ_{pred} should be an over-estimate, because it includes the statistical uncertainty related to a finite number of samples per bin. For $N \gg 1$ this contribution should become negligible. With this procedure we effectively sample $N \times M$ points in phase space.

Following Eq. (4.19), we can also fix the phase space points, so instead of sampling for each weight sample another set of phase space points, we use the same phase space points for each weight sampling. This should stabilize the statistical fluctuations, but with the drawback of relying only on an effective number of N phase space points. Both approaches lead to the same σ_{pred} for sufficiently large N, which we typically set to $10^5 \dots 10^6$. For the Bayesian weights we find stable results for $M = 30 \dots 50$.

In Fig. 4.8 we show the marginalized densities and predictive uncertainties for the kicker ramp. In y-direction the density and the predictive uncertainty show the expected flat behavior. The only exceptions are the phase space boundaries, where the density starts to deviate slightly from the training data and the uncertainty correctly reflects that instability. In x-direction, the marginalized density and uncertainty can be compared to their one-dimensional counterparts in Fig. 4.4. While we expect the same peak-dip

structure, the key question is if the numerical values for $\sigma_{\text{pred}}(x)$ change. If the network learns the y-direction as uncorrelated additional data, the marginalized uncertainty should decrease through a larger effective training sample. This is what we typically see for Monte Carlo simulations, where a combination of bins in an unobserved direction leads to the usual reduced statistical uncertainty. On the other hand, if the network learns that the y-directions is flat, then adding events in this direction will have no effect on the uncertainty of the marginalized distribution. This would correspond to a set of fully correlated bins, where a combination will not lead to any improvement in the uncertainty. In Fig. 4.8 we see that the $\sigma_{\text{pred}}(x)$ values on the peak, in the dip, and to the upper end of the phase space boundary hardly change from the one-dimensional results in Fig. 4.4. This confirms our general observation, that the (B)INN learns a functional form of the density in both directions, in close analogy to a fit. It also means that the uncertainty from the generative network training is not described by the simple statistical scaling we observed for simpler networks [1, 60] and instead points towards a GANplification-like [115] pattern.

4.4 LHC events with uncertainties

As a physics example we consider the Drell-Yan process

$$pp \to Z \to e^+e^- \,, \tag{4.22}$$

with its simple $2 \to 2$ phase space combined with the parton density. The training set consists of an unweighted set of 4-vectors simulated with `Madgraph5` [161] at 13 TeV collider energy with the NNPDF2.3 parton densities [162]. We fix the masses of the final-state leptons and enforce momentum conservation in the transverse direction, which leaves us with a four-dimensional phase space. In our discussion we limit ourselves to a sufficiently large set of one-dimensional distributions. For these marginalized uncertainties we follow the procedure laid out in Sec. 4.3.5 with 50 samples in the BINN-weight space. In Tab. 4.2 we give the relevant hyper-parameters for this section.

To start with, we show a set of generated kinematic distributions in Fig. 4.9. The positron

Parameter	Flow
Hidden layers (per block)	2
Units per hidden layer	64
Batch size	512
Epochs	500
Trainable weights	$\sim$ 182k
Number of training events	$\sim$ 1M
Optimizer	Adam
(α, β_1, β_2)	(1×10^{-3}, 0.9, 0.999)
Coupling layers	20
Prior width	1

Table 4.2: Hyper-parameters for the Drell-Yan data set, implemented in `Pytorch` (v1.4.0) [64].

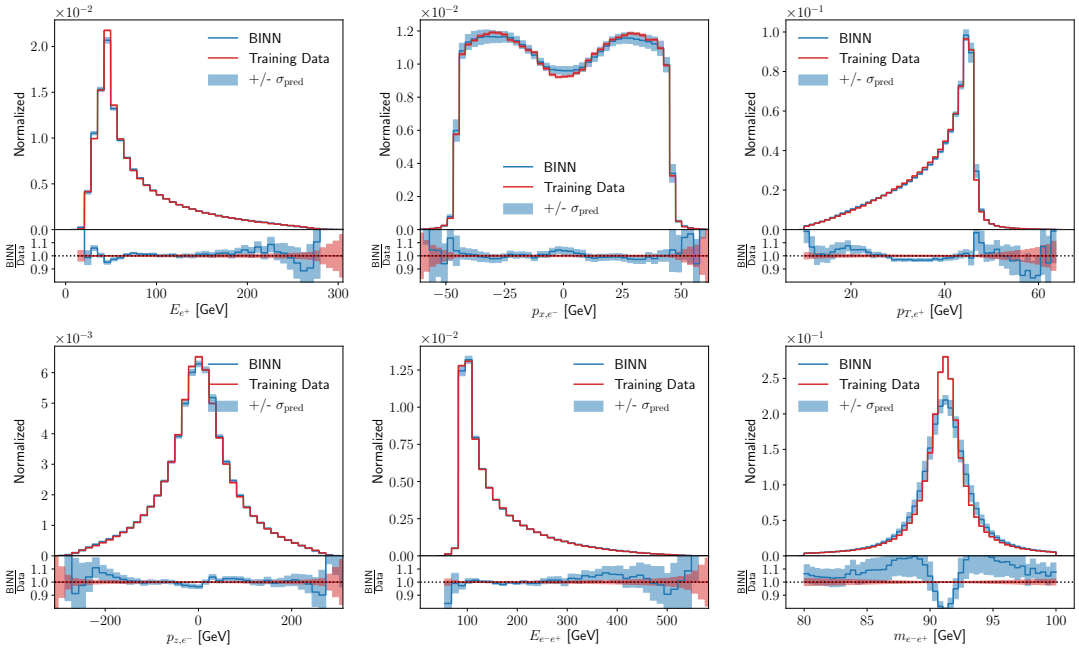


Figure 4.9: One-dimensional (marginalized) kinematic distributions for the Drell-Yan process. We show the central prediction from the BINN and include the predictive uncertainty from the BINN as the blue band. The red band indicates the statistical uncertainty of the training data per bin in the Gaussian limit.

energy features the expected strong peak from the Z-resonance. Its sizable tail to larger energies is well described by the training data to $E_e \approx 280$ GeV. The central value learned by the BINN becomes unstable at slightly lower values of 250 GeV, as expected. The momentum component p_x is not observable given the azimuthal symmetry of the detector, but it's broad distribution is nevertheless reproduced correctly. The predictive uncertainty covers the slight deviations over the entire range. What is observable at the LHC is the transverse momentum of the outgoing leptons, with a similar distribution as the energy, just with the Z-mass peak at the upper end of the distribution. Again, the predictive uncertainty determined by the BINN covers the slight deviations from the truth on the pole and in both tails. In the second row we show the p_z component as an example of a strongly peaked distribution, similar to the Gaussian toy model in Sec. 4.3.3.

While the energy of the lepton pair has a similar basic form as the individual energies, we also show the invariant mass of the electron-positron pair, which is described by the usual Breit-Wigner peak. It is well known that this intermediate resonance is especially hard to learn for a network, because it forms a narrow, highly correlated phase space structure. Going beyond the precision shown here would for instance require an additional MMD loss, as described in Ref. [128] and in more detail in Ref. [33]. This resonance peak is the only distribution where the predictive uncertainty does not cover the deviation of the BINN density from the truth. This apparent failure corresponds to the fact that generative networks always overestimate the width and hence underestimate the height of this mass peak [128]. This is an example of the network being limited by the expressive power in phase space resolution, generating an uncertainty which the Bayesian version cannot account for.

In Fig. 4.10 we show a set of absolute and relative uncertainties from the BINN. The

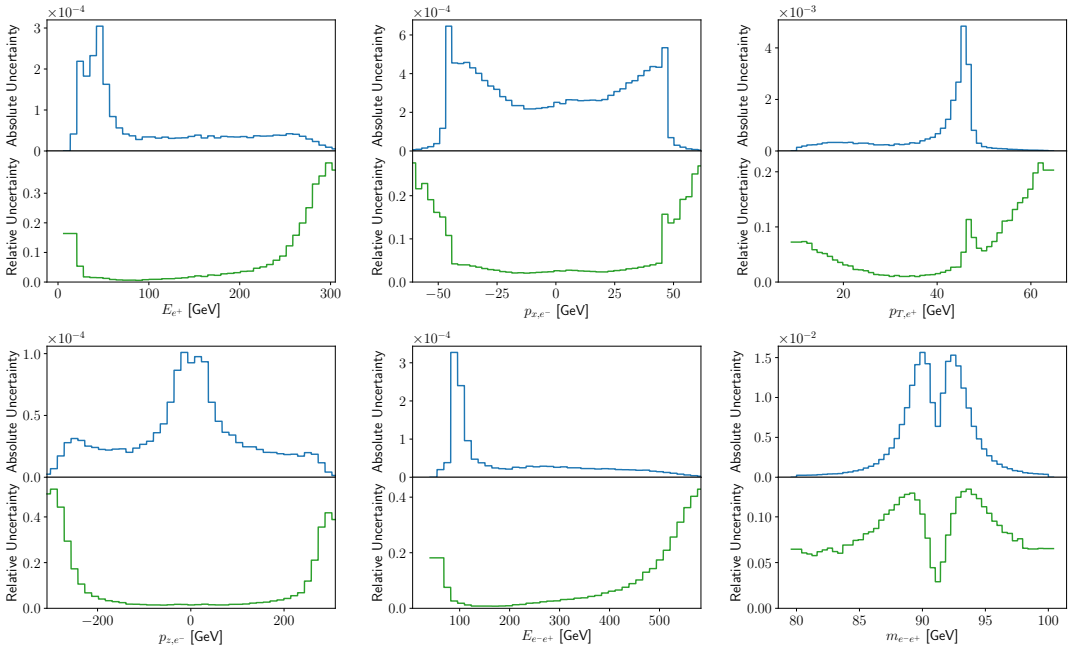


Figure 4.10: Absolute and relative uncertainties as a function of some of the kinematic Drell-Yan observables shown in Fig. 4.9.

strong peak combined with a narrow tail in the E_e distribution shows two interesting features. Just above the peak the absolute uncertainty drops more rapidly than expected, a feature shared by the wedge and kicker ramps at their respective upper phase space boundaries. The shoulder around $E_e \approx 280$ GeV indicates that for a while the predictive uncertainty follows the increasingly poor modeling of the phase space density by the BINN, to a point where the network stops following the truth curve altogether and the predictive uncertainty is limited by the expressive power of the network. Unlike the absolute uncertainty, the relative uncertainty keeps growing for increasing values of E_e. This behavior illustrates that in phase space regions where the BINN starts failing altogether, we cannot trust the predictive uncertainty either, but we see a pattern in the intermediate phase space regime where the network starts failing.

The second kinematic quantity we select is the (unobservable) x-component of the momentum. It forms a relative flat central plateau with sharp cliffs at each side. Any network will have trouble learning the exact shape of such sharp phase space patterns. Here the BINN keeps track of this, the absolute and the relative predictive uncertainties indeed explode. The only difference between the two is that the (learned) density at the foot of the plateau drops even faster than the learned absolute uncertainty, so their ratio keeps growing.

Finally, we show the result for the Breit-Wigner mass peak, the physical counterpart of the Gaussian ring model of Sec. 4.3.3. Indeed, we see exactly the same pattern, namely a distinctive minimum in the predictive uncertainty right on the mass peak. This pattern can be explained by the network learning the general form of a mass peak and then adjusting the mean and the width of this peak. Learning the peak position leads to a minimum of the uncertainty right at the peak, and learning the width brings up two maxima on the shoulders of the mass peak. In combination Figs. 4.9 and 4.10 clearly show that the BINN traces uncertainties in generated LHC events just as for the toy models. Again, some distinctive patterns in the predictive uncertainty can be explained

by the way the network learns the phase space density.

4.5 Outlook

Controlling the output of generative networks and quantifying their uncertainties is the main task for any application in LHC physics, be it in forward generation, inversion, or inference. We have proposed to use a Bayesian invertible network (BINN) to quantify the uncertainties from the network training for each generated event. For a series of two-dimensional toy models and an LHC-inspired application we have shown how the Bayesian setup indeed generates an uncertainty distribution, over the full phase space and over marginalized phase spaces. As expected, the learned uncertainty shrinks with an improved training statistics. Our method can be trivially extended from unweighted to weighted events by adapting the simple MLE loss.

An intriguing result from our study is that the combined learning of the density and uncertainty distributions allows us to draw conclusions on how a normalizing-flow network like the BINN learns a distribution. We find that the uncertainty distributions are naturally explained by a fit-like behavior of the network, rather than a patch-wise learning of the density. For the LHC, this can be seen for instance in the non-trivial uncertainty for an intermediate Breit-Wigner resonance. These results are another step in understanding GANplification patterns [115] and might even allow us to use INNs to extrapolate in phase space.

Obviously, it remains to be seen how our observations generalize to other generative networks architectures. For the LHC, the next step should be an in-depth study of INN-like networks applied to event generation.

Chapter 5

Event generation with Bayesian neural networks – a different angle

The research presented in this chapter has been published in Ref. [4]. Most of the figures, tables and text are completely identical to the content of this publication.

5.1 Introduction

Combining our expectations of a vastly increased dataset from the upcoming LHC runs with novel analysis methods and ever-improving theory predictions, we are looking at exciting times for particle physics. One of the keys to make optimal use of the LHC data is to consistently employ modern techniques, inspired by data science and further developed for particle physics application. Inference based on precision predictions from first principles critically rests on the assumption that we can provide theory predictions over the full phase space fast, precisely, and with flexible model assumptions. To meet the speed and precision expectations from HL-LHC we can use modern machine learning (ML) throughout the event generation and simulation chain [163].

A straightforward ML-task is regression of loop amplitudes, represented as a smooth scalar function over a relatively simple phase space. For simple $(2 \to 2)$-processes learning a non-divergent loop amplitude does not even require deep networks [124] and has been achieved with conventional interpolation methods [164, 165] as well. For higher final-state multiplicities [166] precision turns into a challenge, which we can try to meet by separating phase space into finite and divergent regions [108] or physics-inspired channels combined with very large training samples [167]. For our di-photon benchmark process at one-loop order current methods have shown to work well, but with limited precision especially in challenging regions of phase space [168].

Like in many physics applications, we would like to complement precision predictions of amplitudes with a reliable uncertainty estimate. Amplitudes are a simple problem because the training data consists of arbitrarily precise numerical values for well-defined phase space points. Once the network has learned all relevant features, we expect the leading uncertainties to reflect a possible local sparsity of the training data. Bayesian networks are perfectly suited to track training-related uncertainties [39]. In LHC physics they have been applied to regression [1], classification [60], ensembling [160], and generation [2, 169]. We will use Bayesian networks as a surrogate for ML-amplitudes because they learn amplitude values together with an uncertainty, and because we can use their likelihood loss to improve the network training.

When we want to train a network on amplitudes over phase space, with the additional condition that large amplitude values should be reproduced well, we need to re-think our training strategy. While usual NN-applications can be viewed as a non-parametric fit, we want to precisely reproduce individual amplitudes in the spirit of an interpolation [170]. We can force the network to reproduce certain amplitudes by boosting these amplitudes in the training. To decide which amplitudes need boosting, we use a Bayesian network with its point-wise uncertainty estimate. We find that moving freely between fit-like and the interpolation-like tasks allows us to improve the uncertainty estimate through a loss-based boosting and the precision though a process-specific performance boosting.

In Sec. 5.2 we introduce our dataset and the benchmark results, before introducing the Bayesian network in Sec. 5.3. The improved training through the two boosting strategies is illustrated for the $\gamma\gamma g$ channel in Sec. 5.4. Finally, we compare a set of 1-dimension kinematic distributions for the training data and the NN-amplitudes including uncertainties and with the different boosting strategies, before we provide an Outlook. In the Appendix we show the corresponding results for the $\gamma\gamma gg$ final state, based on the same concepts and architectures, but using a larger network and more training data.

5.2 Dataset and benchmark results

As an example process for our surrogate NN-amplitudes we use the partonic one-loop process [168]

$$gg \to \gamma\gamma g(g) \,, \tag{5.1}$$

generated with SHERPA [171] and the NJET amplitude library [172][2]. We apply a basic set of detector-inspired cuts on the partons in the final state,

$$\begin{aligned} &p_{T,j} > 20 \text{ GeV} && |\eta_j| < 5 && R_{jj,j\gamma,\gamma\gamma} > 0.4 \\ &p_{T,\gamma} > 40, 30 \text{ GeV} && |\eta_\gamma| < 2.37 \,. \end{aligned} \tag{5.2}$$

These cuts reduce the originally produced dataset from 1.1M $(2 \to 3)$-amplitudes of Ref. [168] to roughly 960k amplitudes. Each data point consists of real amplitude values as a function of the external 4-momenta, defining a 20-dimensional phase space. We divide our dataset into 90k training amplitudes and 870k independent test amplitudes acting as a high-statistics truth. The results shown in this chapter focus on the $2 \to 3$-process. For the $2 \to 4$-process we refer to Appendix A.2.

Transition amplitudes play a special role in the applications of neural networks to LHC physics, because they can be computed as functions of phase space with essentially arbitrary precision. The combination of high precision with limited training data is the challenge for the corresponding regression networks. Implicitly, it is assumed that the NN-amplitude networks will be faster than even the evaluation of leading-order amplitudes with state-of-the-art methods. The main figure of merit compares the true and the NN-amplitudes for a set of training or test data points,

$$\Delta_j^{(\text{train})} = \frac{A_{j,\text{NN}}}{A_{j,\text{train}}} - 1 \qquad \text{or} \qquad \Delta_j^{(\text{test})} = \frac{A_{j,\text{NN}}}{A_{j,\text{test}}} - 1 \,, \tag{5.3}$$

[2] In the main body of the paper we work with the $(2 \to 3)$-process, the corresponding results for the 4-body final state are presented in the Appendix. The interface between NJET and SHERPA is provided with Ref. [168] and available at `https://github.com/JosephPB/n3jet`.

where j runs over amplitude data points and we subtract 1 compared with the original paper [168]. Typical distributions of these Δ for existing calculations come with a width of 10% or more for the one-jet process of Eq. (A.1) [168]. For the tree-level process $e^+e^- \to q\bar{q}g$ the width of the Δ-distribution can be reduced to the per-mille level, using a training dataset of 60M amplitudes and a rather complex, physics-inspired architecture of networks [167].

Our approach follows a different strategy from the physics-inspired architectures mentioned above. We will use a relatively simple and small network, enhanced by a Bayesian network structure, and target the precision requirements with a new training strategy. The goal is to show that small and hence fast networks are expressive enough to describe a scattering amplitude over phase space. Enforcing and controlling the required precision leads us to, essentially, an appropriate loss function and the corresponding network training strategy.

5.3 Bayesian networks for amplitudes

Deterministic networks, usually trained by minimizing an MSE loss function, exhibit several weaknesses when it comes to LHC applications and controlled precision predictions. First, they only learn the amplitude value over phase space, without any information on if they have learned all features and how precise their estimate is. Second, their conceptually weak MSE loss function limits their performance. We will show how a Bayesian network with a likelihood comes with a whole range of conceptual and practical benefits.

Bayesian network set up

The loss function and network construction is equivalent to the one discussed in Sec. 2.2.2. In regression a BNN provides us with two types of uncertainties: σ_{stoch} and σ_{pred}. However, to avoid mis-understanding we refer to the former one in this chapter as a model-related uncertainty giving it the symbol σ_{model} rather than the usual σ_{stoch}. The dataset used for the study presented in this chapter is non-stochastic. The true amplitudes can be computed analytically and are, therefore, up to floating point precision exactly known. Thus, it is more accurate to say that σ_{model} captures the limitation of the expressive power of our model. As discussed in Sec. 2.2.2 with infinite training data σ_{pred} should go to zero, while σ_{model} approaches a constant. To have σ_{model} approach zero as well, we need to to go to the limit of having an extremely expressive model, a perfect training and infinite training data.

Network architecture

We use one Bayesian network trained on the entire training dataset. We train on amplitudes as a function of phase space with logarithmic preprocessing,

$$A_j \;\to\; \log\left(1 + \frac{A_j}{\sigma_A}\right) \;, \tag{5.4}$$

where σ_A is given by the distribution of the amplitude values. In addition, all phase space directions are preprocessed by subtracting the respective mean and dividing by the respective standard deviation.

The network describing the $(2 \to 3)$-part of the reference process in Eq. (A.1) consists of four hidden layers with 20 kinematic input dimensions, $\{20, 20, 30, 40\}$ nodes, and two output dimensions corresponding to the amplitude and its uncertainty, as illustrated in Eq. (2.16). The network has around 6k parameters. Between the hidden layers we use a tanh activation function, while for the last layer we find that a SoftPlus activations outperforms GeLU slightly and ReLU significantly. The network is trained on 90k amplitudes for 400000 epochs with a batch size of 8192 and learning rate of 10^{-4}, after which we observe no significant improvement in the loss. We use the Adam optimizer [114] with standard parameters.

BNN performance

As a first test of our BNN, we check the precision with which it approximates the true amplitudes in the training and test datasets, as defined in Eq. (5.3). For Fig. 5.1 we split the amplitudes by their absolute values, to see the effect of the limited training statistics in the collinear phase space regions. For the complete set of amplitudes the precision follows an approximate Gaussian with a width of a few per-mille, for the training and for

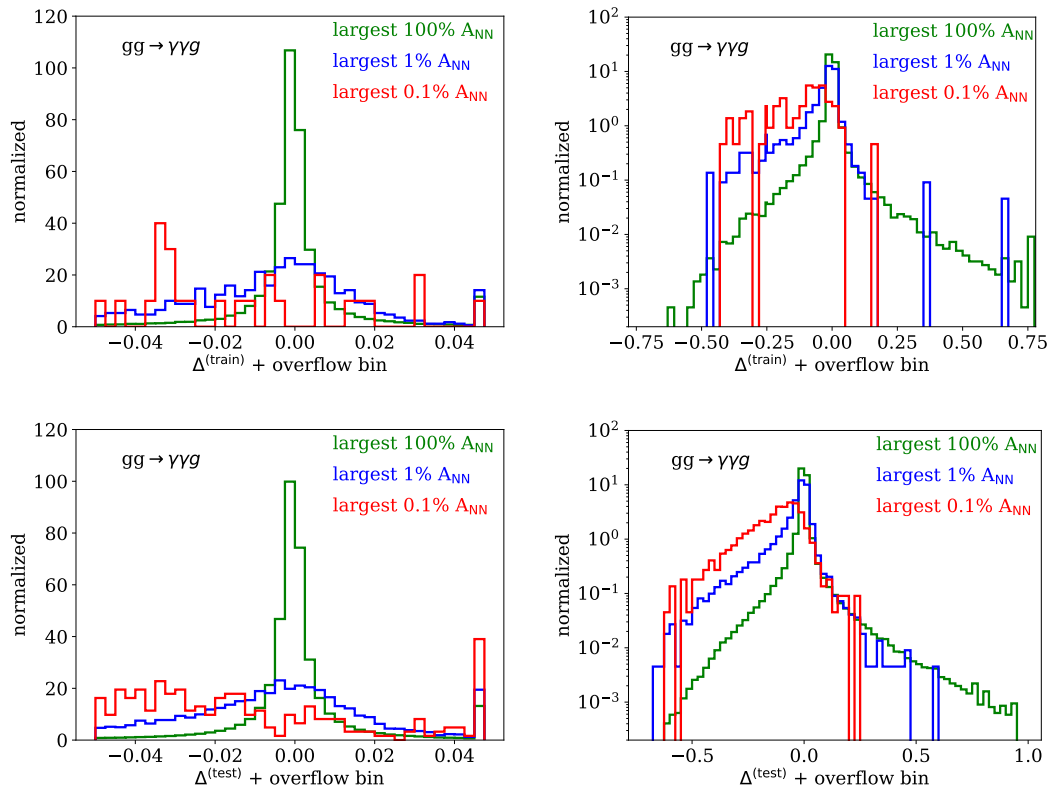


Figure 5.1: Performance of the BNN in terms of the precision of the generated amplitudes, Eq. (5.3), evaluated on the training (upper) and test datasets (lower).

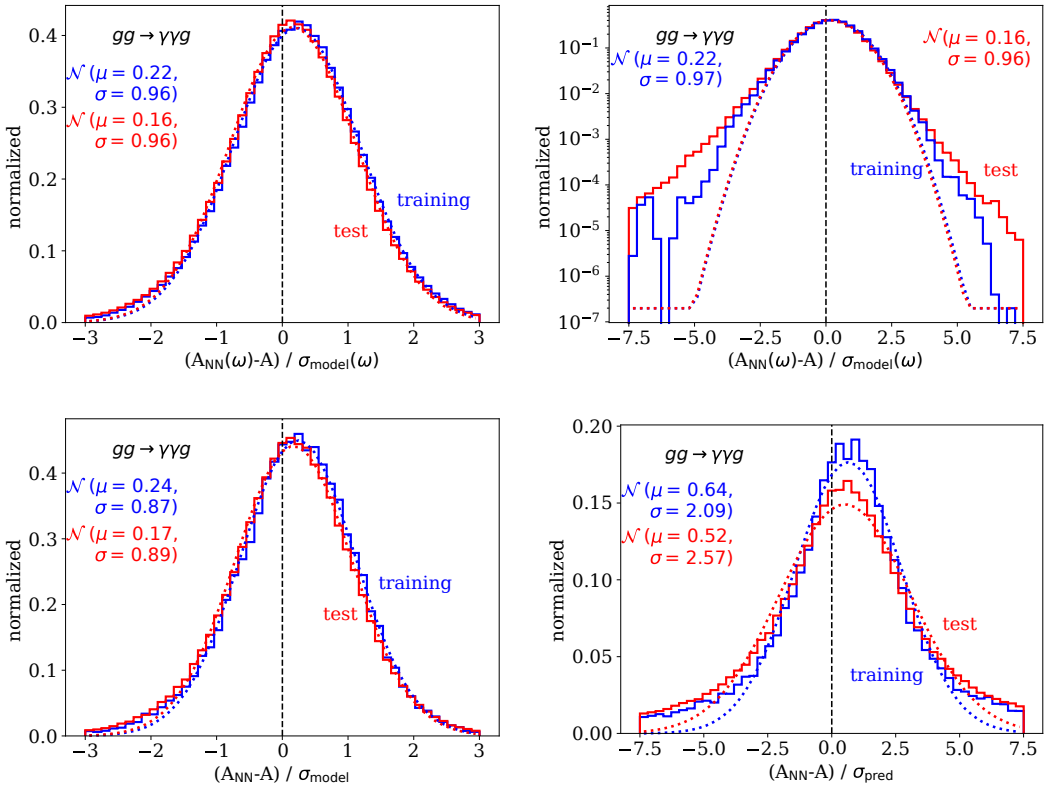


Figure 5.2: Pulls for the BNN, defined in Eq. (5.6) and evaluated on the training and test data. The two upper panels show the same curve for the weight-dependent pull on a linear and a logarithmic axis.

the test data. This matches the best available performance from the literature [167], but with a very compact and fast network.

In the logarithmic panels of Fig. 5.1 we see that the tails of the Δ-distributions for the full datasets are clearly enhanced. The picture changes when we only consider the phase space points with large amplitudes. For the 0.1% largest amplitudes the network is consistently less accurate with a slight tendency of underestimating the amplitudes. This is the motivation for training a separate network on the divergent phase space region(s) [168]. As we will see, the BNN offers an alternative approach which allows the full amplitude to be accurately described by a single network.

Because the BNN provides us with an uncertainty estimate for the NN-amplitude, we can define pull variables after integrating over the weight distributions,

$$t_{\text{model},j} = \frac{\langle A \rangle_j - A_j^{(\text{truth})}}{\sigma_{\text{model},j}} \qquad \text{or} \qquad t_{\text{pred},j} = \frac{\langle A \rangle_j - A_j^{(\text{truth})}}{\sigma_{\text{pred},j}} \,, \tag{5.5}$$

where the point-wise "truth" refers to the training or test datasets we use to evaluate the pulls. Neither of these pulls have an ω-dependent counterpart, because their numerators and denominators are sampled over the network weights independently. In the upper panels of Fig. 5.2 we see that the two pulls follow an approximate Gaussian shape, but with a much broader distribution for the σ_{pred}-based pull because of the smaller estimated uncertainty. We note that because of the log-rescaling of Eq. (5.4) it is not actually the

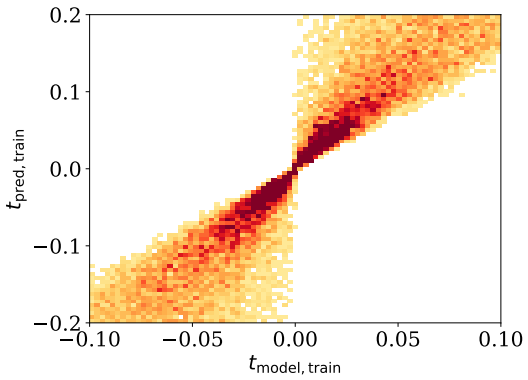


Figure 5.3: Correlation between the two pulls for the BNN, evaluated on the training data.

amplitudes A which should define Gaussian pulls, but their logarithms. We have explicitly checked that indeed the $\log A$ lead to a Gaussian, but that given our limited range of relevant amplitudes, the Gaussian shape translates into an approximately Gaussian shape for the amplitudes themselves.

Making use of the Gaussian likelihood loss of the BNN, Eq. (2.17), we can also define the weight-dependent pull

$$t_{\text{model},j}(\omega) = \frac{\overline{A}_j(\omega) - A_j^{(\text{truth})}}{\sigma_{\text{model},j}(\omega)} \,. \tag{5.6}$$

As part of the loss we can use its distribution as a consistency condition during network training. Given the Gaussian likelihood loss we expect a Gaussian distribution of $t_{\text{model},j}(\omega)$, sampled over ω according to the Gaussian $q(\omega)$ and over phase space points x. In the upper panels of Fig. 5.2 we see that, again, the pull distribution is Gaussian in the center, but develops symmetric, enhanced tails roughly two standard deviations from the mean.

Finally, we need to go back to the definition of the network uncertainties and understand how the split $\sigma^2_{\text{tot}} = \sigma^2_{\text{model}} + \sigma^2_{\text{pred}}$ can affect improved ways of training the amplitude network. We show the two sampled pulls defined in Eq. (5.5) in the lower panels of Fig. 5.2. Both are approximately Gaussian, and the width of the $t_{\text{model},j}$ distribution is much smaller than for $t_{\text{pred},j}$. This is an effect of the general observation that for a well-trained model

$$\sigma_{\text{tot}} \approx \sigma_{\text{model}} > \sigma_{\text{pred}} \,. \tag{5.7}$$

The only issue with all pulls shown in Fig. 5.2 is that they come with a slight bias towards positive values, which means the network slightly overestimates the amplitudes as a whole. This is in contrast to the underestimation of the 0.1% largest amplitudes observed in Fig. 5.1, and we will target it by improving the network training.

Figure 5.3 shows very strong correlations between the two pulls defined in Eq. (5.5). Both pulls correctly identify the training data points which are not described by the network well. For our regression tasks with exact amplitudes both uncertainties are largely induced by the lack of training statistics especially in the divergent phase space regions, so this correlation is expected.

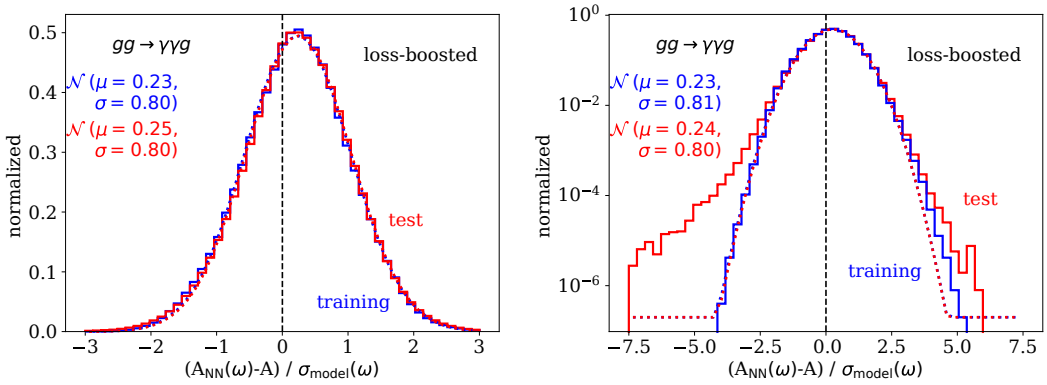


Figure 5.4: Pulls for the loss-boosted BNN, defined in Eq. (5.6) and evaluated on the training and test data. The two panels show the same results on a linear and a logarithmic axis. All curves can be compared to the BNN results without boosting in Fig. 5.2.

5.4 Network boosting

While the BNN-amplitude results described in the previous section are promising, the distribution of amplitudes and the pull distributions indicate potential improvements. We know that for generative networks we can employ an additional discriminator network to identify poorly learned phase space regimes [169], the solution is much simpler for a regression network. In the BNN loss we can compute the relative deviations between data and network output, or large pulls, and target these amplitudes directly. Once we control the network and its uncertainties, we can even think about further enhancing the training in the direction of an interpolation.

5.4.1 Loss-based boosting

Because the BNN loss in Eq. (2.17) represents a Gaussian log-likelihood, we can modify it and require a higher precision for those phase space points which according to the BNN uncertainty are not yet learned well. In practice, this is equivalent to feeding these training data points n_j times into the computation of the BNN loss

$$\mathcal{L}_{\text{Boosted BNN}} = \int d\omega \; q_{\mu,\sigma}(\omega) \sum_{\text{points } j} n_j \times \left[\frac{\left| \overline{A}_j(\omega) - A_j^{(\text{truth})} \right|^2}{2\sigma_{\text{model},j}(\omega)^2} + \log \sigma_{\text{model},j}(\omega) \right] + \frac{\sigma_q^2 - \sigma_p^2 + (\mu_q - \mu_p)^2}{2\sigma_p^2} + \log \frac{\sigma_p}{\sigma_q} \; . \tag{5.8}$$

As mentioned for Eq. (2.17), the regularization has to be adjusted for the additional amplitudes in the boosted training sample. This feedback training is similar to simple boosting algorithms for decision trees, where amplitudes for which the decision tree gives a wrong result are enhanced with additional weights. In our simple approach we duplicate some training amplitudes, or equivalently increases their weights in discrete steps.

In a first stage, we improve the self-consistency of the network with the initial assumptions and boost the network training for amplitudes with a large ω-dependent pull, Eq. (5.6).

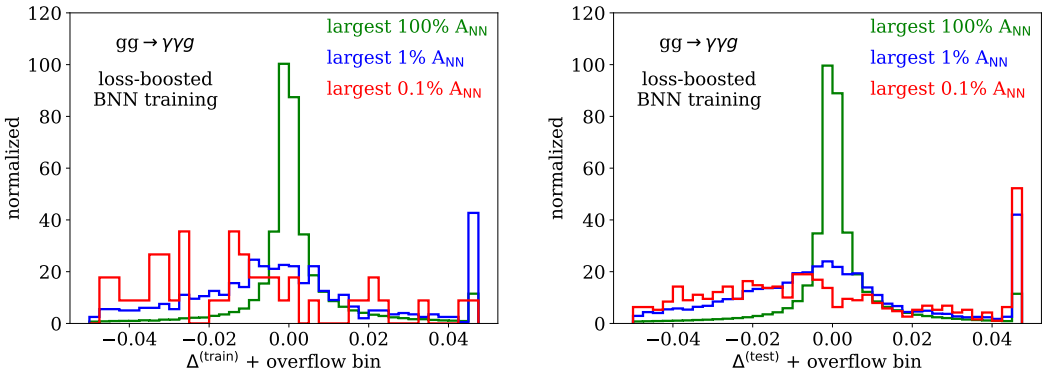


Figure 5.5: Performance of the loss-boosted BNN in terms of the precision of the generated amplitudes, Eq. (5.3), evaluated on the training and test datasets on a linear (left) and logarithmic (right) axis. The curves can be compared to the BNN results without boosting in Fig. 5.1.

In five iterations we identify the amplitudes with values of $t_{\text{model},j}(\omega)$, which are more than two standard deviations away from the mean and increase their contribution to the loss function in Eq. (5.8) by values $n_j = 5$. This is done four times. After adding the weights we continue the training on the enlarged datasets. For the next training cycle we again add weights to the amplitudes which now have pulls more than two standard deviations away. Each training ends when we see no more significant change to the loss which usually takes around 2000 epochs. This boosting forces the network towards a more self-consistent description of the tails of the pull distributions. We checked that small variations of n_j or the number of cycles do not have a significant impact on these improvements.

In Fig. 5.4 we show the pulls from the boosted Bayesian neural network, boosted based on the self-consistency of the loss measured by the pull. We see a significant improvement for $t_{\text{model}}(\omega)$, the parameter we target with our boosting. One would naively expect the corresponding distribution to assume a Gaussian shape with unit width. However, first of all our loss-based boosting only moves amplitudes from the tails into the bulk, which means that the tails of the boosted pull distributions should be low. Second, the pulls entering the loss and the pulls shown in Fig. 5.4 are different because the loss includes weights for high-pull amplitudes. In combination, both effects explain the narrower Gaussian for $t_{\text{model}}(\omega)$. In the logarithmic version we also see a visible over-training though loss-boosting.

Moving on to the precision for the amplitudes, we see in Fig. 5.5 that the loss-boosting only has a mild impact on the Δ-distributions. It does not significantly improve the precision of the amplitudes compared to Fig. 5.1, so we need a second boosting step.

5.4.2 Performance boosting

Given that the loss-boosting in the previous section worked for the uncertainty estimate but only had a modest effect on the performance of our amplitude network, we proceed to a more powerful boosting strategy. Independent of the self-consistency of the network, we know at the training level which amplitudes challenge the network. This means we can select them with the goal of improving the training for the largest amplitudes. The

difference between a general loss boosting and this process-dependent strategy is that now we target the largest and most poorly learned amplitudes by selecting them based on σ_{tot}. We choose the 200 amplitudes with the largest uncertainty σ_{tot} and add three additional copies to the training dataset. This process is repeated 20 times, where each training ends when we see no more significant change to the loss which is usually around 2000 epochs.

In Fig. 5.6 we first see that the process-specific performance boosting broadens the pull distributions and this way reverses some of the beneficial effects of the loss-boosting on $t_{\text{model}}(\omega)$. However, the widths of the boost distributions remains below one, and the bias towards larger amplitudes is removed. This is true for the training data and for the test data. In addition, the consistency with the Gaussian shape is broken symmetrically for too small and too large amplitudes, again consistently for training and test data. Given that the two boostings target different amplitudes and effectively compete with each other, this pattern is expected.

The positive impact on the large amplitudes can be seen more clearly in Fig. 5.7. Evaluated on the training data, the 0.1% largest amplitudes now show a clear peak at small Δ^{train}, consistent with all other amplitudes. This means the network has learned all amplitudes in the training dataset equally well. This effect translates to the test sample qualitatively, so the performance on the test data improves after performance-boosting, but this improvement is less pronounced than for the training data. This means that, at the expense of an overtraining, we have improved our network from a fit-like description to an interpolation-like description of the largest amplitudes.

The pattern observed by performance-boosting points to a conceptual weakness of standard network training when it comes to precision applications. If we stop the network training at the point where the performance on a training sample exceeds the performance on the test sample, we miss the opportunity of improving the network on the test and training data, but at a different rate. Overtraining is, per se, not a problem, as we know from applications of interpolation to describe data. The only challenge from such a network overtraining is a reliable uncertainty estimate from the generalization, for which we propose an appropriate scheduling of loss-boosting and performance-boosting.

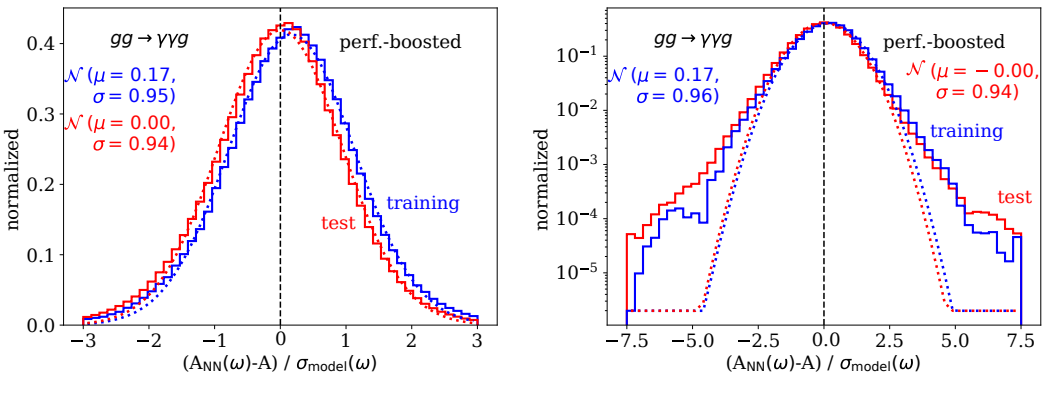


Figure 5.6: Pulls for the performance-boosted BNN, defined in Eq. (5.6) and evaluated on the training and test data. The two panels show the same results on a linear and a logarithmic axis. All curves can be compared to the BNN results without boosting in Fig. 5.2 and the loss-boosted results in Fig. 5.4.

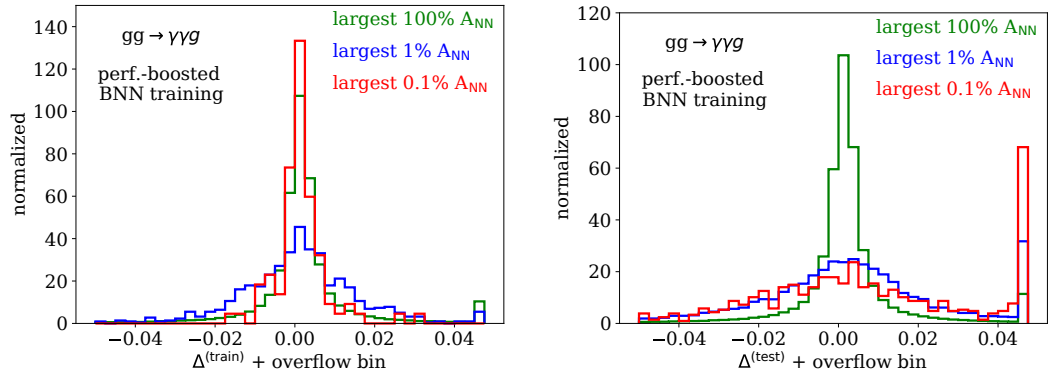


Figure 5.7: Performance of the performance-boosted BNN in terms of the precision of the generated amplitudes, Eq. (5.3), evaluated on the training (left) and test datasets (right). All curves can be compared to the BNN results without boosting in Fig. 5.1 and the loss-boosted results in Fig. 5.5.

5.4.3 Effect of training statistics

Given that our amplitude-BNN has successfully learned the amplitudes for the partonic process $gg \to \gamma\gamma g$ well below the percent level, with a small and simple network and 90k training points, we can ask the question how much training data we actually need for a precision amplitude network. For this study we use the same BNN as before, including loss-boosting and performance-boosting, but trained on a reduced dataset of

$$10\% \quad (9.000 \text{ amplitudes}) \quad \cdots \quad 100\% \quad (90.000 \text{ amplitudes}) \,. \tag{5.9}$$

In Fig. 5.8 we show the corresponding Δ-distributions for the test dataset. Our smallest training dataset contains 9000 amplitudes, which turn out sufficient to train our network with its 6192 parameters. The corresponding network reproduces the test data well, albeit with sizable overflow bins. Increasing the amount of training data improves the precision of the network, but relatively slowly. We observe the same level of improvement for all amplitudes and for the 1% largest amplitudes. For the latter we only show results after process boosting, without any boosting the quality of the low-statistics training is comparably poor.

5.5 Kinematic distributions

After illustrating the performance of the amplitude network in a somewhat abstract manner, we can also show 1-dimensional kinematic distributions. The integration of the remaining phase space dimension requires a little care, because we cannot just integrate the uncertainties together with the central values for the amplitudes.

For the central values we combine the amplitudes with phase space sampling. For example applying the simple RAMBO [173] algorithm we identify the phase space weights with A. A 1-dimensional distribution is generated through bins which collect the sum of the

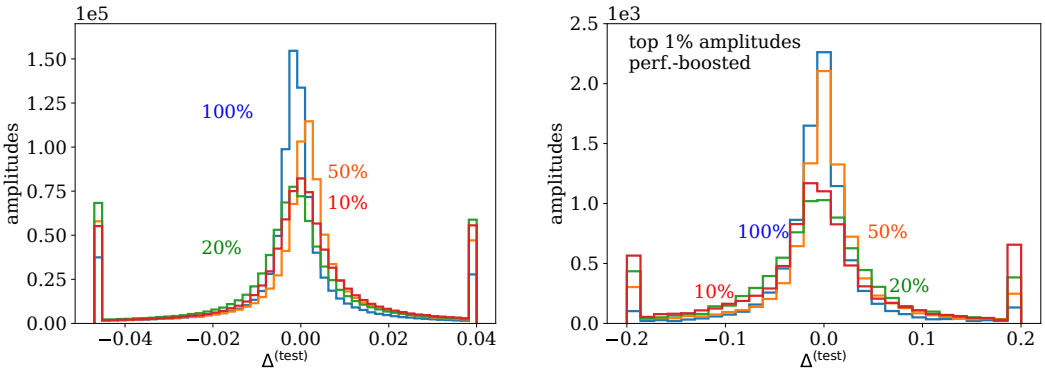


Figure 5.8: Performance of the BBN for all amplitudes (left) and a performance-boosted BNN for the largest 1% of all amplitudes (right), after training on different fractions of the full training dataset.

amplitudes in the remaining phase space directions. The histogram value for a bin k is

$$h_k = \sum_{j=1}^{N} A_j \; . \tag{5.10}$$

To use the amplitudes predicted by the BNN we have to add the sampling over the weights ω. By replacing the truth amplitudes with the NN-amplitudes we can target the uncertainties from the modeling of the amplitudes through the BNN. In analogy to Eq. (2.19) and omitting the index k for the histogram we first extract a central histogram value as

$$\begin{aligned} \langle h \rangle &= \int d\omega \; q(\omega) \sum_j \overline{A}_i(\omega) \\ &= \int d\omega \; q(\omega) \, \overline{h}(\omega) \qquad \text{with} \qquad \overline{h}(\omega) = \sum_j \bar{A}_j(\omega) \; . \end{aligned} \tag{5.11}$$

Again in analogy to Eqs. (2.21) and (2.22) we define the absolute uncertainties on the bin entry as

$$\begin{aligned} \sigma^2_{h,\text{pred}} &= \int d\omega \; q(\omega) \left[\overline{h}(\omega) - \langle h \rangle \right]^2 \\ \sigma^2_{h,\text{model}} &= \int d\omega \; q(\omega) \left[\overline{h^2}(\omega) - \overline{h}(\omega)^2 \right] \; . \end{aligned} \tag{5.12}$$

The total uncertainty is again $\sigma^2_{h,\text{tot}} = \sigma^2_{h,\text{model}} + \sigma^2_{h,\text{pred}}$. We can simplify $\sigma_{h,\text{model}}$ further. In all of the above formulas h is just a sum over amplitudes. If we assume that the corresponding σ_model values are uncorrelated, we can relate $\sigma_{h,\text{model}}$ to σ_model by exchanging the sum and the variance,

$$\begin{aligned} \sigma^2_{h,\text{model}} &= \int d\omega \; q(\omega) \, \text{Var}(\overline{h}(\omega)) \\ &= \int d\omega \; q(\omega) \, \text{Var} \left(\sum_j A_j(\omega) \right) = \int d\omega \; q(\omega) \sum_j \text{Var} \left(A_j(\omega) \right) \end{aligned}$$

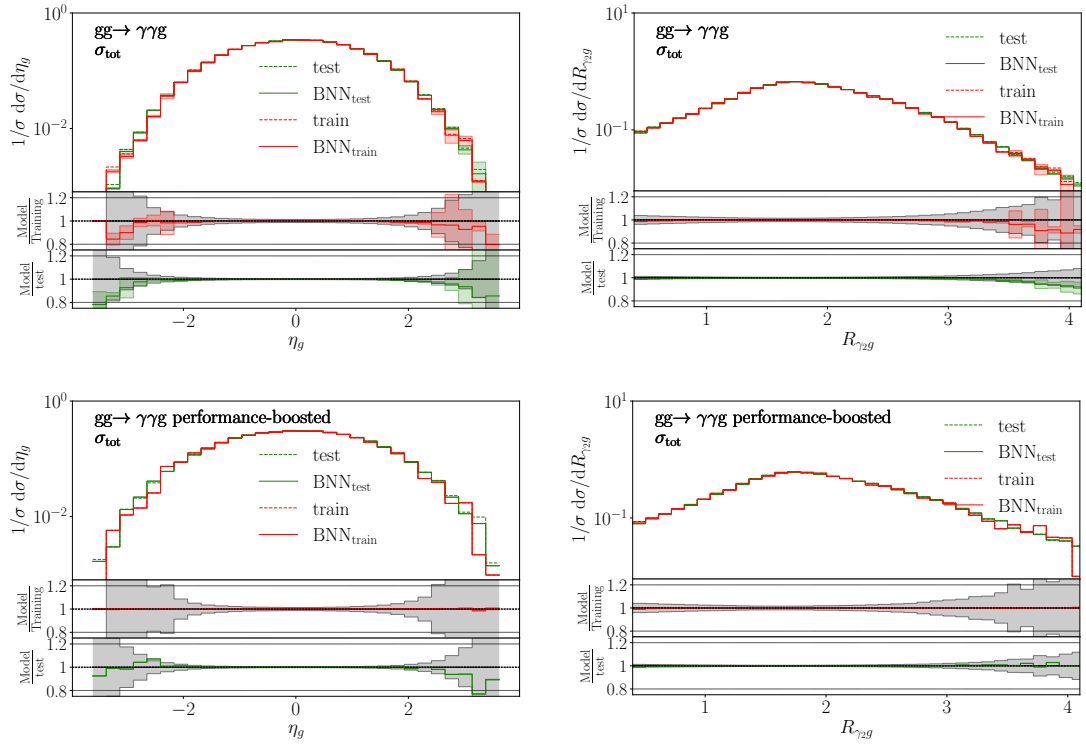


Figure 5.9: Kinematic distributions from the BNN without boosting (upper) and after performance-boosting (lower). The grey error bars in the lower panels indicate the statistical limitation of the training and test data.

$$= \int d\omega \; q(\omega) \sum_j \sigma^2_{\text{model},j}(\omega) \equiv \left\langle \sum_j \sigma^2_{\text{model},j}(\omega) \right\rangle . \tag{5.13}$$

While we assume uncorrelated uncertainties for σ_{model} we cannot do the same for σ_{pred}. To compute σ_{pred} we first sample a set of weight configurations, which turns our BNN into an ensemble of neural networks, and then use each of these neural networks to compute the corresponding histogram value. Computing the standard deviation of these values gives us an estimate for $\sigma_{h,\text{pred}}$. By sampling from the weight distributions we change the neural network itself and all of its predictions. To assume that these changes are uncorrelated for different amplitudes seems not exceptionally well justified.

Based on this procedure we show BNN-amplitude results for a set of kinematic distributions in the upper panels of Fig. 5.9. We see the effect of limited training data towards the end of the different kinematic distributions, where the agreement between the NN-amplitudes and truth deteriorates. For our reference process this happens for $|\eta_g| \gtrsim 2.5$ or $|\eta_\gamma| \gtrsim 1.5$. Still, the BNN uncertainty estimate covers the deviation from the truth reliably.

In the lower panels of Fig. 5.9 we see that after performance-boosting the BNN predictions agree with the training data spectacularly well. This is the goal of the boosting and leads to the network learning all features in the training data extremely well. In the phase space regions where the regular BNN precision is limited by sparse and large training amplitudes, the improved agreement between NN-amplitudes and the training data carries over to the test data at a level that the network prediction is significantly

improved. The uncertainties for the training data still cover the deviations from the truth, but unlike the central values this uncertainty estimate does not generalize correctly to the test data. This structural issue with process boosting could be ameliorated by alternating between loss-boosting and performance-boosting, until the specific requirements of a given analysis in precision and uncertainty estimates are met.

5.6 Outlook

Learning loop-amplitudes for LHC simulations is a classic ML-problem, because we need to train a precision network only once to provide a much faster simulation tool which can be used many times. In this application neural networks really work like better fits to the training data. Unlike for many other network applications, the training amplitudes are not noisy, which means we would like to reproduce the training amplitudes exactly, supplemented with a controlled uncertainty over all of phase space. To provide a reliable uncertainty map over phase space, we can rely on Bayesian regression networks [1].

The precision task reminds us of an interpolation rather than a fit, which means we need modify our ML-approach conceptually. If we are willing to accept a certain amount of overtraining, we can significantly improve the network training through boosting certain amplitudes. Because the Bayesian network provides a reliable uncertainty estimate, we can select the to-be-boosted amplitudes based on their deviation from the training data in units of the uncertainty. This loss-based boosting simply improves the self-consistency of the Bayesian network training. In a second step, we can boost training amplitudes just based on their absolute uncertainty. This selection helps with the precision for a given process, and because we use the absolute uncertainty we typically focus on the largest amplitude values.

We have applied Bayesian network training and the two strategies of amplitude boosting to the partonic process $gg \to \gamma\gamma g$ [168]. We have first found that the network amplitudes agree with the true amplitudes at the sub-percent level, for the training data and for a test dataset. For the 1% largest amplitudes an agreement at the percent level required process-specific performance boosting. For 1-dimensional kinematic distributions we have seen that the performance-boosting allows for extremely precise predictions in kinematic tails, albeit with a somewhat reduced performance in the uncertainty estimate for the test dataset. This can be improved by alternating between process and loss boosting in order to retain improved uncertainty estimation and increased performance which will be subject of future studies.

Finally, we have checked what happens with our boosted Bayesian network training when we reduce the number of training amplitudes from 90k to 9k and found that thanks to the boosting this only leads to a mild decrease in the network precision. This leaves us confident that boosted amplitude training with its shift from a fit-like to interpolation-like objective provides us with highly efficient surrogate models whenever the generation of training data is CPU-intensive.

Chapter 6

From models to SMEFT and back?

The research presented in this chapter has been published in Ref. [3]. Most of the figures, tables and text are completely identical to the content of this publication.

6.1 Introduction

The Higgs discovery [9,10] and many measurements of the Higgs Lagrangian [174] indicate that the Standard Model with its single, weakly interacting Higgs boson might well be the correct effective theory around the electroweak scale. However, the Standard Model is extremely unlikely to be the full story. Many theoretical considerations, including electroweak baryogenesis, dark matter, or neutrino mass generation, point to an extended electroweak or scalar sector. To avoid a bias through a specific, pre-selected signal hypothesis, modern LHC searches for beyond the Standard Model (BSM) physics are often conducted in the Standard Model effective theory (SMEFT) [175]. Because of its vast operator landscape, the corresponding experimental searches [176, 177] and global analyses [178–184] provide a comprehensive probe of rates and kinematic patterns in LHC processes.

One of the complications of SMEFT analyses of LHC data is that the effective theory truncated at dimension six has a limited validity range, and that LHC measurements span a large energy range. Moreover, even if we assume the SMEFT to be generally valid, it is not clear how much information on a full BSM model is lost when we confront it with LHC data via a truncated SMEFT Lagrangian rather than the original full model. Combining these questions, it is instructive to consider concrete, albeit simplified, BSM models and examine the limits extracted through a SMEFT interpretation matched to these models in comparison with the constraints obtained from direct searches [183, 185–192].

The naive expectation behind SMEFT analyses is that we can use the complete, correlated information on the Wilson coefficients from a global analysis and derive limits on any BSM model through matching. However, if the BSM scale is not sufficiently well-separated from the electroweak (EW) scale, an interpretation based on the SMEFT Lagrangian truncated at dimension six will likely give inaccurate results [193, 194]. The theory uncertainties related to the matching to full models are usually not accounted for in global analyses, which instead take their Lagrangian as a fixed interpretation framework. In general, limits derived on BSM models through a SMEFT framework using the same data and with all uncertainties accounted for will differ from limits derived on the full model directly, where the former can be significantly weaker or stronger than the latter.

This work aims at exploring the complementarity of the two analysis strategies and at highlighting general aspects that emerge when the SMEFT results are related to a concrete BSM scenario. We address this question for a global analysis of electroweak, di-boson and Higgs measurements, matching the relevant Wilson coefficients to the UV-model at one loop, using functional matching methods. We use the SFitter framework and include a proper estimate of a new and non-negligible theory uncertainty from the variation of the matching scale. As a UV-model we use a triplet-extended gauge sector [185, 195–198] a standard scenario when it comes to motivating the SMEFT approach to the Higgs and electroweak sector. Such a triplet model can be linked for instance to the weakly coupled gauge group $SU(3) \times SU(2) \times SU(2) \times U(1)$ [199] or deconstructed extra dimensions [200].

The paper is organized as follows: in Sec. 6.2 we review the basics of functional one-loop matching, we define the gauge triplet model under study, and we provide details about the SFitter setup. In Sec. 6.3 we discuss the decoupling limit of the new heavy states and the relevance of the matching scale choice. The impact of these two aspects on the global analysis is illustrated via simplified fits. In Sec. 6.4 we present the results of a global fit to the full vector triplet model, based on the dimension-6 SMEFT Lagrangian, and compare our results with limits obtained from direct searches. We conclude in Sec. 6.5.

6.2 Basics

In this section, we briefly review the one-loop matching procedure, the UV-model, as well as the SFitter setup. Experienced readers are welcome to skip this section.

6.2.1 One-loop matching: generic approach

The methods of constructing and matching effective-field theories [201, 202] have been in use for more than four decades [203–206]. Generic expressions for the low-energy effective action of a gauge theory at the one-loop order were derived in the 80s [207]. More recently, the approach has been further explored, particularly within the context of SMEFT [208–221].

We consider a UV model which can be defined in terms of light fields ψ and heavy fields Ψ, and which supports a perturbative expansion based on a local Lagrangian. Heavy fields are characterized by the condition that the support of their spectral functions vanishes below a certain threshold. We may identify the threshold with a mass M, typically the lightest mass that belongs to the heavy spectrum. The remaining fields are understood as light fields.

The UV model is expressible in terms of an effective action $\Gamma_{\rm UV}[\psi, \Psi]$, the generating functional of its one-particle irreducible (1PI) vertex functions. If fields of spin higher than 1/2 are involved, or if global symmetries are present, it is constrained to be a solution of a Slavnov-Taylor identity. By assumption, $\Gamma_{\rm UV}$ is calculable in a loop expansion from a local Lagrangian $\mathcal{L}_{\rm UV}(\psi(x), \Psi(x))$ with a finite number of fields and parameters. The parameters depend on the choice of a regularization and renormalization scheme and are redefined order by order by suitable renormalization conditions. This includes resolving inherent ambiguities associated with field reparameterizations, such as wave-function renormalization and terms vanishing by equations of motion.

The EFT is likewise expressible in terms of an effective action $\Gamma_{\rm EFT}[\psi]$, a functional of the light fields only. Again, we assume that a perturbative loop expansion is possible, and that it can be computed from a local Lagrangian $\mathcal{L}_{\rm EFT}(\psi(x))$. The number of parameters of $\mathcal{L}_{\rm EFT}$ is intended to be finite, but it increases without bounds with the accuracy that we want to implement via matching conditions. To keep the EFT parameter set manageable, we have to define an organizing principle which amounts to a series of approximations, and a prescription to truncate this series at a certain order.

To find the EFT Lagrangian iteratively, one introduces the one-light-particle irreducible (1LPI) effective action $\Gamma_{\rm L,UV}[\psi]$. Formally, this is a double Legendre transform of $\Gamma_{\rm UV}[\psi,\Psi]$; in practice, it amounts to absorbing a maximal set of independent heavy-field propagators in the skeleton expansion of S-matrix elements. This results in redefined light-field effective vertices. By contrast, the light-field propagators are kept explicit. In general they still carry a mixture of light and residual heavy degrees of freedom, depending on the precise definition of the original UV model. Like the original effective action, $\Gamma_{\rm L,UV}[\psi]$ depends on conventions regarding renormalization and handling the equations of motion. In terms of this entity, the matching condition reads

$$\Gamma_{\rm L,UV}[\psi] = \Gamma_{\rm EFT}[\psi] + \Delta\Gamma[\psi] \,. \tag{6.1}$$

The matching error $\Delta\Gamma[\psi]$ describes a set of vertex-function corrections $\Delta\Gamma_i(x)$ that are not calculable from a local Lagrangian involving light fields only. The matching procedure succeeds if, in momentum space, all contributions to this error are sufficiently power-suppressed at low energy,

$$\Delta\Gamma_i(p) < c|p|^k \,, \tag{6.2}$$

where p is any light-particle mass or momentum component.

At the tree level, the 1LPI effective action $\Gamma^{(0)}_{\rm L,UV}[\psi]$ of the UV model can be derived by simple variable changes, applying the equations of motion. Unless the ψ multiplets are incomplete under a symmetry, the result satisfies the tree-level Slavnov-Taylor identity with only light fields taken into account. The tree-level effective action $S_{\rm EFT}[\psi] = \Gamma^{(0)}_{\rm EFT}[\psi]$ is evaluated, to arbitrary order, by means of a momentum-space Taylor expansion of the 1LPI effective action on the l.h.s. of Eq. (6.1). In this expansion, residual heavy degrees of freedom are naturally removed from the tree-level light-field propagators. The latter assume their canonical tree-level form while any extra terms are shifted to the interaction part of $S_{\rm EFT}[\psi]$.

The operator content of the tree-level effective action $S_{\rm EFT}[\psi]$ can be determined independently by algebraic methods. Their coefficients are fixed by a term-by-term comparison with the vertices of $\Gamma^{(0)}_{\rm L,UV}$. The symmetries are preserved in this expansion if covariant derivatives are used consistently. At one loop, new contributions to the UV effective action arise which are generically non-local, and can be formally summarized as

$$\Gamma^{1\ell}_{\rm UV}[\psi,\Psi] = ic_s \,{\rm Tr}\log\left(-\frac{\delta^2 S_{\rm UV}[\psi,\Psi]}{\delta^2(\psi,\Psi)}\right) \,, \tag{6.3}$$

where the trace is integrated over all field components at all space-time points and c_s accounts for the statistics of the fields that are integrated over. This evaluates to the sum of all one-loop Feynman graphs with external fields attached. In expressions of this kind, the external field insertions act as bookkeeping devices, or background fields [222–230].

This allows for employing gauges and conventions that distinguish between internal and external lines, a generic feature of working with 1PI vertex functions. This means in particular that, with respect to the background fields, a manifestly gauge-invariant effective action can be computed [231, 232]. The trace is in general UV divergent and requires the application of a regularization scheme and the addition of local counterterms, such as dimensional regularization and minimal subtraction.

To match the UV model to the EFT at the one-loop order, we have to evaluate Eq. (6.1) again. Initially,

$$\Delta\Gamma^{1\ell}[\psi] = ic_s \,\mathrm{Tr}\left[\log\left(-\frac{\delta^2 S_{\mathrm{UV}}[\psi,\Psi]}{\delta^2(\psi,\Psi)}\right) - \log\left(-\frac{\delta^2 S^{(0)}_{\mathrm{EFT}}[\psi]}{\delta^2\psi}\right)\right]\Bigg|_{\Psi=0}, \tag{6.4}$$

where the formal trace includes the integral over all one-loop diagrams which are 1LPI and do not contain open external Ψ lines. Because $S^{(0)}_{\mathrm{EFT}} = \Gamma^{(0)}_{\mathrm{EFT}} = \Gamma^{(0)}_{\mathrm{L,UV}} + \mathcal{O}(|p|^k)$, the difference is well-behaved in the IR. Loops of canonical light propagators only would exactly cancel between the two terms, but since the light-field propagators need not coincide between the two Lagrangians, we have to be careful to take all terms into account. In any case, due to the IR cancellation the one-loop functional Eq. (6.4) again admits a Taylor expansion up to the order of the previous tree-level truncation. The result can be expressed as a finite set of local terms that modify the coefficients of terms which are already present in the generic effective Lagrangian of the tree-level EFT. They are absorbed in $S_{\mathrm{EFT}}[\psi]$,

$$S^{1\ell}_{\mathrm{EFT}}[\psi] = -\Delta\Gamma^{1\ell}[\psi]\Big|_{\text{local, truncated}}, \tag{6.5}$$

and disappear from Eq. (6.4). In effect, the remainder still contains all non-local parts of the matching error but satisfies Eq. (6.2), to one-loop order.

By the same reasoning, the difference in Eq. (6.4) is not well-behaved but divergent in the UV, and therefore requires regularization and renormalization. The renormalization conditions are given by the matching conditions themselves and thus indirectly refer to the renormalization conditions of the UV model. All free parameters of the EFT are fixed, order by order, in terms of the original parameters of the UV model. Nevertheless, a practical scheme such as dimensional regularization with minimal subtraction may introduce an intermediate renormalization which depends on an arbitrary scale μ_R. The implications of this additional mass scale will be discussed in detail below.

In analogy with the tree-level matching procedure, in order to manifestly preserve the symmetries of the theory one should consistently work with covariant derivatives in the one-loop matching calculation, as discussed in the following subsection. However, due to the presence of UV divergences in the matching conditions the Slavnov-Taylor identity need not be compatible with a local Taylor expansion of the one-loop vertex functions, and the separation of the UV effective action into a gauge-invariant low-energy effective action and a remainder like in Eq. (6.1) may fail [207, 233, 234]. In the current paper, we assume that such an obstruction does not critically affect our argument.

6.2.2 One-loop matching: implementation

Instead of constructing the difference in Eq. (6.4) in terms of Feynman graphs explicitly, the subtraction may be accounted for in the integrand by employing the method of

regions [235–237]. The matching correction (Eqs. (6.4) and (6.5)) is replaced by

$$S^{1\ell}_{\rm EFT}[\psi] = ic_s\ {\rm Tr}\log\left(-\frac{\delta^2 S_{\rm UV}[\Psi,\psi]}{\delta(\Psi,\psi)^2}\right)\bigg|_{\rm hard}\,. \tag{6.6}$$

The label "hard" has to be understood in the following way: the functional trace is computed in momentum space. Two different regions are of interest in the matching, the hard and the soft region. If q denotes the typical size of a loop momentum, the hard region is defined by $q \sim M \gg m$, whereas the soft region is defined by $q \sim m \ll M$. Here m stands for the typical mass scale of the light sector. As discussed above, only the hard region is relevant while in the soft region the matching integral is well behaved. It has been shown that the tree-level induced EFT contribution to the matching cancels the soft region contributions from the UV-theory in the difference in Eq. (6.4) [213, 238]. Therefore, the integrands of the loop integrals in Eq. (6.6) are expanded only in the hard region. The evaluation of the functional trace then reduces to computing integrals of the form

$$\int \frac{{\rm d}^d q}{(2\pi)^d}\frac{q^{\mu_1}\dots q^{\mu_{2n_c}}}{(q^2-M_{i_1}^2)^{n_1}\dots(q^2-M_{i_m}^2)^{n_m}(q^2)^{n_0}}. \tag{6.7}$$

Here, all masses $M_{i_1},\dots,M_{i_m}$ are of the order of M. This implies that the dependence of Eq. (6.6) on any external momentum or mass $|p|$ is analytic, and no logarithms of the form $\log(m/|p|)$ or $\log(|p|/M)$ can appear. The only logarithm possible is $\log(M/\mu_R)$, and to avoid large logarithms in the relation between EFT and UV parameters we need to choose $\mu_R \sim M$.

Apart from the prescription "hard", the second derivatives of the UV-action evaluated at the background field configurations appear in the matching. To derive a universal result these derivatives are split into a part that contains the gauge-kinetic term of the field and its mass term, generating the propagator of the field, and a pure interaction contribution that appears in the final result. For the field ψ this latter piece is given by

$$X_{\psi\psi} = -\frac{\delta^2 S_{\rm UV,int.}}{\delta\psi^2}\ , \tag{6.8}$$

where only the interaction part of the action excluding the interactions with gauge bosons through the covariant derivative appears. The interactions with the gauge bosons are included in the propagator part of the functional derivative, which allows for an evaluation in which only gauge covariant objects appear at every step and the final result is manifestly gauge invariant. The price to be paid for this manifest gauge covariance is that every occurrence of a covariant derivative has to be shifted by a loop momentum in the evaluation of the functional trace in Eq. (6.6). We therefore have to parameterize Eq. (6.8) as

$$X_{\psi\psi} = U_{\psi\psi} + iD_\mu Z^\mu_{\psi\psi} + iZ^{\dagger\mu}_{\psi\psi}D_\mu + \dots, \tag{6.9}$$

where D_μ is the covariant derivative of the UV-model. The quantities $U_{\psi\psi}, Z^\mu_{\psi\psi}$ and $Z^{\dagger\mu}_{\psi\psi}$ only depend on covariant derivatives through commutators whereas the explicit covariant derivatives appearing in Eq. (6.9) are so-called open covariant derivatives that act on everything to their right. The ellipsis denotes terms with further open covariant derivatives. Importantly, contributions with one open covariant derivative arise at dimension six whenever there is a scalar field charged under the gauge group

and therefore they contribute to the matching through the presence of the Higgs field. Consequently, for our matching computations we use an extension of the results of Ref. [215], adding gauge bosons and the heavy resonance of our model. Since the gauge boson fluctuations appear in loops they have to be gauge fixed. This gauge fixing does not disturb the manifest gauge invariance at the level of the background fields and the gauge-fixing parameter can be chosen at convenience. Choosing Feynman gauge allows for easy incorporation of these operators into the results of Ref. [215], since we can treat gauge bosons like scalar fields with an extra index. Care has to be taken to account for the overall sign in the propagator. For the resonance this choice is not available since it does not have a gauge-fixing term and some operators with up to two open covariant derivatives have to be computed for the matching.

6.2.3 Triplet model

The UV model we study in this paper is a gauge-triplet extension of the Standard Model [185, 195–198]. In the unbroken electroweak phase, the Lagrangian reads

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{SM}} &- \frac{1}{4}\widetilde{V}^{\mu\nu A}\widetilde{V}^{A}_{\mu\nu} - \frac{\tilde{g}_M}{2}\,\widetilde{V}^{\mu\nu A}\widetilde{W}^{A}_{\mu\nu} + \frac{\tilde{m}_V^2}{2}\widetilde{V}^{\mu A}\widetilde{V}^{A}_{\mu} \\ &+ \sum_f \tilde{g}_f\,\widetilde{V}^{\mu A}J^{fA}_{\mu} + \tilde{g}_H\,\widetilde{V}^{\mu A}J^{HA}_{\mu} + \frac{\tilde{g}_{VH}}{2}\,|\phi|^2\widetilde{V}^{\mu A}\widetilde{V}^{A}_{\mu}\,,\end{aligned} \tag{6.10}$$

where $\widetilde{V}^A_\mu$ is a new, massive vector field transforming as a triplet of $SU(2)_L$, $\widetilde{W}^A_\mu$ are the SM weak gauge bosons, and ϕ is the SM Higgs doublet. The kinetic term of the vector field includes a covariant derivative,

$$\widetilde{V}^A_{\mu\nu} = \widetilde{D}_\mu\widetilde{V}^A_\nu - \widetilde{D}_\nu\widetilde{V}^A_\mu \quad \text{with} \quad \widetilde{D}_\mu\widetilde{V}^A_\nu = \partial_\mu\widetilde{V}^A_\nu - g_2 f^{ABC}\widetilde{W}^B_\mu\widetilde{V}^C_\nu\,. \tag{6.11}$$

where A, B, C are $SU(2)_L$ indices and the covariant derivative carries a tilde to indicate that it contains the fields $\widetilde{W}^A_\mu$. The currents coupling the heavy vector to the SM-fields are given by

$$J^{lA}_\mu = \bar{l}_i\gamma_\mu t^A l_j\,\delta^{ij}\,, \qquad J^{qA}_\mu = \bar{q}_i\gamma_\mu t^A q_j\,\delta^{ij}\,, \qquad J^{HA}_\mu = \phi^\dagger i\overleftrightarrow{D}^A_\mu\phi\,, \tag{6.12}$$

with l, q being the SM lepton and quark doublets, $t^A = \sigma^A/2$ the $SU(2)$ generators and σ^A the Pauli matrices. i, j are flavor indices and the Lagrangian is defined in a flavor-symmetric limit. In the Higgs current, $(\phi^\dagger i\overleftrightarrow{D}^A_\mu\phi) = i\phi^\dagger t^A(D_\mu\phi) - i(D_\mu\phi^\dagger)t^A\phi$. As pointed out in [238], the theory cannot be quantized in a self-consistent way for $\tilde{g}_{VH} < 0$.

The gauge mixing described by the triplet model is familiar from the general case of extra-$U(1)$ bosons [239]. A special feature is the explicit $\widetilde{V}$-mass term, which would have to be generated by some kind of symmetry breaking and likely involve additional fields; we ignore these additional fields, for instance in their effect on $\tilde{g}_M$. The Higgs doublet ϕ is yet to develop a VEV, which means we are working in the unbroken electroweak phase. Underlying this choice is the assumption that a SMEFT expansion for the EFT exists. This is the case unless there are additional sources of electroweak symmetry breaking, or a heavy particle obtains all of its mass from the Higgs VEV [240]. Even in the weakly coupled UV-completion of the triplet model there are no additional sources of electroweak symmetry breaking, because the additional scalar breaks $SU(2) \times SU(2)$ to $SU(2)_L$ and leaves the electroweak symmetry completely intact.

To remove the kinetic mixing, we can re-define the SM-gauge field as [196, 197]

$$W^{\mu\nu A} = \widetilde{W}^{\mu\nu A} + \tilde{g}_M \widetilde{V}^{\mu\nu A} = \partial^\mu(\widetilde{W}^{\nu A} + \tilde{g}_M \widetilde{V}^{\nu A}) - \partial^\nu(\widetilde{W}^{\mu A} + \tilde{g}_M \widetilde{V}^{\mu A}) + \cdots \tag{6.13}$$

For the triplet field we only allow for a re-scaling, $\widetilde{V}^{\mu A} = \alpha V^{\mu A}$, so that the triplet mass does not get transferred into the SM-gauge sector. The triplet mass also fixes the phase of the real vector field $\widetilde{V}^A_\mu$, such that α has to be real. Requiring a canonical normalization of the new kinetic term $V^{\mu\nu A} V^A_{\mu\nu}$ we find $\alpha^2 = 1/(1-\tilde{g}_M^2)$. This relation requires $\tilde{g}_M \neq \pm 1$, to ensure a valid model with a propagating heavy vector. Furthermore, as we will see in Sec. 6.3, we need to require $|\tilde{g}_M| < 1$ for the squared pole mass of the resonance to be positive. The final form of the gauge field re-definition in Eq. (6.13) becomes

$$\widetilde{W}^{\mu A} = W^{\mu A} - \frac{\tilde{g}_M}{\sqrt{1-\tilde{g}_M^2}} V^{\mu A} \qquad \text{and} \qquad \widetilde{V}^{\mu A} = \frac{1}{\sqrt{1-\tilde{g}_M^2}} V^{\mu A} \,, \tag{6.14}$$

and brings the Lagrangian into the form

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{SM}} &- \frac{1}{4} V^{\mu\nu A} V^A_{\mu\nu} + \frac{m_V^2}{2} V^{\mu A} V^A_\mu \\ &+ \sum_f g_f \, V^{\mu A} J^{fA}_\mu + g_H V^{\mu A} J^{HA}_\mu + \frac{g_{VH}}{2} \, |H|^2 V^{\mu A} V^A_\mu \\ &+ \frac{g_{3V}}{2} f^{ABC} \, V^{\mu A} V^{\nu B} V^C_{\mu\nu} - \frac{g_{2VW}}{2} f^{ABC} \, V^{\mu B} V^{\nu C} W^A_{\mu\nu} \,, \end{aligned} \tag{6.15}$$

which has the same structure as Eq. (6.10), but additional triple and quartic gauge couplings between the weak and triplet sectors. The Lagrangian parameters are related through

$$\begin{aligned} m_V^2 &= \frac{\tilde{m}_V^2}{1-\tilde{g}_M^2} \,, & g_H &= \frac{\tilde{g}_H + g_2 \tilde{g}_M}{\sqrt{1-\tilde{g}_M^2}} \,, & g_f &= \frac{\tilde{g}_f + g_2 \tilde{g}_M}{\sqrt{1-\tilde{g}_M^2}} \,, \\ g_{VH} &= \frac{2\tilde{g}_{VH} + g_2^2 \tilde{g}_M^2 + 2 g_2 \tilde{g}_H \tilde{g}_M}{2(1-\tilde{g}_M^2)} \,, & g_{3V} &= -\frac{2 g_2 \tilde{g}_M}{(1-\tilde{g}_M^2)^{1/2}} \,, & g_{2VW} &= \frac{g_2 \tilde{g}_M^2}{1-\tilde{g}_M^2} \,, \end{aligned} \tag{6.16}$$

where g_2 denotes the $SU(2)_L$ gauge coupling. The heavy vector triplet couples to the weak gauge bosons not only via the $g-$couplings in Eq. (6.15), but also through the non-abelian component of the covariant derivative Eq. (6.11), that leads to interaction terms of the form $(\partial V)VW$ and $VVWW$. These interactions are weighted by the weak gauge coupling, and therefore are present even if $g_i\,(\tilde{g}_i) \equiv 0$.

6.2.4 SFitter setup

The SFitter framework [241] has been long employed for global analyses of LHC measurements in the context of Higgs couplings and EFTs [178, 242–246], including a comprehensive study of an analysis in terms of Higgs couplings and its UV-completion [247]. The approach is unique in that it allows a comprehensive treatment of uncertainties: SFitter uses a likelihood set up that includes a broad set of statistical, systematic, and theory uncertainties. Statistical and most systematic ones are described by a Poisson- or Gauss-shaped likelihood. Theoretical uncertainties lack a frequentist interpretation, and

are described by flat likelihoods in SFitter, corresponding to a range of equally likely theory predictions. An important difference between employing a flat likelihood compared to a Gaussian one is that the uncorrelated profile likelihood adds the uncertainties from the flat distributions linearly, while Gaussian error bars are added in quadrature. The profile likelihood combination of a flat and a Gaussian uncertainty gives the well-known RFit prescription [248]. Correlations among certain classes of systematic uncertainties are also included.

From the technical point of view, the new aspect of the SFitter analysis presented in this paper is the translation of the SMEFT likelihood into the parameter space of the UV model. In the fit, all observables are parameterized in the SMEFT using the operator set provided in Tab. B.1, that is based on the HISZ basis [249]. All SMEFT predictions are at LO in QCD and scaled by the same corrections as the SM-rates used for the actual experimental analysis. Terms obtained from squaring amplitudes with one operator insertion, that are quadratic in the Wilson coefficients, are retained. The Wilson coefficients are then expressed in terms of $\tilde{g}_i$ parameters of the UV model, Eq. (6.10), using the one-loop matching expressions onto the Warsaw basis provided in Ref. [250] and the Warsaw-to-HISZ basis translation in Appendix B.1. In this way, the likelihood can be directly sampled in the parameter space of the UV model.

In addition, we employ a new likelihood sampling method [5] compared to previous SFitter analyses, that ensures a much more efficient sampling close to the SM point, where all Wilson coefficients vanish. By contrast, the previous sampling method was optimized for the detection of potential secondary maxima in the likelihood, by giving higher weight to the edges of the parameter space.

Dataset

The SMEFT analysis presented in this work builds directly on the dataset employed in Ref. [178], which includes electroweak precision observable (EWPO) at LEP (14 measurements), Higgs measurements (275) and di-boson measurements at the LHC (43). The latter contain results from both Run 1 and Run 2 [245]. In addition, we include differential measurements from three resonance searches by ATLAS, that reach up to invariant masses in the multi-TeV range and that we re-interpret within the SMEFT framework. One of these [251] was already included in the analysis of Ref. [178]. The other two [252,253] are more recent and have been added specifically for this work. These measurements are not usually included in the SMEFT analyses and are not covered by the simplified template cross section framework [254]. Nevertheless, it can be instructive to explore their sensitivity, particularly to operators that induce momentum-enhanced corrections. Moreover, all the resonance searches considered here target heavy vector triplets decaying into WH or WW as a potential signal. Therefore they allow to compare directly the constraining power of the SMEFT analysis to that of the direct search.

Theory uncertainties

In view of the upcoming LHC runs and their rapidly growing data sets, the treatment of theory uncertainties in global analyses is becoming critical. In our analysis, we include theory uncertainties associated to parton distribution functions, to missing higher orders in the SM or SMEFT predictions, and to the matching scale to the EFT. The latter will be discussed in more detail in Sec. 6.3.2.

For the time being, we do not include uncertainties associated to missing SMEFT operators due to the truncation of the SMEFT Lagrangian [255] or to symmetry assumptions, such as CP-conservation. Nevertheless, the impact of missing higher orders in the EFT expansion becomes obviously manifest in the comparison between constraints extracted from the SMEFT analysis and from direct searches.

Concerning higher orders in the loop expansion, Higgs analyses in SFitter currently adopt the most accurate SM predictions available, which are implemented so as to match the state-of-the art predictions reported in the experimental analyses. The corresponding K-factors are then applied onto the tree-level SMEFT predictions as well, which is tantamount to assuming that QCD corrections scale evenly for all SMEFT operators and in the same way as in the SM. Although this assumption is, strictly speaking, not correct [256], for the rate measurements considered here we do not expect large variations in the K-factors between different operators. For some kinematic distributions these effects can be larger. We therefore assign conservative theory uncertainties in order to reduce the numerical impact of these effects. A proper SMEFT simulation of Higgs and di-boson production up to NLO in QCD is postponed to a future work.

6.3 Toy fits and matching uncertainty

In this section we discuss two aspects of the vector triplet model and of its matching onto the SMEFT, that are preliminary to a correct SMEFT global analysis. The first is the decoupling limit of the model, and the second is the numerical impact of varying the scale at which the 1-loop matching is performed. Both issues are analyzed via simplified toy fits.

6.3.1 Decoupling

The decoupling limit of the vector triplet model considered in this work is most easily identified starting from the Lagrangian of Eq. (6.15), where, as long as the EW symmetry is unbroken, the heavy triplet and the SM gauge bosons do not mix. In this case, it is easy to see that the BSM states decouple for large values of the physical mass, $m_V \to \infty$. This is directly reflected in the matching formulas, which give $\lim_{m_V\to\infty} C_i \equiv 0$ for all the dimension-6 Wilson coefficients. At the level of a global fit, the decoupling limit can be visualized by setting the central values of all the measurements included to match the corresponding SM predictions. Figure 6.1 shows the results obtained in this way from SFitter: the likelihood is first computed as a function of 4 free parameters in the Lagrangian of Eq. (6.10)

$$\{\tilde{m}_V, \tilde{g}_M, \tilde{g}_H, \tilde{g}_l\}\,, \tag{6.17}$$

setting other parameters $\tilde{g}_q$, $\tilde{g}_{VH}$ to zero. We then project them onto the 7 parameters for the rotated Lagrangian of Eq. (6.15),

$$\{m_V, g_H, g_l, g_q, g_{VH}, g_{3V}, g_{2VW}\}\,. \tag{6.18}$$

At this stage, we fix the matching scale to $Q = m_V = 4\,\text{TeV}$. For each of the couplings we see that, as expected, the range of allowed values increases as $m_V^{-1} \to 0$. It is worth noting that the rate at which this happens varies between the g-parameters. This is due to the

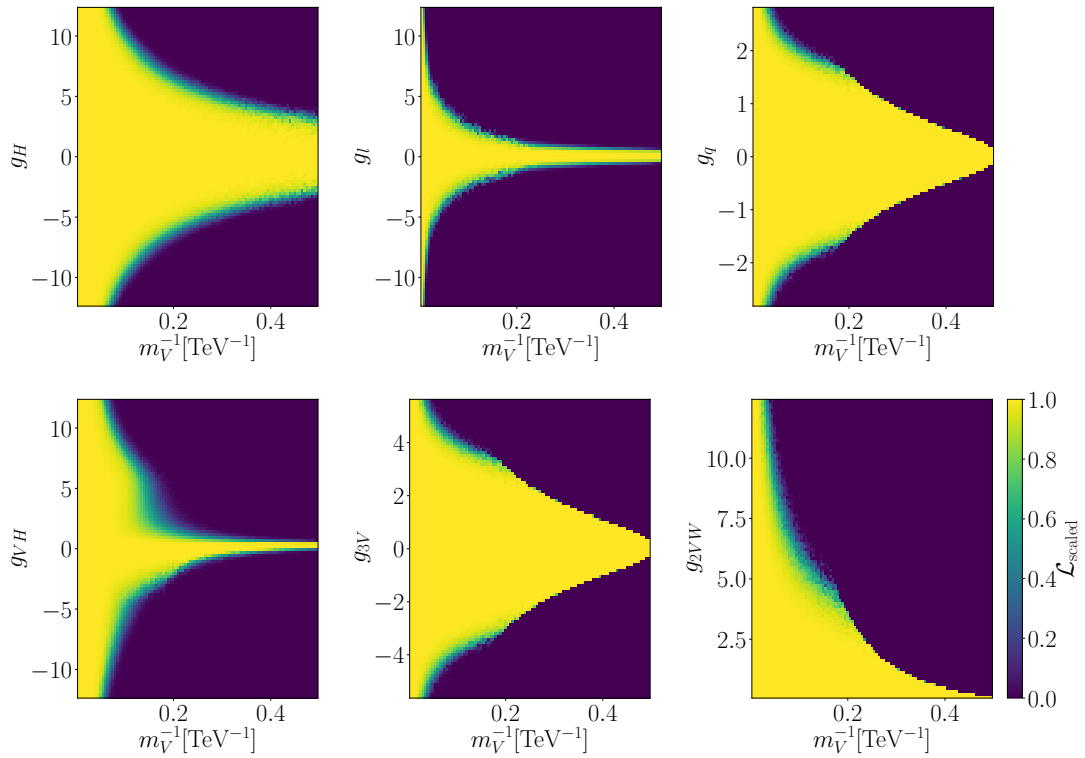


Figure 6.1: Decoupling pattern for the vector triplet model. Global fit with all measurements at their SM values and to the 4 free parameters $\tilde{m}_V, \tilde{g}_M, \tilde{g}_H, \tilde{g}_l$, and subsequently projected onto the 7 parameters of the unmixed Lagrangian of Eq. (6.15).

fact that the matching expressions do not scale homogeneously with g_i^2/m_V^2, but generally have a more complex polynomial structure. The degeneracy between g_i and $1/m_V$ in these expressions is also broken by the $V-W$ interactions proportional to the weak gauge coupling. The homogeneity of the yellow regions indicates that there the likelihood is flat and no point is preferred. Setting all measurements to their actual measured values, which generally depart from the SM predictions, has the effect of introducing a substructure in the likelihood, thereby identifying a more restricted preferred region. This is shown, for a subset of panels, in Fig. 6.2. Here, for instance, the best fit point moves to finite m_V and prefers non-vanishing values of g_H. Note that, to good approximation, the entire region highlighted in green is allowed at 68%CL. The yellow points simply identify a best-fit region and should not be interpreted as statistically significant. Finally, the reduced number of parameters in the Lagrangian Eq. (6.10) as compared to the setup without kinetic mixing induces strong correlations through $\tilde{g}_M$, as illustrated in the $g_{2VW} - g_{3V}$ plane of Fig. 6.2.

As the matching procedure highlighted in Sec. 6.2.1 requires a separation between light and heavy degrees of freedom, defining the decoupling limit in the notation of Eq. (6.10) requires some more care, due to the explicit kinetic mixing between the heavy triplet and the SM gauge fields.

From Eq. (6.16), we see that $m_V \to \infty$ can be achieved for $\tilde{m}_V \to \infty$ or for $|\tilde{g}_M| \to 1$. However, the condition $|\tilde{g}_M| = 1$ does not lead to a well-defined decoupling condition, because in this limit $\tilde{V}_\mu^A$ become auxiliary fields, i.e. the theory loses three dynamical

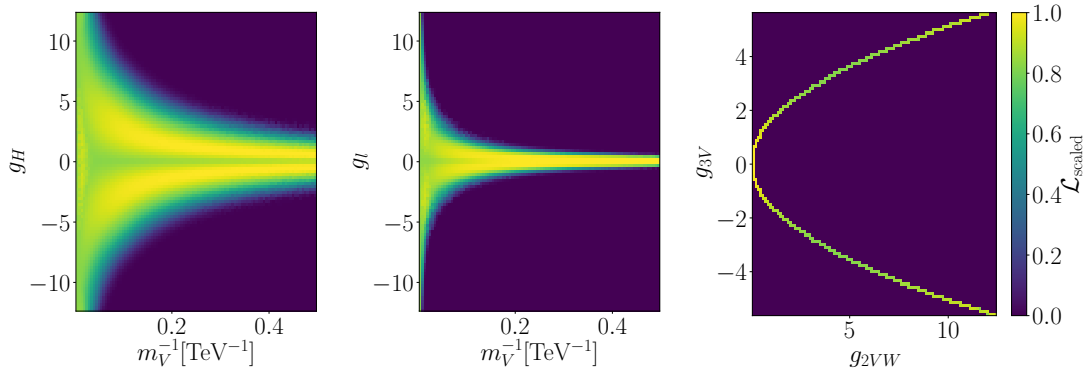


Figure 6.2: Results of the same global analysis as in Fig. 6.1, but with measurements set to their actual values.

degrees of freedom. This is not sufficient for a proper decoupling in the EFT sense because even as an auxiliary field $\widetilde{V}_\mu^A$ still induces mass-suppressed vertices that enter correlation functions and we enter a strongly interacting regime where our perturbative approach fails.

To see the impact of $\tilde{g}_M$ we resum insertions of gauge mixing into the $\widetilde{W}_\mu^A$ and $\widetilde{V}_\mu^A$ propagators. The corrected propagators of these fields become

$$\begin{aligned}\hat{D}^{\widetilde{V}}_{\mu\nu} &= -\frac{i}{p^2-\tilde{m}_V^2-\tilde{g}_M^2 p^2}\left(g_{\mu\nu}-(1-\tilde{g}_M^2)\frac{p_\mu p_\nu}{\tilde{m}_V^2}\right)\\ \hat{D}^{\widetilde{W}}_{\mu\nu} &= -\frac{i}{p^2}\left(g_{\mu\nu}-(1-\xi)\frac{p_\mu p_\nu}{p^2}\right)-\frac{i\tilde{g}_M^2}{p^2-\tilde{m}_V^2-\tilde{g}_M^2 p^2}\left(g_{\mu\nu}-\frac{p_\mu p_\nu}{p^2}\right)\,. \end{aligned} \tag{6.19}$$

It is easy to see that for $|\tilde{g}_M|=1$ the resummed $\widetilde{V}_\mu^A$ propagator loses its momentum dependence, which is indicative of the field becoming auxiliary. For $|\tilde{g}_M|>1$, $\widetilde{V}_\mu^A$ becomes tachyonic while, for $|\tilde{g}_M|<1$, $\widetilde{V}_\mu^A$ is a dynamical degree of freedom. In this case its propagator has a physical pole at $p^2=m_V^2$ as defined in Eq. (6.16), and it can be expanded in $p^2/m_V^2\ll 1$. The resummed $\widetilde{W}_\mu^A$ propagator includes a term with a pole at $p^2=m_V^2$, contaminating $\widetilde{W}_\mu^A$ with a contribution from $\widetilde{V}_\mu^A$. Therefore this field cannot be directly identified with the SM weak bosons. However, in the tree-level matching procedure, once the 1LPI effective action is expanded in p^2/m_V^2, the component associated with the $\widetilde{V}_\mu^A$ pole is shifted from the propagators to the interaction terms, which are unambiguously fixed at this order by the matching condition of Eq. (6.2). At one loop, the fact that the EFT is the low-energy limit of the UV model is manifest in the fact that only the "hard" region of the momentum integral contributes to the functional trace in the matching formula of Eq. (6.6). As a consequence, the first term of the $\widetilde{W}_\mu^A$ propagator cancels against the corresponding EFT contributions, while the second term genuinely contributes to the matching in the hard region. Equivalently, one can match in the shifted basis directly identifying W_μ^A in the UV model with the corresponding weak bosons in the SMEFT.

In the top (bottom) panels of Fig. 6.3 we again show the results of a global analysis where all measurements are set to their SM prediction (to their actual values), this time projected onto a subset of the $\tilde{g}$-parameters and onto the combination $\tilde{g}_M/\sqrt{1-\tilde{g}_M^2}$ that drives most $\tilde{g}-g$ relations, see Eq. (6.16). For reference, the right panels also show lines

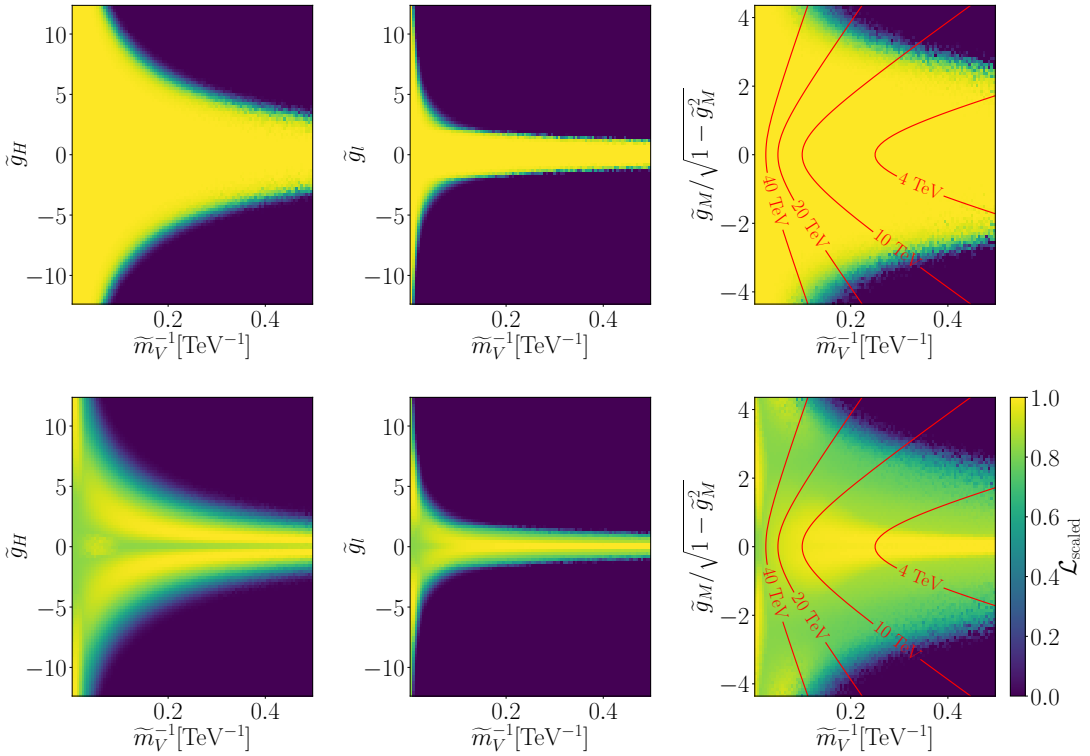


Figure 6.3: Results of the same global analyses as in Fig. 6.1 (upper) and Fig. 6.2 (lower), projected on $\tilde{m}_V, \tilde{g}_H, \tilde{g}_l$ and the combination $\tilde{g}_M/\sqrt{1-\tilde{g}_M^2}$.

of constant m_V, such that the decoupling limit $m_V \to \infty$ flows orthogonally to the lines. Consistent with the results in the unmixed basis (Fig. 6.1), the expected likelihood is mostly flat in the entire preferred region, while the observed one exhibits a substructure that identifies a best-fit region where $\tilde{g}_H \neq 0$ and both m_V and $\tilde{m}_V$ are finite. The reason can be identified in a few EWPO measurements that exhibit small ($< 1\sigma$) deviations from the SM expectation: $A_l(SLD)$ and m_W.

For $|\tilde{g}_M| \to 1$ the theory becomes strongly interacting and some perturbative unitarity considerations are therefore pertinent. Requiring the couplings of the unmixed UV theory to remain perturbative, the most stringent constraints on $\tilde{g}_M$ stem from g_{2VW}

$$g_{2VW} \approx \frac{g_2 \tilde{g}_M^2}{1-\tilde{g}_M^2} < 4\pi \qquad \Leftrightarrow \qquad |\tilde{g}_M| < 0.975\ . \tag{6.20}$$

Therefore for all our fits we require $|\tilde{g}_M| < 0.975$.

6.3.2 Matching scale

In perturbative predictions of LHC observables, at least two unphysical scales are known to reflect a theory uncertainty, the factorization scale and the renormalization scale. Both arise from a separation of an observable into different regimes with different perturbative expansions, and the scale dependence would vanish if we would include all orders in all predictions. For a calculation at finite perturbative order we instead use the scale

variation as one measure of a theory uncertainty and treat it as an unphysical nuisance parameter in theory predictions [246, 257].

One unphysical scale is the renormalization scale, which in the context of dimensional regularization appears as a free parameter. In more physical terms, the renormalization scale is the energy scale associated with those observables that we select for defining the numerical parameters of the theory, the renormalization conditions. Whenever scale choices are arbitrary, we often identify them with each other and a typical energy scale of the scattering process to avoid large logarithms. Clearly, this does not work if renormalization conditions involve widely distinct energy scales, such as in the relation of UV-model parameters to the low-energy observables of the SM.

The renormalization group equation apparently solves this problem. It relates observables at different scales, properly resumming logarithms and absorbing them into running parameters. However, it works only in the absence of mass thresholds. This strongly suggests to match a UV model with a heavy mass M to a low-energy EFT even if the algebraic simplifications of the latter are not essential for a specific calculation.

In a one-loop matching calculation that uses dimensional regularization, the matching scale enters as an additional parameter. However, in contrast to the original renormalization scale this parameter is not entirely arbitrary. If we want to avoid large logarithms, its reasonable range is bounded from above and below. In line with the generic discussion of one-loop matching above, we illustrate this property in the following section. We consider examples of increasing complexity, starting from the QCD coupling, turning next to the SM extended by a scalar singlet and finally returning to the vector triplet model of Sec. 6.2.3.

Running strong coupling

We can illustrate the appearance of the matching scale using the simple example of the running strong coupling. It provides the key ingredients to understanding the EFT matching scale: the separation of low-energy and UV regimes and contributions beyond tree level. In general, the relation between the bare coupling and the renormalized coupling in the $\overline{\text{MS}}$ scheme is

$$\alpha_s^{\text{bare}} = \alpha_s(p^2)\left[1 - \alpha_s b_0\left(\frac{1}{\bar{\epsilon}} + \log\frac{\mu_R^2}{p^2}\right)\right] \quad \text{with} \qquad b_0^{(n_f)} = \frac{1}{4\pi}\left(\frac{11}{3}N_c - \frac{2}{3}n_f\right) . \tag{6.21}$$

Here, p^2 is the energy scale of the scattering, μ_R^2 is introduced by dimensional regularization, and $1/\bar{\epsilon} = 1/\epsilon - \gamma_E + \log 4\pi$. We identify our UV-regime as momenta above the top mass, with six propagating quark flavors, and the low-energy regime as described by five propagating quark flavors. The running of α_s in the two regimes is described by the beta function with five or six flavors, respectively. The UV-divergences in the low-energy and full UV-theories arise from five or six propagating flavors, so the renormalization prescription Eq. (6.21) is different in the two regimes.

The low-energy and UV-regimes are separated by a matching scale Q, which we choose to be of the order of the top mass to avoid large logarithms or inconsistent symmetry structures. Matching conditions guarantee that the two predictions for any observable are the same at least at this scale. Instead of looking at a full set of amplitudes or correlation functions, we limit ourselves to the quasi-observable α_s. Following Eq. (6.21),

the definitions of $\alpha_s(p^2)$ in relation to the bare parameter are different, but they have to agree when evaluated at the matching scale. This defines a threshold correction

$$1 - \frac{\alpha_s b_0^{(6)}}{4\pi}\left(\frac{1}{\bar\epsilon} + \log\frac{\mu_R^2}{p^2}\right)\bigg|_{Q^2} = 1 - \frac{\alpha_s b_0^{(5)}}{4\pi}\left(\frac{1}{\bar\epsilon} + \log\frac{\mu_R^2}{p^2}\right)\bigg|_{Q^2} + \frac{\alpha_s}{6\pi}\log\frac{\mu_R^2}{Q^2}\,. \tag{6.22}$$

The relation of the threshold correction to loop effects is reflected in the logarithmic form $\log\mu_R^2/Q^2$. Together with the five-flavor $\overline{\text{MS}}$ counter term it defines α_s in the low-energy regime as

$$\alpha_s^{\text{bare}} = \alpha_s(p^2)\left[1 - \frac{\alpha_s b_0^{(5)}}{4\pi}\left(\frac{1}{\bar\epsilon} + \log\frac{\mu_R^2}{p^2}\right) + \frac{\alpha_s}{6\pi}\log\frac{\mu_R^2}{Q^2}\right]\,. \tag{6.23}$$

This definition includes three scales for a given scattering process, the physical scale p^2, the renormalization scale μ_R^2, and the matching scale Q^2. In simple problems, the renormalization scale and the physical scale can be identified to avoid potentially large logarithms. The matching scale is usually set to the mass of the decoupled particle, $Q = m_t^2$, leading to a threshold correction that is non-zero in general.

From our toy example we can immediately see the role of the threshold correction at the matching scale and the renormalization group running. If we start from the UV, all parameters of the theory evolve based on the full particle spectrum. In the low-energy theory part of the spectrum decouples also from the running, which can even break the underlying symmetries [258], and we will follow a completely different renormalization group flow. The matching corrections adjust for this effect. They move us to the same flow line in the EFT, independent of the choice of matching scale and with all the caveats of maintaining perturbative control, accounting for changes of the spectrum, changing symmetries, etc.

Singlet extension

When we interpret a SMEFT calculation for an LHC process as a low-energy approximation to a UV-prediction, we again break the phase space of the scattering process into two parts. We first illustrate SMEFT matching using the singlet-extended SM [259, 260],

$$\mathcal{L} \supset \frac{1}{2}\left(\partial_\mu S\right)\left(\partial^\mu S\right) - \frac{1}{2}M^2S^2 - A|\phi|^2 S - \frac{\kappa}{2}|\phi|^2S^2 - \frac{\mu}{3!}S^3 - \frac{\lambda_S}{4!}S^4\,. \tag{6.24}$$

The singlet mass is given by $M_S^2 = M^2 + \mathcal{O}(v^2)$; we integrate it out under the condition $M_S \sim M \gg v$, ensuring a consistent expansion in v/M [240]. As a simplification, we also assume A to be of the order of M. The leading term in v/M is defined by $v = 0$ and can be obtained by matching in the unbroken phase. In the broken phase the Higgs VEV enters via the masses of the SM-particles which properly belong to the EFT Lagrangian, below the matching scale. Matching in the broken phase would allow us to include partial higher-order corrections in the EFT expansion [194]. Since the mass scales in question are not widely separated, it depends on the detailed numerics which setup yields a more reliable approximation. The SMEFT Lagrangian reads

$$\mathcal{L}_{\text{SMEFT}} = \mathcal{L}_{\text{SM}} + \sum_i f_i(p/\mu_R)\,\mathcal{O}_i\,, \tag{6.25}$$

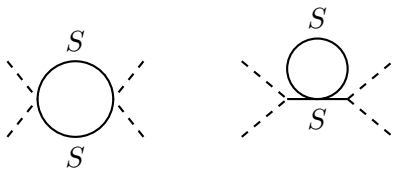
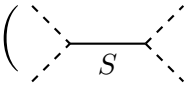


Figure 6.4: Feynman diagrams contributing to $f_{\phi,2}^{(1)}$. Left: Diagram yielding a κ^2-contribution. Right: Diagram yielding a $A^2\lambda_S/M^2$-contribution. The dashed line corresponds to the Higgs field, whereas the solid line corresponds to the singlet.

where the Wilson coefficients are scale dependent. Specifically, we want to define these coefficients such that the SMEFT reproduces all low-energy observables of the UV-theory up to $\mathcal{O}(v^3/M_S^3)$. As matching condition we use Eq. (6.1). In the functional approach we compute this once and for all using functional traces. To illustrate some features related to the matching scale, we compute some contributions to the Wilson coefficient $f_{\phi,2}$ of the operator $\mathcal{O}_{\phi,2} = \partial_\mu(\phi^\dagger\phi)\partial^\mu(\phi^\dagger\phi)/2$ diagrammatically. As discussed in Appendix B.1, it is related to $Q_{\phi\Box} = |\phi|^2\Box|\phi|^2$ as $c_{\phi\Box} \approx -f_{\phi,2}/2$, modulo fermionic operators. The operator contributes to the correlation function with two external fields ϕ and two external fields $\phi^\dagger$ and depends on p^2, so we fix it by requiring

$$\partial_{p^2}\Gamma_{\text{SMEFT}}(\phi^\dagger,\phi^\dagger,\phi,\phi)\Big|_{p^2=0} = \partial_{p^2}\Gamma_{\text{L,UV}}(\phi^\dagger,\phi^\dagger,\phi,\phi)\Big|_{p^2=0}\,, \tag{6.26}$$

order by order in the coupling. With some abuse of notation we also denote specific correlation functions by Γ, arguments indicating the external fields. Since both sides of the equation involve running parameters, the matching has to be imposed at a given scale,

$\partial_{p^2}\Big($  $+\,\text{SM}\Big) = \partial_{p^2}\Big($  $+\,t\text{-channel} + \text{SM}\Big)$ at $p^2 = 0$.

The SM-contributions contain the same diagrams on both sides, with appropriately adjusted parameters through the matching conditions, so their contributions cancel. Only diagrams with at least one heavy propagator actually contribute to the matching, so Eq. (6.26) becomes

$$\partial_{p^2}\left(8p^2 f_{\phi,2}^{(0)}\right)\Big|_{p^2=0} = \partial_{p^2}\frac{2A^2}{4p^2-M^2}\Big|_{p^2=0} \qquad \Rightarrow \qquad f_{\phi,2}^{(0)} = \frac{A^2}{M^4}\,. \tag{6.27}$$

At tree level, the scale dependence only appears implicitly for A and for $f_{\phi,2}^{(0)}$.

Next, we compute the κ^2-contribution to $f_{\phi,2}^{(1)}$ at one loop. This contribution is induced by the diagram on the left in Fig. 6.4, where the external particles are as specified in Eq. (6.26). We again set all external scales to p^2 and find for the diagram

$$\kappa^2\mu_R^{4-d}\int\frac{d^dq}{(2\pi)^d}\frac{1}{((2p+q)^2-M^2)(q^2-M^2)} = \kappa^2\frac{i}{16\pi^2}B_0(4p^2,M,M)$$
$$\text{with}\quad B_0(4p^2,M,M) = \frac{1}{\bar\epsilon} - \log\frac{M^2}{\mu_R^2} + \frac{2p^2}{3M^2} + \mathcal{O}\left(\frac{p^4}{M^4}\right)\,. \tag{6.28}$$

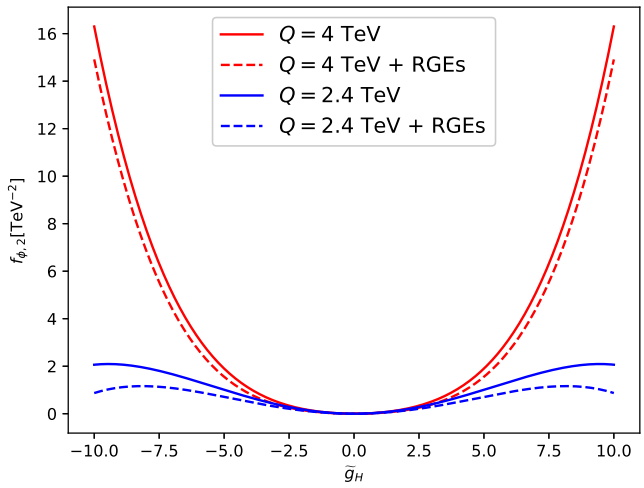


Figure 6.5: Wilson coefficient $f_{\phi 2}$ as a function of $\tilde{g}_H$ at different values of the matching scale Q for fixed $m_V = 4$ TeV and all other UV couplings set to zero. The dashed lines include approximate RG running.

In the full expression the renormalization scale appears, but taking the derivative in the matching condition for this contribution to $f_{\phi,2}$ removes it,

$$\partial_{p^2} B_0(4p^2, M, M)\Big|_{p^2=0} = \frac{2}{3M^2} \qquad \Rightarrow \qquad f_{\phi,2}^{(1)} \supset \frac{1}{16\pi^2}\frac{\kappa^2}{12M^2} \,. \tag{6.29}$$

Just as at tree level, the matching scale does not appear explicitly.

Finally, we compute the $A^2\lambda_S/M^2$-contribution to $f_{\phi,2}^{(1)}$ to illustrate the appearance of matching scale logarithms. This contribution arises from the diagram on the right in Fig. 6.4. The diagram is not 1PI, but is 1LPI and therefore has to be included in the matching. With all external scales again set to p^2 this diagram gives

$$-\frac{\lambda_S A^2}{(4p^2 - M^2)^2}\mu_R^{4-d}\int \frac{\mathrm{d}^d q}{(2\pi)^d}\frac{1}{q^2 - M^2} = -\frac{\lambda_S A^2}{16\pi^2}\frac{M^2}{(4p^2 - M^2)^2}\left(\frac{1}{\bar{\epsilon}} + 1 - \log\frac{M^2}{\mu_R^2}\right) . \tag{6.30}$$

Taking the derivative with respect to p^2 and evaluating it at $p^2 = 0$ we find the one-loop matching condition

$$f_{\phi,2}^{(1)} \supset -\frac{1}{16\pi^2}\frac{\lambda_S A^2}{M^4}\left(-1 + \log\frac{M^2}{Q^2}\right) , \tag{6.31}$$

where the Wilson coefficient explicitly depends on the matching scale. This scale dependence is expected since the corresponding correlation function is divergent. As mentioned before, in models with one new mass scale, we can of course avoid these logarithms by identifying $Q = M$.

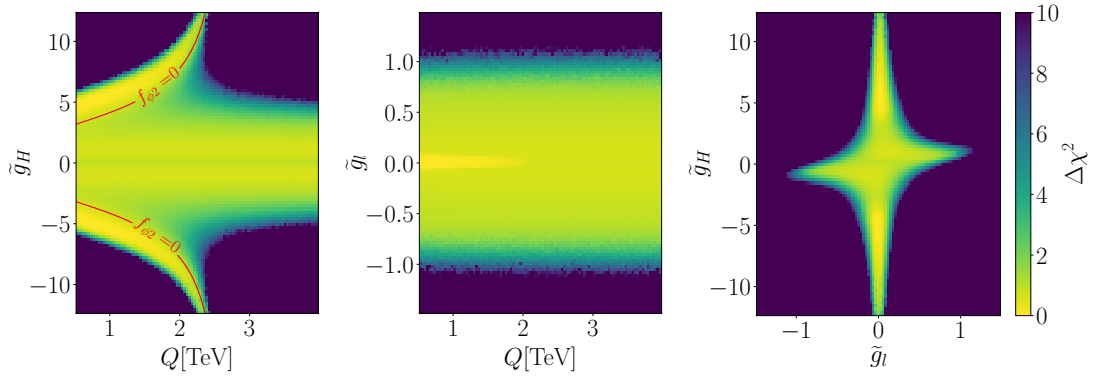


Figure 6.6: The impact of the variation of the matching scale Q at a mass of $m_V = 4\,\text{TeV}$ for a reduced model with free $\tilde{g}_M, \tilde{g}_H, \tilde{g}_l$, expressed in the unmixed Lagrangian Eq. (6.15) with actual measurements.

Vector triplet

Moving to the triplet model defined by the Lagrangian of Eq. (6.10), we will not attempt to show analytic results and instead illustrate the matching scale dependence for one finite coupling $\tilde{g}_H$ and a mass term $\tilde{m}_V$ numerically. In this simplified setup, $m_V = \tilde{m}_V$. Among the various Wilson coefficients, it is instructive to consider $f_{\phi,2}$, as its dependence on the matching scale exhibits interesting features. Including both tree and loop contributions, the matching expression has the form

$$\frac{f_{\phi,2}}{\Lambda^2} \simeq \frac{1}{m_V^2}\left[g_2^4\left(c_0 + c_1 \log\frac{m_V}{Q}\right) + \tilde{g}_H^2\left(c_2 + c_3 \log\frac{m_V}{Q}\right) + \tilde{g}_H^4\left(c_4 + c_5 \log\frac{m_V}{Q}\right)\right], \tag{6.32}$$

where $c_0 = c_1/2$ emerges from 1-loop diagrams inducing the operator structure $(D_\mu W^{\mu\nu})^2$, which maps to $\mathcal{O}_{\phi,2}$ via the equations of motion. Of the additional constants, the $\tilde{g}_H^2$-coefficient is dominated by the tree-level contribution to c_2, while the $\tilde{g}_H^4$-coefficient is completely determined by the one-loop matching. Numerically, we find

$$\begin{aligned} &c_0 = \frac{c_1}{2} = \frac{3}{128\pi^2} = 0.0024\,, \\ &c_2 = 0.75\,, \qquad c_3 = 0.0069\,, \qquad c_4 = 0.019\,, \qquad c_5 = -0.045\,. \end{aligned} \tag{6.33}$$

In Fig. 6.5 we show the numerical dependence of $f_{\phi,2}$ on $\tilde{g}_H$ for different choices of Q. For $Q = m_V = 4$ TeV the Wilson coefficient has a simple power dependence on $\tilde{g}_H$ driven by c_4. For $Q \approx 0.66\, m_V = 2.6\,\text{TeV}$ the $\tilde{g}_H^4$-term cancels exactly. For Q below this threshold, the coefficient in front of $\tilde{g}_H^4$ becomes negative, which flips the sign of $f_{\phi,2}$ at $\tilde{g}_H \gg 1$ and allows a solution of $f_{\phi,2} = 0$ for $\tilde{g}_H \neq 0$. For $Q \lesssim 2.4\,\text{TeV}$ the solution is within the range $|\tilde{g}_H| < 4\pi$ and leads to visible effects in our global analysis.

Figure 6.6 shows the results of the same global analysis as in Sec. 6.3.1, where now we fix $m_V = 4\,\text{TeV}$. The free parameters are

$$\{\tilde{g}_H, \tilde{g}_l, \tilde{g}_M, Q\}\,, \tag{6.34}$$

where the matching scale is varied in the range $Q = 500\,\text{GeV}$ ... $4\,\text{TeV}$. The left panel shows a central allowed region for $|\tilde{g}_H| \lesssim 4$ that is independent of Q. In addition, a

beautiful *fleur-de-lis shape* arises in $\tilde{g}_H$ vs Q for $Q < 2.4$ TeV. It roughly follows the curves along which $f_{\phi,2} = 0$ marked in red. The Wilson coefficients f_t, f_b, f_τ have a similar behavior and vanish approximately in the same region, because they are induced by the same or similar loop contributions. As these are the operators that dominate the constraint on $\tilde{g}_H$, the fleur-de-lis feature persists in the full global fit, see Sec. 6.4. When we profile over Q as a nuisance parameter, this correlation broadens the 1-dimensional and 2-dimensional profile likelihood in $\tilde{g}_H$ by roughly a factor 2. As shown in the second and third panels of Fig. 6.6, the broadening affects significantly only the constraints in the $\tilde{g}_H$ direction, while those on $\tilde{g}_l$ are essentially unchanged compared to when $Q = m_V$. Although not shown, this is also verified for $\tilde{g}_M$.

We emphasize that the tree-loop cancellations that drive this effect are only very slightly affected by the renormalization group evolution of $f_{\phi,2}$, as illustrated by the dashed lines including approximate RGE contributions in Fig. 6.5. They really correspond to a choice of the unphysical matching scale, which cannot be compensated by the well-defined change of renormalization scale of the low-energy SMEFT description. Adding higher orders in the loop expansion to the matching decreases the sensitivity to the matching scale. Similar effects, but with a much smaller numerical impact have been observed in Ref. [259].

6.4 SMEFT global analysis

In this section we discuss the results of the SMEFT global analysis, mapped to the parameter space of the heavy vector triplet model defined in Sec. 6.2.3 using 1-loop matching relations. We derive constraints on the UV-parameters $\{\tilde{g}_H, \tilde{g}_q, \tilde{g}_l, \tilde{g}_M, \tilde{g}_{VH}\}$ defined by the Lagrangian in Eq. (6.10) for fixed values of the heavy vector triplet mass. We consider two benchmark values: $m_V = 4\,\text{TeV}$, to be compared with direct resonance searches by the ATLAS Collaboration, and $m_V = 8\,\text{TeV}$ for a consistent SMEFT analysis safely below any on-shell pole.

6.4.1 Resonance searches at high invariant masses

As mentioned in Sec. 6.2.4, in addition to more standard Higgs measurements, the global analysis includes constraints from searches for exotic particles in the WH and WW channels by the ATLAS Collaboration. In particular, two of these analyses [252, 253] have been newly implemented in SFitter.

WH search

We consider the m_{WH} invariant mass distribution measured in Ref. [252] in the WH 1-tag category, and we compare it to a WH signal including dimension-6 corrections. This kinematic distribution extends up to $m_{WH} = 5\,\text{TeV}$ and the strongest constraints on BSM effects stem from the region around $m_{WH} = 2 - 2.5\,\text{TeV}$, where the measurement exhibits large under-fluctuations. A detailed description of the implementation of this analysis will be provided in a future work [5].

For equal values of the Wilson coefficients, the largest correction to the m_{WH} spectrum is induced by the operator $\mathcal{O}_{\phi Q}^{(3)}$ [254, 261–266], that contributes via corrections to the

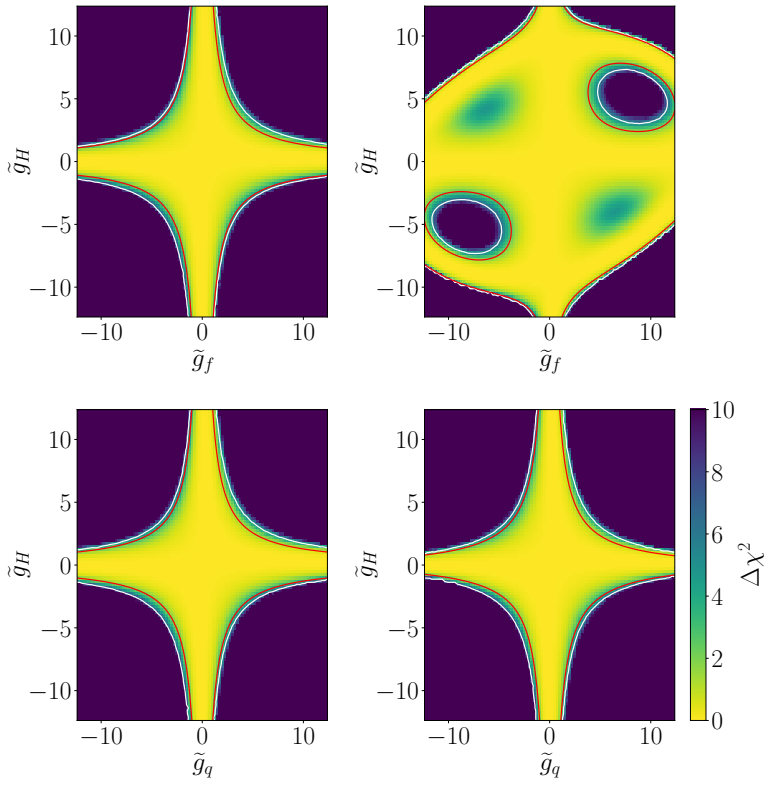


Figure 6.7: 2D fits of the WH resonance search of Ref. [252] only. We fix $m_V = 4\,\text{TeV}$ and $\tilde{g}_M = \tilde{g}_{VH} = 0$. Left: tree-level matching. Right: Loop-level matching. Top: with $\tilde{g}_l = \tilde{g}_q = \tilde{g}_f$. Bottom: with $\tilde{g}_l = 0$. In the top (bottom) row, red contours indicate $f_W = \pm 4$ ($f_{\phi Q}^{(3)} = \pm 0.8$) with $\Lambda = 1\,\text{TeV}$ and white contours indicate $\Delta\chi^2 = 5.991$.

qqV vertex and via a 4-point $qqVH$ interaction. The latter exhibits an enhancement at large partonic energies due to the missing s-channel propagator and is therefore dominant in the high-invariant-mass regime. Further significant corrections, albeit less momentum-enhanced, are induced by $\mathcal{O}_W$. All other SMEFT operators in the HISZ basis do not contribute significantly to WH production in the high-energy regime.

Figure 6.7 shows the results from a 2D-analysis of the m_{WH} distribution alone, fixing the matching scale $Q = m_V = 4\,\text{TeV}$ and considering only two $\tilde{g}$-couplings at a time. The top row in Fig. 6.7 shows $\tilde{g}_f \equiv \tilde{g}_q = \tilde{g}_l$ vs $\tilde{g}_H$, which matches the benchmark considered in the ATLAS analysis [252]. In this limit, the matching contribution to $f_{\phi Q}^{(3)}$ cancels exactly, both at tree and loop levels. As a consequence, the constraints are driven by f_W, whose matching expressions reduce to

$$\begin{aligned} \frac{f_W}{\Lambda^2} &= 4.76\,\frac{\tilde{g}_H \tilde{g}_l}{m_V^2} && \text{(tree)} \\ \frac{f_W}{\Lambda^2} &\simeq \tilde{g}_l \tilde{g}_H \frac{4.71 \;+ 0.019\,\tilde{g}_l \tilde{g}_H - 0.023\,\tilde{g}_l^2 - 0.057\,\tilde{g}_H^2}{m_V^2} && \text{(tree+loop).} \end{aligned} \tag{6.35}$$

The red contours in the plots indicate $f_W/\Lambda^2 = \pm 4\,\text{TeV}^{-2}$, which is representative of the 2σ boundaries $f_W/\Lambda^2 \in [-3.6, 4.4]\ \text{TeV}^{-2}$ found in a 1D fit to the SMEFT parameters. In a slight abuse of language, here and in the following the $\Delta\chi^2 \leq 1\,(2.3)$ and $\Delta\chi^2 \leq$

3.841 (5.991) regions in 1D (2D) fits are sometimes referred to as 1σ and 2σ intervals, respectively. The fact that these lines coincide to a very good approximation with the 2σ contours (indicated in white) in Fig. 6.7 shows that the constraint on f_W is indeed the leading one. The bottom row shows $\tilde{g}_q$ vs $\tilde{g}_H$ for $\tilde{g}_l = 0$. In this case the cancellation in $f_{\phi Q}^{(3)}$ is spoiled and the constraints are dominated by this Wilson coefficient. Numerically, the matching expression is

$$\begin{aligned} \frac{f_{\phi Q}^{(3)}}{\Lambda^2} &= \frac{\tilde{g}_H(\tilde{g}_l - \tilde{g}_q)}{m_V^2} && \text{(tree)} \\ \frac{f_{\phi Q}^{(3)}}{\Lambda^2} &\simeq 0.99\, \frac{\tilde{g}_H(\tilde{g}_l - \tilde{g}_q)}{m_V^2} && \text{(tree+loop)}, \end{aligned} \tag{6.36}$$

and the bottom panels in Fig. 6.7 show contours for $f_{\phi Q}^{(3)}/\Lambda^2 = \pm 0.8\,\text{TeV}^{-2}$, which is representative of the 2σ interval $f_{\phi Q}^{(3)}/\Lambda^2 \in [-0.90, 0.76]$ TeV^{-2} obtained in a 1D fit.

Finally, comparing the left and right panels in Fig. 6.7, it is worth noting that the impact of loop contributions to the matching is negligible in the case $\tilde{g}_l = 0$, but significant for $\tilde{g}_l = \tilde{g}_q$. This is a direct consequence of the form of the matching expression in the particular model considered. Loop terms only induce a very minor overall rescaling in the expression of $f_{\phi Q}^{(3)}$, Eq. (6.36), but they introduce a series of new terms in the expression of f_W, Eq. (6.35). Although numerically subdominant, the latter have a strong impact on the likelihood structure.

WW search

We consider the m_{WW} distribution measured in Ref. [253] in the WW 1-lepton category and ggF/DY merged, high-purity signal region, that targets neutral resonances decaying to $W^\pm W^\mp$ pairs and covers invariant masses up to $m_{WW} = 4\,\text{TeV}$. We compare the measured distribution to a $W^\pm W^\mp$ production signal including SMEFT corrections. Again, we postpone a detailed discussion of the implementation to a later paper.

The $W^\pm W^\mp$ production process exhibits a greater complexity in the SMEFT compared to $W^\pm H$ in the high-energy limit. We find that, fixing all Wilson coefficients to the same numerical value, the largest corrections are induced by the operators $\mathcal{O}_{\phi u}, \mathcal{O}_{\phi d}, \mathcal{O}_{\phi Q}^{(1)}, \mathcal{O}_{\phi Q}^{(3)}$ at quadratic level, that exhibit a large enhancement $\propto m_{WW}^2$. The origin of this behavior can be identified as a $qq\phi\phi$ contact interaction between two quarks and two Goldstone bosons induced by these operators, that dominates at high energies due to the equivalence theorem [267]. Effects induced by $\mathcal{O}_W, \mathcal{O}_B$, and $\mathcal{O}_{WWW}$ have a weaker momentum-enhancement and are roughly two orders of magnitude smaller. Nevertheless, they were retained in the fit, as they are relevant for the global analysis in terms of both SMEFT and UV model parameters. In the former case, this measurement contributes significantly to improving the constraints on f_W, by roughly a factor two [5]. In the latter, it is important to stress that the matching expressions for a given UV model generally do *not* give homogeneous values for the Wilson coefficients. Therefore a suppression of two orders of magnitude in the SMEFT predictions can be easily compensated in the matching, and the corresponding contributions to the signal may lead to significant constraints on the UV model parameters. In fact, for the WW analysis implemented here we find that the constraints projected on the $\tilde{g}_q - \tilde{g}_H$ and $\tilde{g}_f - \tilde{g}_H$ planes are entirely dominated by the contributions of f_W and $f_{\phi Q}^{(3)}$, the same two operators that lead in the WH case.

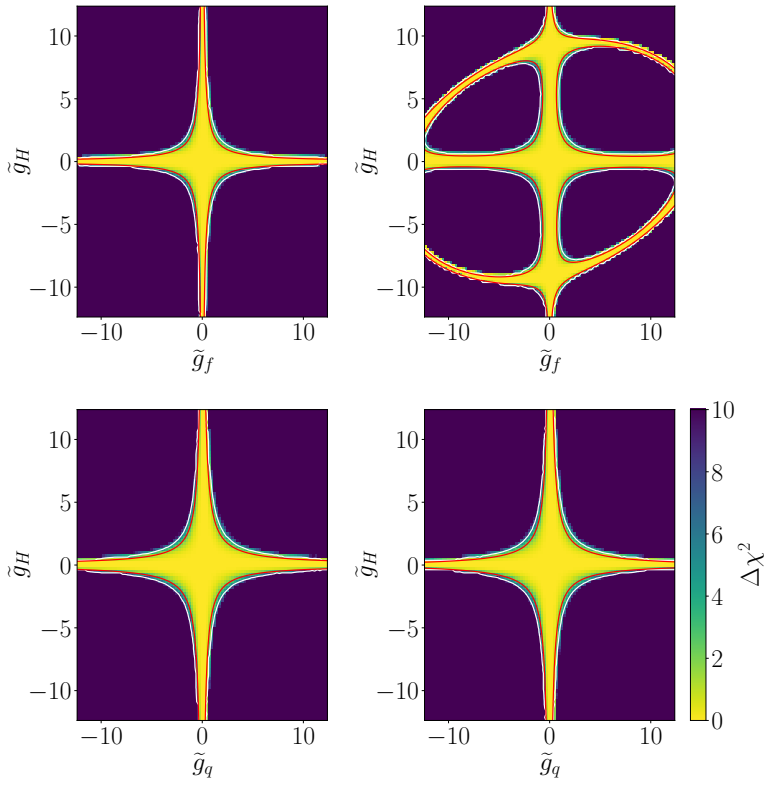


Figure 6.8: 2D fits of the WW resonance search of Ref. [253] only. We fix $m_V = 4\,\text{TeV}$ and $\tilde{g}_M = \tilde{g}_{VH} = 0$. Left: tree-level matching. Right: Loop-level matching. Top: with $\tilde{g}_l = \tilde{g}_q = \tilde{g}_f$. Bottom: with $\tilde{g}_l = 0$. In the top (bottom) row, red contours indicate $f_W = \pm 0.7$ ($f_{\phi Q}^{(3)} = 0.2$ or $f_{\phi Q}^{(3)} = -0.3$) with $\Lambda = 1\,\text{TeV}$ and white contours indicate $\Delta\chi^2 = 5.991$.

Figure 6.8 shows the results from a 2D-analysis of the m_{WW} distribution alone, fixing $Q = m_V = 4\,\text{TeV}$ and considering the same benchmarks as in Fig. 6.7. The red curves in Fig. 6.8 are again given by Eqs. (6.35) and (6.36), but for different values of f_W and $f_{\phi Q}^{(3)}$, namely $f_W/\Lambda^2 = \pm 0.7\,\text{TeV}^{-2}$ and $f_{\phi Q}^{(3)}/\Lambda^2 = -0.27, +0.23\,\text{TeV}^{-2}$. Again, these values correspond to the 2σ-boundaries identified in 1D fits.

This analysis yields stronger bounds compared to WH because in this particular case the constraints are dominated by the tail of the distribution, in the region around $m_{WW} = 2.5 - 4$ TeV, which exhibits under-fluctuations. Again, the effect of introducing loop contributions to the matching expressions is only visible in the scenario dominated by f_W, for the same reasons as described above.

6.4.2 Global analysis results

Figure 6.9 shows the results of our global analysis, including the full data set described in Sec. 6.2.4 as well as the resonance searches discussed in Sec. 6.4.1, for a fixed value of the heavy vector triplet mass $m_V = 4\,\text{TeV}$. The analysis is performed varying $\tilde{g}_M$ and $\tilde{g}_{VH}$ within the physical region $\tilde{g}_M = -1$ ... 1, $\tilde{g}_{VH} > 0$ and all other coupling parameters in the perturbative range $\tilde{g} = -4\pi$ ... 4π.

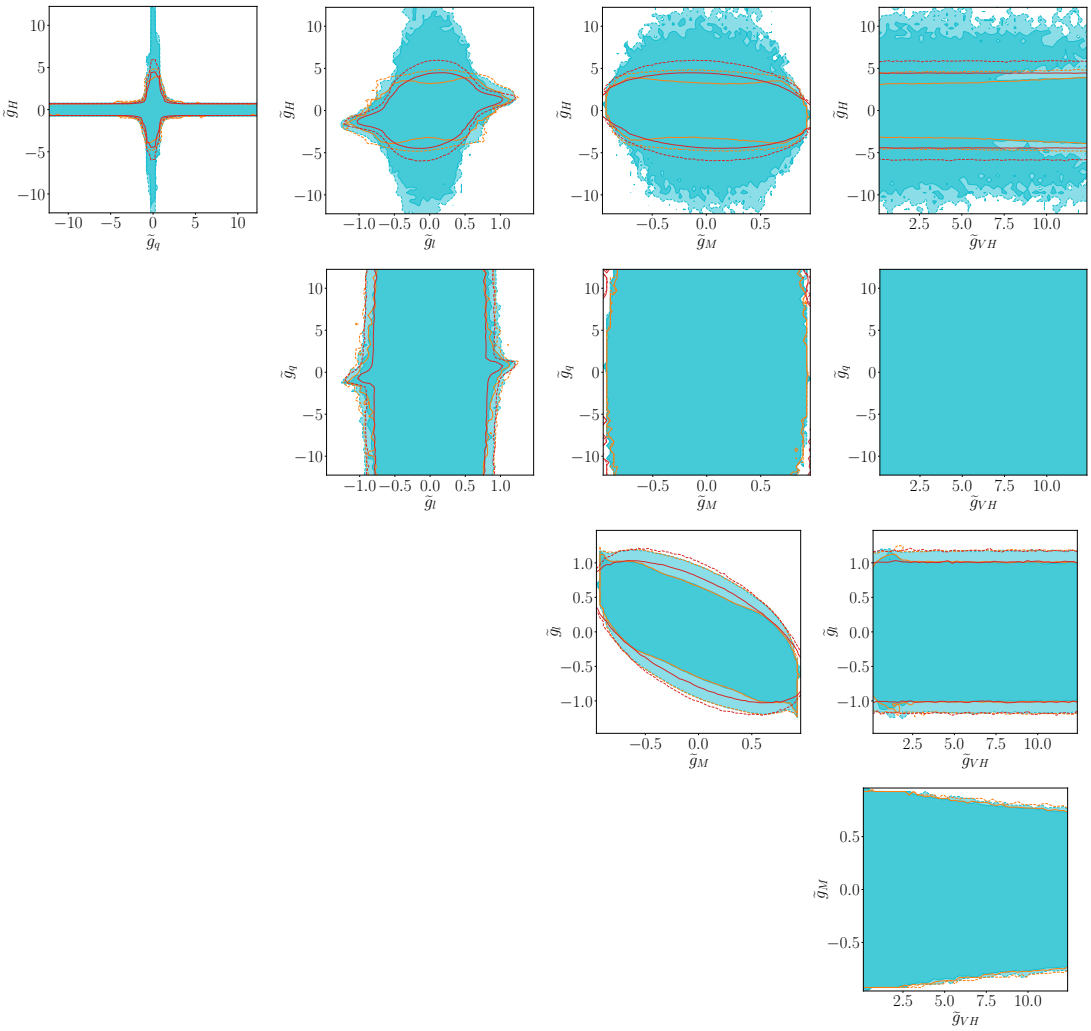


Figure 6.9: 5-parameter global fit of the full data set to the model parameters from Eq. (6.10) for fixed $m_V = 4\,\text{TeV}$. Profiled $\Delta\chi^2 = 2.3$ ($\Delta\chi^2 = 5.991$) contours are shown as solid (dashed) lines. Red (orange) curves indicate the results obtained with tree (1-loop) matching onto the SMEFT and a fixed matching scale $Q = m_V$. The light blue region shows the results from 1-loop matching, profiled over $Q = 500\,\text{GeV}$ ... m_V.

Fixed matching scale

For a fixed matching scale $Q = m_V$ (red and orange lines in Fig. 6.9), we find that the SMEFT fit constrains significantly $\tilde{g}_l$ and $\tilde{g}_H$, while $\tilde{g}_M$, $\tilde{g}_q$, and $\tilde{g}_{VH}$ are essentially unconstrained. The striking difference between the constraints on the vector triplet couplings to leptons and to quarks is largely due to the fact that the SMEFT fit is dominated by EWPO constraints extracted at LEP, on which the leptonic interactions have a much stronger impact. We have verified that, indeed, removing EWPO constraints from the fit relaxes significantly the constraint on $\tilde{g}_l$.

The 2D projections show that $\tilde{g}_l$ is also anti-correlated to $\tilde{g}_M$. The reason is that, at tree-level, $\tilde{g}_l$ enters the matching expressions only in the combination $\tilde{g}_l + g_2\tilde{g}_M$, where g_2 is the $SU(2)$ coupling constant. Specifically, we find that the constraints in the $\tilde{g}_M - \tilde{g}_l$ plane are dominated by the constraint on f_{LLLL}, whose tree-level matching expression is

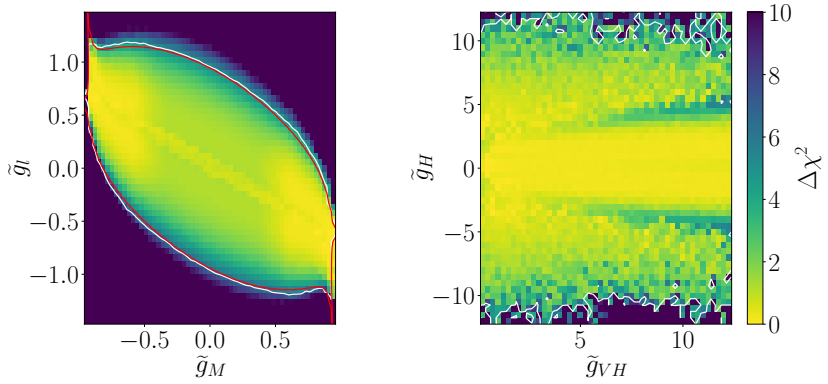


Figure 6.10: Heat map of the profiled $\Delta\chi^2$ distribution from the same fit as in Fig. 6.9, with 1-loop matching and profiling over the matching scale. The red contours indicate $f_{LLLL}/\Lambda^2 = -0.014, +0.017$ TeV^{-2} and the white contours indicate $\Delta\chi^2 = 5.991$.

quadratic in the relevant combination

$$\frac{f_{LLLL}}{\Lambda^2} = -\frac{(\tilde{g}_l + g_2\tilde{g}_M)^2}{4\tilde{m}_V^2} \,. \tag{6.37}$$

Therefore, for most values of $\tilde{g}_M$ and $\tilde{g}_l$, the constraints are driven by the limit for negative values of this Wilson coefficient. At 1-loop, the matching expression is more complex and allows for positive values of f_{LLLL} in a region close to $|\tilde{g}_M| \simeq 1$ and $|\tilde{g}_l| \simeq 1$. The right panel in Fig. 6.10 shows that the 2σ boundary from the 5D likelihood (in white) matches very well the contours for $f_{LLLL}/\Lambda^2 = -0.014, +0.017\,\text{TeV}^{-2}$ (in red), corresponding to the 2σ interval derived from a 2D fit of f_{LLLL} and f_{BW}. Here a 2D fit is necessary owing to the strong correlation between f_{LLLL} and f_{BW}. A 1D fit would lead to an over-estimation of the constraints.

There are no major differences between tree and loop level matching when keeping the matching scale fixed $Q = m_V$. Only slight differences can be observed in the limits on $\tilde{g}_M$ and $\tilde{g}_H$. The effect on $\tilde{g}_H$ is completely washed out once the matching scale is allowed to vary, as we discuss below. Although less visible due to the different scales, an analogous anti-correlation is present in the $\tilde{g}_M - \tilde{g}_H$ plane, as $\tilde{g}_H$ also enters tree-level matching expressions exclusively in the combination $\tilde{g}_H + g_2\tilde{g}_M$. Because $\tilde{g}_H$ enters many Wilson coefficients, both at tree and loop level, in this case it is not possible to identify one particular SMEFT parameter, or combination thereof, that drives the global bounds.

The constraint on $\tilde{g}_q$, on the other hand, is driven by that on $f_{\phi Q}^{(3)}$, whose matching expression is given in Eq. (6.36). This is consistent with the fact that $\tilde{g}_q$ only shows a non-trivial interplay with $\tilde{g}_H$. The cross-like shape emerging in the $(\tilde{g}_q, \tilde{g}_H)$ panel results from the superposition of the hyperbola-like shape expected from the $f_{\phi Q}^3$ matching expression, and of additional constraints on $\tilde{g}_H$ that introduce extra suppressions away from the two axes. Finally, $\tilde{g}_{VH}$ does not contribute to any dimension-6 operator at tree-level, so, in this limit, the likelihood is exactly flat in the corresponding direction. At 1-loop $\tilde{g}_{VH}$ gives contributions to $f_W, f_{WW}, f_{\phi 2}, f_{t,b,\tau}$ and $f_{\phi Q}^{(3)}$. Among these, the dominant constraint stems from $f_{\phi 2}$, leading to the orange contours in the $\tilde{g}_{VH} - \tilde{g}_M$ and $\tilde{g}_{VH} - \tilde{g}_H$ planes.

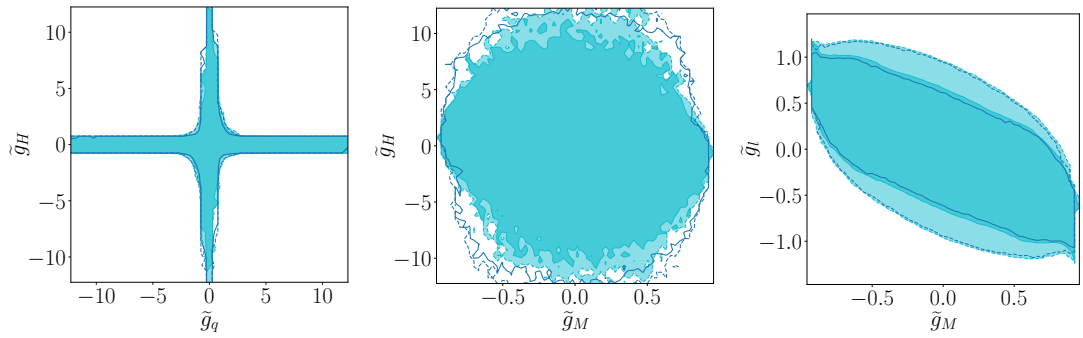


Figure 6.11: Impact of the high-energy kinematic distributions [251–253] on the global 5-parameter SMEFT fit for fixed $m_V = 4\,\text{TeV}$. The solid regions include the full data set (same as Fig. 6.9), while the dark blue lines exclude the high-energy kinematic distributions. Solid (dashed) lines mark the $\Delta\chi^2 = 2.3$ ($\Delta\chi^2 = 5.991$) contours.

Variable matching scale

Varying the matching scale as $Q = 500\,\text{GeV}$ ... $m_V = 4\,\text{TeV}$, as shown as light blue region in Fig. 6.9, affects the constraints on $\tilde{g}_H$, while for the other parameters the dependence is negligible. This is what we expect from the toy results in Sec. 6.3.2 and Fig. 6.6, and we have verified that extending the range to $Q \gtrsim m_V$ does not add any significant feature to the results. As for the 5-parameter fit, the main consequence of variable Q is that, for $Q \lesssim 2.4\,\text{TeV}$, the matching expressions of $f_{\phi 2}$ and $f_{t,b,\tau}$ acquire a new zero. Because these operators are the dominant source of constraints on $\tilde{g}_H$, this results in a broader allowed region for this parameter, which is largest close to the $Q \simeq 2.4\,\text{TeV}$ threshold. This effect washes out the correlation between $\tilde{g}_H$ and $\tilde{g}_M$ mentioned above.

At $Q \simeq 2.4\,\text{TeV}$, the most constraining Wilson coefficient is $f_{\phi 2}$, which is responsible for the outermost region of the 2σ contours for $\tilde{g}_H$ in Fig. 6.9. The inner structure of the likelihood, including the 1σ contour, cannot be explained in terms of a single Wilson coefficient. It is the result of a non-trivial interplay between several effects, including $\tilde{g}_H$ entering a large number of Wilson coefficients and profiling over the matching scale.

It is also interesting to look at the finer structure of the profiled likelihood. In Fig. 6.10 we show $\Delta\chi^2$ for the same 2D projections as before. We can see that the best-fit points are focused in regions where $|\tilde{g}_M| > 0.5$. This effect emerges in the 5-parameter fit with 1-loop matching, irrespective of whether the matching scale is fixed or varied. It is the same effect as observed for the 3-parameter fit varying the heavy vector mass in Fig. 6.3, and it is due to the EWPO preferring a best-fit point away from the SM. In particular, we have checked that the observed substructures are entirely dominated by less than 1σ deviations in $A_l(SLD)$ and m_W. In addition, the measurements of σ_h^0, R_l^0, $A_{FB}^{0,l}$, A_c reinforce the deviation through correlations. If future measurements with reduced uncertainties confirmed the present deviations from the SM, this would lead to exclusion limits with intricate patterns.

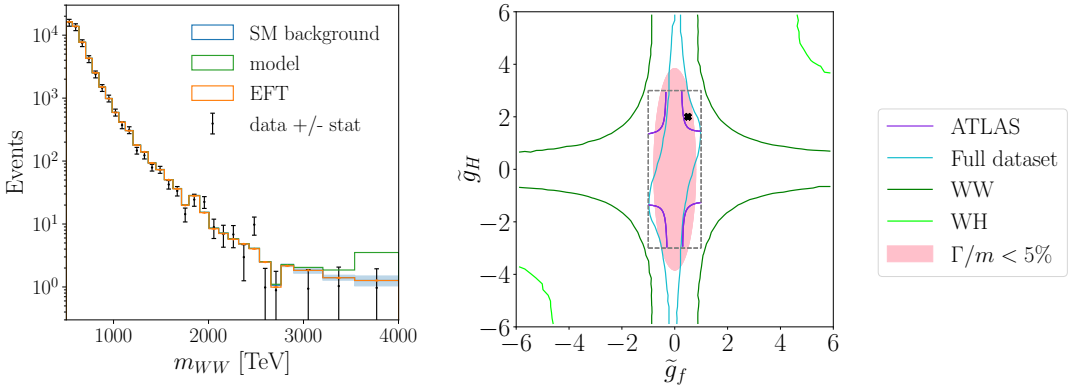


Figure 6.12: Left: Z' prediction for $m_V = 4\,\text{TeV}$, $\tilde{g}_H = 2$, $\tilde{g}_f = 0.5$ (shown by a star in the right panel) for the WW search [253], compared to the SMEFT prediction. Right: SMEFT limits ($\Delta\chi^2 = 5.991$) for $m_V = 4\,\text{TeV}$ and profiled over the matching scale, for the WW and WH distributions alone and the full dataset. We also show the 95%CL exclusion from the WH resonance search [252]. The gray box marks the ATLAS search region, the narrow-width is shaded in pink.

Impact of high energy measurements

It is well known [178] that kinematic distributions probing high invariant masses have significant impact on global fits to the SMEFT parameters. In our analysis, we confirm this behavior for the two analyses described in Sec. 6.4.1, which are found to constrain significantly f_W, $f_{\phi d}$ and f_{WWW}. Unfortunately, once the SMEFT is mapped onto the heavy vector triplet model, the constraining power of these measurements is diminished. This is shown in Fig. 6.11, where the results of Fig. 6.9 are compared to those from a 5-parameter fit where the three analyses of Refs. [251–253] are removed (dark blue line). The lack of visible impact of the high-energy kinematic distributions is very much due to the specific model and the corresponding numerical behavior of the matching formulae. As discussed above, the main constraints on the vector triplet parameter space are dominantly associated to those on $f_{LLLL}, f_{\phi 2}$ and $f^3_{\phi Q}$, which are only marginally improved by these searches.

SMEFT vs direct searches

A key question we would like to address in this work is whether a global SMEFT analysis can be competitive with direct searches in constraining a given UV model. Figure 6.12 compares the constraints in the $(\tilde{g}_f, \tilde{g}_H)$ plane obtained in the direct search of WH resonances by ATLAS, Ref. [252], and from 2D SMEFT fits to different sets of observables. In particular, the light green line indicates the SMEFT constraints obtained from the same distribution as in the direct search. For all lines in this plot, the heavy triplet mass is fixed to $m_V = 4\,\text{TeV}$, the maximum value accessible by the resonance search. Strictly speaking, the direct and indirect constraints extracted from the same measurement apply to complementary regions of the parameter space: the former are valid for masses $m_V \lesssim 4\,\text{TeV}$ and for narrow vector triplets within the pink-shaded region of Fig. 6.12, while the latter hold for $m_V \gg 4\,\text{TeV}$ irrespective of the resonance width. Obviously, a comparison should be taken with a grain of salt.

Nevertheless, it can be instructive to examine the interplay between the signals produced by a heavy resonance and by its corresponding SMEFT approximation. The left panel of Fig. 6.12 shows the m_{WW} resonant distribution obtained for a benchmark point at $m_V = 4\,\text{TeV}$, $\tilde{g}_H = 2$, $\tilde{g}_f \equiv \tilde{g}_l = \tilde{g}_q = 0.5$, compared to the ATLAS measurement [252] (black data points) and the SMEFT signal matched to this benchmark model at dimension six. This point is indicated by a cross in Fig. 6.12 (right), and it is excluded at 95%CL by both the ATLAS WH and WW searches, but falls within the 2σ-allowed region of our SMEFT global analysis. This discrepancy is obvious from the high-energy m_{WW} tail, where aside from the mass peak the dimension-6 SMEFT also misses the initial rise of the distribution. Among the Wilson coefficients that contribute to WW production, only $f_W/\Lambda^2 = 0.28\,\text{TeV}^{-2}$ takes a value above the permille level, while $f_{\phi Q}^{(3)} = 0$ because $\tilde{g}_q = \tilde{g}_l$. This results in SMEFT signals of only a few percent across the entire m_{WW} distribution, which are always well within the uncertainties. It is worth pointing out that in such a situation the best place to look for the SMEFT signal might not just be the bins where the energy enhancement is largest, but rather those where the uncertainties are smallest.

While not surprising, these conclusions do not extend to arbitrary BSM scenarios. One characteristic of the case examined here is that the resonance is narrow. As a consequence, the effect in m_{WW} is only visible close to m_V, where the SMEFT expansion immediately breaks down. The situation improves when we include higher-dimensional operators [193, 268]. At dimension six, the matching to our specific model suppresses all energy-enhanced SMEFT contributions to WW production, so the signal is underestimated across the em_{WW} distribution. This does not have to be the case in other BSM models. For instance, it is possible that the dimension-6 approximation over-estimates the model predictions, in which case the dimension-8 contributions need to be large and negative, and the truncated SMEFT constraints appear more stringent than those from direct searches.

Going beyond the comparison of resonance searches and SMEFT analyses for one measurement, the true power of the SMEFT approach is that it allows to combine a large number of different measurements. This will always improve the sensitivity of the SMEFT analyses and, on the other hand, it allows to derive more general conclusions, by constraining all model parameters simultaneously, as shown in Fig. 6.9. The light blue lines in Fig. 6.12 show the constraints from a 2-parameter SMEFT fit to the entire dataset employed in this work. Consistent with the discussions above, these limits are dominated by EWPO, for which the SMEFT expansion is valid. In particular, the constraint on $\tilde{g}_f$ is dominated by the leptonic component $\tilde{g}_l$, which in turn is mostly associated to the f_{LLLL} Wilson coefficient. Comparing these limits to those from the ATLAS WH-search, we find that the latter are slightly stronger for $|\tilde{g}_H| \gtrsim 1$ (with the caveat that they are only valid in the narrow width regime), while the former dominate for $|\tilde{g}_H| \lesssim 1$. Here, the WH search has an unconstrained direction along the $\tilde{g}_H = 0$ axis, that is broken by the EWPO in the SMEFT fit [196].

Heavy vector results

One of the main motivations for the SMEFT formalism is that it allows us to derive constraints on new particles with masses beyond the reach of direct searches. In this spirit, we can extend our SMEFT constraints on the $\tilde{g}$ parameters for a heavy triplet mass to $m_V = 8\,\text{TeV}$. Now, the dimension-6 SMEFT approximation is valid all over the

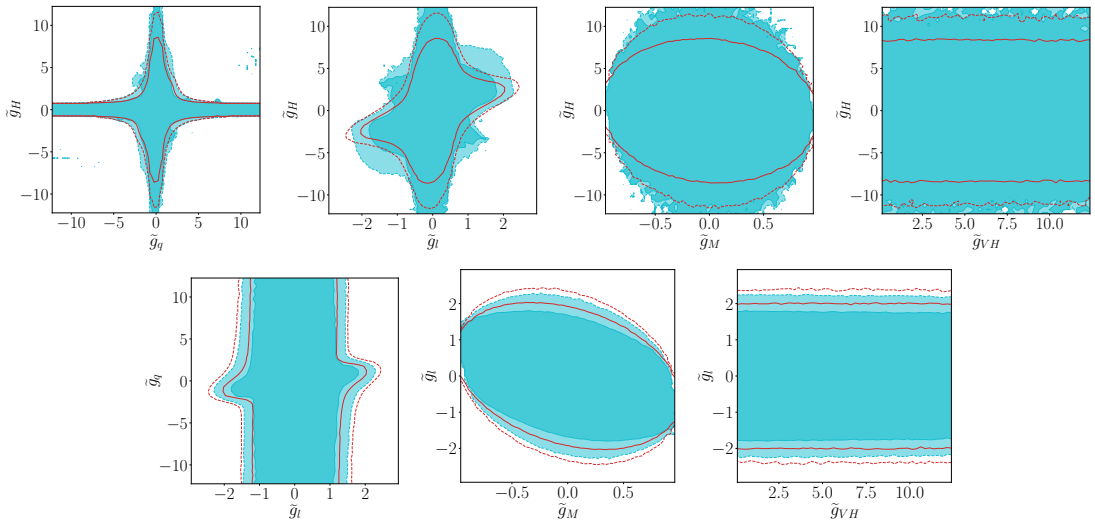


Figure 6.13: 5-parameter fit to the full data set for the model parameters in Eq. (6.10) for fixed $m_V = 8\,\text{TeV}$. Each panel shows profiled $\Delta\chi^2 = 2.3$ (solid) and $\Delta\chi^2 = 5.991$ (dashed) contours. Red curves correspond to tree-level matching, the light blue region to 1-loop matching, profiled over three $\tilde{g}$ parameters plus the matching scale $Q = 500\,\text{GeV}$ ... m_V. The panels for $\tilde{g}_M - \tilde{g}_q$ $\tilde{g}_{VH} - \tilde{g}_q$ and $\tilde{g}_{VH} - \tilde{g}_M$ are not shown as they are unconstrained in the explored ranges.

kinematic measurements discussed above. The corresponding results in Fig. 6.13 can be directly compared to those in Fig. 6.9 for $m_V = 4\,\text{TeV}$. As expected, all the bounds on the model parameters are weaker for heavier values of m_V (see also Fig. 6.3). However, a notable feature is that the limits do not simply scale with a factor proportional to m_V, as one would naively expect from the SMEFT analysis at dimension six. The reason is that the matching expressions that relate the Wilson coefficients to the model parameters are generally non-trivial and do not scale universally with $(\tilde{g}_i/m_V)$, as can be seen for instance in Eq. (6.32). Moreover, as we are considering a BSM state that is not a singlet under $SU(2)$, the EW gauge coupling g_2 contributes to the matching independently of the $\tilde{g}$ parameters. The result is that the degeneracy between $\tilde{g}_i$ and m_V is largely broken in the matching, leading to a complex likelihood structure that changes significantly with m_V.

6.5 Outlook

We have presented a global analysis of a Standard Model extension with a gauge-triplet vector resonance in terms of the dimension-6 SMEFT Lagrangian. We have performed a global SFitter analysis including electroweak precision observables, Higgs and di-boson measurements as well as resonance searches at the LHC, and have compared our results with limits obtained from direct searches. To relate the full model and the SMEFT we have employed one-loop matching with a focus on the theory uncertainties from the choice of the matching scale.

First, we have shown that the theory uncertainty due to the choice of the matching scale can have a large effect on the global analysis. In particular, the bounds on the coupling of the new vector to the SM-Higgs are significantly weakened once we profile over a variable

matching scale, illustrating how all theory uncertainties need to be taken into account at least once we translate SMEFT results back into models.

Comparing the SMEFT results with direct searches reveals an intriguing complementarity. Direct and SMEFT searches are reliable in different parameter regions; while direct searches are sensitive to narrow resonances with kinematically accessible masses, SMEFT searches apply to energies sufficiently below the resonance mass. The SMEFT analysis can be sensitive to the onset of the resonance, but a reliable description of this region requires a tower of higher-dimensional operators. Specifically for the vector-triplet model, the SMEFT model for the high-energy tail of kinematic distributions turned out less sensitive than the resonance search, and therefore provided conservative constraints. On the other hand, the SMEFT analysis can probe vector masses beyond the reach of resonance searches. Here, we found that the one-loop matching dampens the sensitivity decrease of the SMEFT analysis compared to the naively expected scaling.

While SMEFT analyses cannot replace model-specific searches for new physics, they add valuable constraints from a large variety of measurements and are sensitive to new physics scales beyond the reach of resonance searches. Only this complementarity of direct and indirect searches allows us to make best use of current and future LHC data.

Chapter 7

To profile or to marginalize? – a SMEFT case study

The research presented in this chapter has been published in Ref. [5]. Most of the figures, tables and text are completely identical to the content of this publication.

7.1 Introduction

Higgs physics at the LHC [174] perfectly illustrates a deep tension in contemporary particle physics: on the one hand, the existence of a fundamental Higgs boson is a direct consequence of describing the electroweak gauge sector in terms of a quantum field theory, specifically a renormalizable gauge theory. It looks like Nature chose the simplest possible realization of the Higgs mechanism, with one light scalar particle and an electroweak vacuum expectation value (VEV) of unknown origin. On the other hand, puzzles like dark matter or baryogenesis seem to point to non-minimal Higgs sectors for convincing solutions based on renormalizable quantum field theory, but without any LHC hint in these directions. The main goal of the LHC Higgs program is to understand if Nature really took the opportunity of a minimal electroweak and Higgs sector solving as many problems as possible, or why she skipped this opportunity in favor of theoretically less attractive alternatives. Or, more practically speaking, we need to study as many Higgs properties as precisely as possible.

Given the vast LHC dataset already after Run 2 and our fundamental ignorance of the correct UV-completion of the Standard Model (SM), we need to measure Higgs-related observables and express them in a consistent, fundamental, and comprehensive theory framework. To provide the necessary precision, this framework has to be defined beyond leading order in perturbation theory, it needs to incorporate kinematic information, and it should allow us to combine as many LHC observables as possible. The EFT extension of the Standard Model (SMEFT) [175] fulfills precisely these three requirements and defines a theoretical path to understanding the entire LHC dataset in terms of a fundamental Lagrangian. Its main shortcoming is the necessary truncation in the operator dimensionality. The truncated SMEFT approximation will hardly describe new physics appropriately, so SMEFT should really be viewed as a systematic, conservative limit-setting tool. Of course, this practical aspect does not cut into the fundamental attractiveness of an effective quantum field theory description of all LHC data.

While ATLAS and CMS have not published properly global SMEFT analyses of the Higgs-electroweak or top sector, there exists a range of phenomenological Higgs-gauge

analyses [178,180–183,269], top analyses [246,270,271], combinations of the two [184,272], and combinations with parton densities [273]. These analyses are typically based on experimentally preprocessed information, including the full range of uncertainties. Given that by assumption any SMEFT analysis will be centered around the renormalizable SM-Lagrangian, the main focus of all global analyses is the uncertainty treatment and the correlations between the different operators. Technically, these two tasks tend to collide. We can choose a conservative uncertainty treatment based on profile likelihoods and nuisance parameters, but it is much more computing-efficient to treat correlations through covariance matrices of marginalized Gaussian likelihoods [274]. The SFitter framework [178,242,243,245,246,257] is unique in the sense that it has mostly been used for profile likelihood analyses, but can provide marginalized limits equally well [257,275].

We make use of this flexibility and study, for the first time, the difference between profiled and marginalized likelihoods of the same global Run 2 dataset. In Sec. 7.3 we find that for the Higgs-electroweak dimension-6 operators and the given dataset the two approaches agree well, so we can use the marginalized setup to treat correlated measurements and uncertainties efficiently. Based on these results, we include a range of recent Run 2 measurements from Higgs studies as well as from exotics resonance searches, again with a focus on a comprehensive and conservative uncertainty treatment, in Sec. 7.4. Finally, we study the impact of these new measurements and the inputs from a global top analysis in Sec. 7.5 and find interesting differences between the profiling and marginalization methods.

7.2 SMEFT Lagrangian

The SMEFT Lagrangian is based on the same field content, global, and gauge symmetries as the SM. It includes interactions with canonical dimension larger than four, organized in a systematic expansion in the inverse powers of a new physics scale [42, 276, 277]. Neglecting lepton number violation at dimension five, the leading beyond-SM effects stem from dimension-six terms,

$$\mathcal{L} = \sum_x \frac{f_x}{\Lambda^2} \mathcal{O}_x \quad . \tag{7.1}$$

There are 59 baryon-number conserving operators, barring flavor structure [278–282]. We use the operator basis of Refs. [178, 283], starting with a set of P-even and C-even operators and then using the equations of motion to define a basis without blind directions in the electroweak precision data. We neglect operators that cannot be studied at the LHC yet, like those changing the triple-Higgs vertex [284–288]. We also neglect operators which are too strongly constrained from other LHC measurements to affect the Higgs-electroweak analysis, like the ubiquitous triple-gluon operator

$$\mathcal{O}_G = f_{ABC} G^\rho_{A\nu} G^\nu_{B\lambda} G^\lambda_{C\rho} \,, \tag{7.2}$$

which is strongly constrained from multi-jet production [289]. In the bosonic sector the relevant operators then are

$$\begin{aligned}
&\mathcal{O}_{GG} = \phi^\dagger\phi \; G^a_{\mu\nu} G^{a\mu\nu} && \mathcal{O}_{WW} = \phi^\dagger \; \hat{W}_{\mu\nu} \hat{W}^{\mu\nu} \; \phi && \mathcal{O}_{BB} = \phi^\dagger \; \hat{B}_{\mu\nu} \hat{B}^{\mu\nu} \; \phi \\
&\mathcal{O}_W = (D_\mu\phi)^\dagger \hat{W}^{\mu\nu} (D_\nu\phi) && \mathcal{O}_B = (D_\mu\phi)^\dagger \hat{B}^{\mu\nu} (D_\nu\phi) && \mathcal{O}_{BW} = \phi^\dagger \; \hat{B}_{\mu\nu} \hat{W}^{\mu\nu} \; \phi
\end{aligned}$$

$$\mathcal{O}_{\phi 1} = (D_\mu \phi)^\dagger \, \phi\phi^\dagger \, (D^\mu \phi) \qquad \mathcal{O}_{\phi 2} = \frac{1}{2}\partial^\mu(\phi^\dagger\phi)\partial_\mu(\phi^\dagger\phi)$$
$$\mathcal{O}_{3W} = \mathrm{Tr}\left(\hat{W}_{\mu\nu}\hat{W}^{\nu\rho}\hat{W}^\mu_\rho\right) \, , \tag{7.3}$$

where $\hat{B}_{\mu\nu} = ig'B_{\mu\nu}/2$ and $\hat{W}_{\mu\nu} = ig\sigma^a W^a_{\mu\nu}/2$. The covariant derivative acting on the Higgs doublet is $D_\mu = \partial_\mu + ig'B_\mu/2 + ig\sigma_a W^a_\mu/2$.

In addition to the purely bosonic operators, we also need to include single-current operators modifying the Yukawa couplings,

$$\mathcal{O}_{e\phi,22} = \phi^\dagger\phi \; \bar{L}_2\phi e_{R,2} \qquad \mathcal{O}_{e\phi,33} = \phi^\dagger\phi \; \bar{L}_3\phi e_{R,3}$$
$$\mathcal{O}_{u\phi,33} = \phi^\dagger\phi \; \bar{Q}_3\tilde{\phi} u_{R,3} \qquad \mathcal{O}_{d\phi,33} = \phi^\dagger\phi \; \bar{Q}_3\phi d_{R,3} \, , \tag{7.4}$$

The main difference to earlier SFitter analyses is that we treat the correction to the muon Yukawa $f_{e\phi,22}$ as an independent parameter, while previously it was tied to $f_{e\phi,33}$ via an approximate flavor symmetry. As LHC Run 2 found experimental evidence for the Higgs coupling to muons, this approximation can now be dropped. However, when including the observed branching ratio to muons, we will not be sensitive to the sign of the muon Yukawa, except for the fact that such a sign flip is not consistent with the SMEFT assumptions.

Other single-current operators modify gauge and gauge-Higgs ($HVff$) couplings [179, 180, 261–266],

$$\mathcal{O}^{(1)}_{\phi u} = \phi^\dagger(i\overleftrightarrow{D}_\mu\phi)(\bar{u}_R\gamma^\mu u_R) \qquad \mathcal{O}^{(1)}_{\phi Q} = \phi^\dagger(i\overleftrightarrow{D}_\mu\phi)(\bar{Q}\gamma^\mu Q)$$
$$\mathcal{O}^{(1)}_{\phi d} = \phi^\dagger(i\overleftrightarrow{D}_\mu\phi)(\bar{d}_R\gamma^\mu d_R) \qquad \mathcal{O}^{(3)}_{\phi Q} = \phi^\dagger(i\overleftrightarrow{D}^a_\mu\phi)\left(\bar{Q}\gamma^\mu\frac{\sigma_a}{2}Q\right)$$
$$\mathcal{O}^{(1)}_{\phi e} = \phi^\dagger(i\overleftrightarrow{D}_\mu\phi)(\bar{e}_R\gamma^\mu e_R) \, . \tag{7.5}$$

The four-lepton operator

$$\mathcal{O}_{4L} = (\bar{L}_1\gamma_\mu L_2)\,(\bar{L}_2\gamma^\mu L_1) \tag{7.6}$$

induces a shift in the Fermi constant. For the operators in Eq. (7.5), we maintain for simplicity a flavor symmetry, and all currents are implicitly defined with diagonal flavor indices. In this limit, the operators $\mathcal{O}^{(1)}_{\phi L}, \mathcal{O}^{(3)}_{\phi L}$, analogous to $\mathcal{O}^{(1)}_{\phi Q}, \mathcal{O}^{(3)}_{\phi Q}$, are redundant with the bosonic set of Eq. (7.3) via equations of motion [178, 283].

Dipole operators and $\mathcal{O}^{(1)}_{\phi ud,ij} = \tilde{\phi}^\dagger(i\,D_\mu\phi)(\bar{u}_{R,i}\gamma^\mu d_{R,j})$ are neglected for two reasons: the approximate flavor symmetry requires them to scale with the SM Yukawa couplings and their interference with the SM is always proportional to the fermion masses. Both factors suppress their effects except for the top quark. The three dipole moments of the top quark — electric, magnetic and chromomagnetic — are not suppressed, so in this work we choose to retain the chromomagnetic operator [180, 290–292]

$$\mathcal{O}_{tG} = ig_s(\bar{Q}_3\sigma^{\mu\nu}T^A u_{R,3})\;\tilde{\phi}G^A_{\mu\nu} \, . \tag{7.7}$$

It affects the Higgs observables at the LHC significantly through the loop-induced production process [184, 272, 293–295].

Our SMEFT Lagrangian is then defined as

$$\begin{aligned}\mathcal{L}_{\text{eff}} = \mathcal{L}_{\text{SM}} &- \frac{\alpha_s}{8\pi}\frac{f_{GG}}{\Lambda^2}\mathcal{O}_{GG} + \frac{f_{WW}}{\Lambda^2}\mathcal{O}_{WW} + \frac{f_{BB}}{\Lambda^2}\mathcal{O}_{BB} + \frac{f_{BW}}{\Lambda^2}\mathcal{O}_{BW} \\ &+ \frac{f_W}{\Lambda^2}\mathcal{O}_W + \frac{f_B}{\Lambda^2}\mathcal{O}_B + \frac{f_{\phi 1}}{\Lambda^2}\mathcal{O}_{\phi 1} + \frac{f_{\phi 2}}{\Lambda^2}\mathcal{O}_{\phi 2} + \frac{f_{3W}}{\Lambda^2}\mathcal{O}_{3W} \\ &+ \frac{f_\mu m_\mu}{v\Lambda^2}\mathcal{O}_{e\phi,22} + \frac{f_\tau m_\tau}{v\Lambda^2}\mathcal{O}_{e\phi,33} + \frac{f_b m_b}{v\Lambda^2}\mathcal{O}_{d\phi,33} + \frac{f_t m_t}{v\Lambda^2}\mathcal{O}_{u\phi,33} \\ &+ \frac{f^{(1)}_{\phi Q}}{\Lambda^2}\mathcal{O}^{(1)}_{\phi Q} + \frac{f^{(1)}_{\phi d}}{\Lambda^2}\mathcal{O}^{(1)}_{\phi d} + \frac{f^{(1)}_{\phi u}}{\Lambda^2}\mathcal{O}^{(1)}_{\phi u} + \frac{f^{(1)}_{\phi e}}{\Lambda^2}\mathcal{O}^{(1)}_{\phi e} + \frac{f^{(3)}_{\phi Q}}{\Lambda^2}\mathcal{O}^{(3)}_{\phi Q} + \frac{f_{4L}}{\Lambda^2}\mathcal{O}_{4L} \\ &+ \frac{f_{tG}}{\Lambda^2}\mathcal{O}_{tG} + \text{invisible decays}\,. \end{aligned} \tag{7.8}$$

It contains 20 independent Wilson coefficients. The branching ratio of the Higgs to invisible final states, BR_{inv}, is treated as a free parameter, to account for potential Higgs decays to a dark matter agent. For the global analysis it is convenient to work with the two orthogonal combinations

$$\mathcal{O}_\pm = \frac{\mathcal{O}_{WW} \pm \mathcal{O}_{BB}}{2} \qquad \Rightarrow \qquad f_\pm = f_{WW} \pm f_{BB}\,. \tag{7.9}$$

The rotation is defined such that only $\mathcal{O}_+$ contributes to the $H\gamma\gamma$ vertex.

If we base our calculation on the Lagrangian like that given in Eq. (7.8), we strictly speaking need to supplement it with a renormalization scheme or a renormalization condition. For each process, a reasonable assumption is that all Lagrangian parameters, including the Wilson coefficients, are evaluated at the same renormalization scale μ_R. For the processes entering our global analysis, an appropriate central scale choice is $\mu_R \simeq m_H/2 \,...\, m_H$. To improve the precision beyond leading order, one should eventually account for the renormalization group evolution [296], and evaluate the SMEFT predictions at the energy scale appropriate for each process. This scale can vary for instance across bins of a kinematic distribution. In this work, all SMEFT predictions are calculated at leading order, so we postpone an in-depth analysis of renormalization group effects to a future work, together with a systematic study of the impact of higher-order corrections to inclusive Higgs production and decay rates.

The truncated Lagrangian of Eq. (7.8) as our fundamental theory hypothesis needs to be put into context. The hypothesis based on a truncated Lagrangian is, strictly speaking, not well defined once we include higher multiplicities of the dimension-6 operators in the amplitude. Therefore, the SMEFT analysis should be interpreted as representing classes of models [297, 298], and the validity of the SMEFT approach rests on the process-dependent assumption that in the corresponding models no new particle is produced on its mass shell [186]. While SMEFT is an excellent framework to interpret global LHC analysis, possible anomalies need to be interpreted by matching it to UV-complete models [187, 188, 191, 299], where for instance WBS signatures of corresponding models might eventually require us to go beyond dimension-6 operators [300].

If global SMEFT analyses should be interpreted as representing classes of UV-complete models for a limited set of observables, we need to consider the interplay between the SMEFT hypothesis and more fundamental models. Given the precision of the SMEFT analysis and its field-theoretical advantages over the naive coupling analysis we can and should perform this matching beyond leading order [185, 209, 210, 212, 215, 301], accounting for matching scale uncertainties [3, 183], rather than ignoring them at leading order.

While this scale uncertainty clearly does not cover all uncertainties induced by matching SMEFT limits to UV-complete models, it also illustrates that such uncertainties exist and have to be taken into account.

7.3 Bayesian SFitter setup

Global SMEFT analyses are a key ingredient to a more general analysis strategy at the LHC, which is to test theory predictions based on perturbative quantum field theory using the full kinematic range of the complete set of LHC measurements. It is worth stressing that SMEFT analyses are currently the only way to systematically probe kinematic LHC measurements beyond resonance searches. They come with two assumptions which greatly simplify the actual analyses

1. experimentally, we know that our SMEFT analysis is not confronted with established anomalies; those should be discussed using properly defined BSM models;

2. theoretically, SMEFT can only describe small deviations from the Standard model, otherwise the dimensional expansion in Eq. (7.1) is not valid.

While global SMEFT analyses with a truncated Lagrangian can translate kinematic measurements into fundamental parameters, these two aspects imply that their outcome will be limit-setting. For our analysis this means that we already know that the global maximum of the SMEFT likelihood lies around the SM-limit $f_x/\Lambda^2 \to 0$. The exact position of the most likely parameter point is of limited interest, the main task of the global analysis is to determine the uncertainty on the values of the Wilson coefficients or, more in general, the finite preferred region in the multi-dimensional SMEFT parameter space.

In this spirit, the goal of the SFitter framework is to enable an independent interpretation of experimental inputs, without relying on pre-processed information and including a comprehensive treatment of statistical, systematic, and theory uncertainties [242,243,257]. The SFitter methodology relies on the construction of a likelihood function in which these uncertainties can be described by nuisance parameters. In all previous SFitter analyses, nuisance parameters are profiled over. The resulting profile likelihood is then profiled over the parameters of interest, to extract one- and two-dimensional limits on the Wilson coefficients. An alternative, Bayesian treatment is based on marginalizing over nuisance parameters and parameters of interest. It has been adopted in several SMEFT analyses [274,302–305] and simplifies greatly the treatment of correlated uncertainties. The goal of this work is to perform an apples-to-apples comparison between a profiled and a marginalized likelihood, employing exactly the same data and uncertainties inputs in both cases.

Marginal likelihood

Since marginalization is new in SFitter, we provide a brief description of the main features. The corresponding profile likelihood treatment is discussed in detail in Refs. [178,242, 245,246,257]. The first step of a global analysis is the construction of the fully exclusive likelihood $\mathcal{L}_{\text{excl}}$, which is a function of the parameters of interest f_x and of a set of nuisance parameters θ_i. This $\mathcal{L}_{\text{excl}}$ is defined with the following uncertainty treatment: (i) statistical uncertainties are included via a Poisson distribution, in some cases approximated using a

Gaussian whenever this stabilizes the numerical evaluation; (ii) systematic uncertainties are assumed to be Gaussian, organized in 31 categories, such that uncertainties within the same category are fully correlated through a covariance matrix or through nuisance parameters. Systematics which do not fit into any of the 31 categories are assumed to be uncorrelated; (iii) theory uncertainties are modelled as flat distributions. Whenever theory uncertainties need to be correlated we use an explicit nuisance parameter.

For a Bayesian analysis we first marginalize over or integrate out the nuisance parameters. This yields the marginal likelihood $\mathcal{L}_{\text{marg}}$, for one counting measurement and one parameter illustrated by

$$\mathcal{L}_{\text{marg}}(f_x) = \int d\theta\, \mathcal{L}_{\text{excl}}(f_x, \theta) = \int d\theta\, \text{Pois}(d|m(f_x, \theta))\, p(\theta)\ . \tag{7.10}$$

Here d stands for the measured number of events, m is the model (theory) prediction, θ is a nuisance parameter and $p(\theta)$ the distribution over the nuisance parameter which, in the Bayesian context, defines the prior. In SFitter, nuisance priors are either Gaussian or flat. Computing $\mathcal{L}_{\text{marg}}$ in SFitter starts with the marginalization procedure over the nuisance parameters, so we omit the dependence on f_x for now.

SFitter provides several options to define the statistical model of a measurement, including a simplified Gaussian likelihood where uncertainties add in quadrature. A more sophisticated and reliable framework starts with a typical LHC measurement as an independent counting experiment, which is modelled by a Poisson distribution. Systematic uncertainties or theory uncertainties then define the completely exclusive likelihood for one measurement

$$\mathcal{L}_{\text{excl}}(\theta) = \text{Pois}(d|m(\theta_1, \theta_2, ..., b))\, p(b) \prod_i p(\theta_i)\ . \tag{7.11}$$

Here d is the measured number of events, b the background estimate, and m the model prediction, that is a function of the nuisance parameters θ_i. The distributions $p(b)$ and $p(\theta_i)$ incorporate our knowledge about these quantities. In general, they can be extracted from auxiliary measurements, simulations, or other possible sources. However, because tracking hundreds of different reference measurements is beyond the scope of SFitter, we simply assume $p(\theta_i)$ to be Gaussian for systematic uncertainties and flat or uniform for theory uncertainties,

$$p(\theta_i) = \begin{cases} \mathcal{N}_{0,\sigma_i}(\theta_{\text{syst},i}) & \text{systematics} \\ \mathcal{F}_{0,\sigma_i}(\theta_{\text{theo},i}) & \text{theory}\ . \end{cases} \tag{7.12}$$

In this step we assume that all prior distributions for θ_{syst} and θ_{theo} are centered around zero, with given half-widths σ.

For $p(b)$, SFitter provides several choices: for measurements where b is extracted from a single control region (CR) measurement we use

$$p(b) = \text{Pois}(b_{\text{CR}}|bk)\ , \tag{7.13}$$

where k is an interpolation factor between CR and signal region, b_{CR} is the measured number of events in the control region, and b is the expected number of background events in the signal region. For measurements with several control regions or with simulated backgrounds we assume the combined $p(b)$ to be a Gaussian. Systematic uncertainties on

the background measurement can also be included, and are assumed to be fully correlated with the uncertainties on the signal region within the same category.

Typically, the dependence of the theory prediction m on the nuisance parameters in Eq. (7.11) is not spelled out or extremely complex to determine. To simplify this task, we assume a leading linear dependence on assumed-to-be small uncertainties

$$m \approx s + b + \theta_{\text{theo},1} + \theta_{\text{theo},2} + \cdots + \theta_{\text{syst},1} + \theta_{\text{syst},2} + \cdots \equiv s + b + \theta_{\text{tot}} \; . \tag{7.14}$$

where s is the expected number of signal events. The exclusive likelihood of Eq. (7.11) can then be written as

$$\mathcal{L}_{\text{excl}}(\theta) \approx \text{Pois}(d|s + b + \Sigma\theta_{\text{theo},j} + \Sigma\theta_{\text{syst},i})\; p(b) \prod_j \mathcal{F}_{0,\sigma_j}(\theta_{\text{theo},j}) \prod_i \mathcal{N}_{0,\sigma_i}(\theta_{\text{syst},i}) \; , \tag{7.15}$$

The marginal likelihood for a single measurement is then constructed by integrating over all nuisance parameters,

$$\begin{aligned} \mathcal{L}_{\text{marg}} &= \int \prod_j d\theta_{\text{theo},j} \int \prod_i d\theta_{\text{syst},i} \int db \; \mathcal{L}_{\text{excl}}(\theta) \\ &= \int \prod_j d\theta_{\text{theo},j} \mathcal{F}_{0,\sigma_j}(\theta_{\text{theo},j}) \int \prod_i d\theta_{\text{syst},i} \mathcal{N}_{0,\sigma_i}(\theta_{\text{syst},i}) \\ &\qquad \times \int db \; \text{Pois}(d|s + b + \Sigma\theta_{\text{theo},j} + \Sigma\theta_{\text{syst},i})\; p(b) \; . \end{aligned} \tag{7.16}$$

The integration over b can be performed analytically if $p(b)$ is a Poisson distribution. In this case, the convolution $\mathcal{P}(d|s + \theta_{\text{tot}})$ of $p(b)$ and $\text{Pois}(d|m)$ gives a so-called Poisson-Gamma model, as Eq. (7.13) is a special case of the Gamma distribution,

$$\mathcal{L}_{\text{marg}} = \int \prod_j d\theta_{\text{theo},j} \mathcal{F}_{0,\sigma_j}(\theta_{\text{theo},j}) \int \prod_i d\theta_{\text{syst},i} \mathcal{N}_{0,\sigma_i}(\theta_{\text{syst},i}) \times \mathcal{P}(d|s + \theta_{\text{tot}}) \; . \tag{7.17}$$

We use θ_{tot} as defined in Eq. (7.14). To solve the remaining integrals over the nuisance parameters we replace one of the integrals, for instance $\theta_{\text{syst},1}$ with $(\theta_{\text{tot}} - \Sigma_{i\neq 1}\theta_{\text{syst},i})$,

$$\begin{aligned} \mathcal{L}_{\text{marg}} = \int d\theta_{\text{tot}} \; &\mathcal{P}(d|s + \theta_{\text{tot}}) \\ &\times \underbrace{\int \prod_j d\theta_{\text{theo},j} \mathcal{F}_{0,\sigma_j}(\theta_{\text{theo},j}) \int \prod_{i\neq 1} d\theta_{\text{syst},i} \mathcal{N}_{0,\sigma_i}(\theta_{\text{syst},i}) \mathcal{N}_{0,\sigma_1}(\theta_{\text{syst},1})}_{\text{solved analytically}} \; . \end{aligned} \tag{7.18}$$

Assuming only Gaussian plus at most three flat priors, all θ-convolutions except for one can be performed analytically. The corresponding closed formulas are implemented in SFitter, speeding up the marginalization (see Appendix B.2). The remaining 1-dimensional integral in Eq. (7.18) is solved numerically with Simpson's method.

Marginalizing over nuisance parameters and profiling over them will not give the same marginalized likelihood. Only for statistical uncertainties described by Poisson statistics and Gaussian systematics, the two lead to the same marginalized result in the limit of large enough statistics. Differences appear when we use flat theory uncertainties. For a

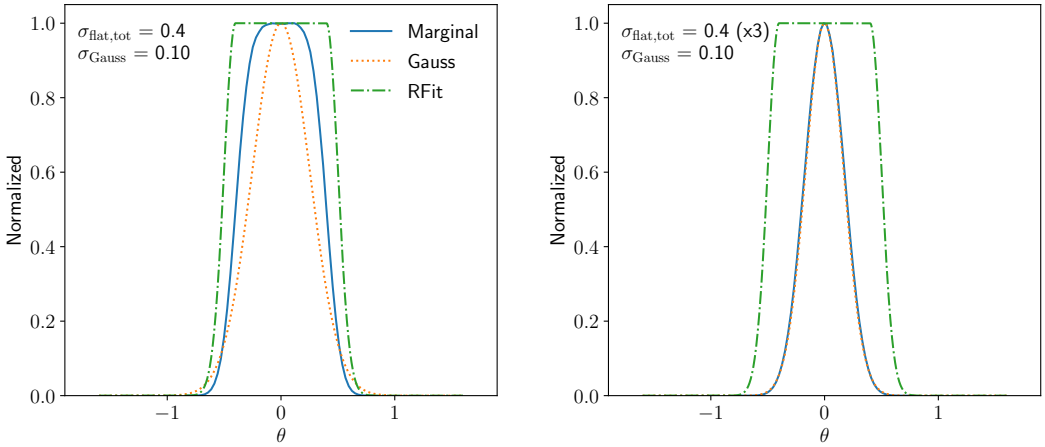


Figure 7.1: Marginalized and profiled likelihoods from the convolution of a Gaussian distribution with one (left) and three (right) flat ones. The orange curve shows, for comparison, the Gaussian obtained adding half-widths in quadrature.

Bayesian marginalization the central limit theorem ensures that the final posterior will be approximately Gaussian. Using a profile likelihood, two uncorrelated flat uncertainties add linearly, while a combination of flat and Gaussian uncertainties give the well-known RFit prescription [248]. Figure 7.1 shows, as an illustration, the distributions obtained combining one Gaussian with one (left) or three (right) flat nuisance parameters. We see that the profile likelihood or RFit result maintains a flat core and is independent of the number of theory nuisances, while the marginalized result varies and is very close to a Gaussian in the right panel.

Combining channels

Unlike probabilities, likelihoods of a set of measurements can simply be multiplied. This means we can generalize Eqs. (7.11) and (7.15) to a set of N measurements by replacing

$$\begin{aligned} \text{Pois}(d|m)p(b) \quad &\longrightarrow \quad \prod_k \text{Pois}(d_k|m_k)p(b_k) \\ \mathcal{N}_{0,\sigma_i}(\theta_{\text{syst},i}) \quad &\longrightarrow \quad \mathcal{N}_{\vec{0},\Sigma_i}(\vec{\theta}_{\text{syst},i}) \\ \mathcal{F}_{0,\sigma_j}(\theta_{\text{theo},j}) \quad &\longrightarrow \quad \prod_k \mathcal{F}_{0,\sigma_{kj}}(\theta_{\text{theo},kj}) \;, \end{aligned} \tag{7.19}$$

with

$$m_k \approx s_k + b_k + \sum_i \theta_{\text{syst},ki} + \sum_j \theta_{\text{theo},kj} \equiv s_k + b_k + \theta_{\text{tot},k} \;. \tag{7.20}$$

Here we assume that the theory uncertainties are uncorrelated, while the systematics can be correlated, so we need to introduce an N-dimensional Gaussian with the covariance matrices Σ_i encoding the correlations between uncertainties of category i entering different measurements k. We use either uncorrelated or fully correlated systematics.

When we compute the marginal likelihood in analogy to Eq. (7.16) the only non-trivial as-

pect are the correlated systematic uncertainties including the covariance matrix. However, the convolution of N-dimensional Gaussians still leads to one N-dimensional Gaussian, where the combined covariance matrix is the sum of the individual covariance matrices. This means, in the last step of Eq. (7.18) we are now left with an N-dimensional integral over $\theta_{\text{tot},k}$, correlated through the covariance matrix appearing in the distribution of the systematic nuisance parameters.

In SFitter, this integral is solved by approximating it with the Laplace method. This is computationally efficient and works well for cases where most of the probability is concentrated around one mode. This is the case when the nuisance parameters are Gaussians or flat. We can then write

$$\int dx^n f(x) = \int dx^n e^{\log f(x)} \,, \tag{7.21}$$

and assume that $f(x)$ has a maximum at $x = x_0$. Then one can expand $\log f(x)$ up to second order around x_0 as

$$\log f(x) \approx \log f(x_0) + \underbrace{\frac{\partial}{\partial x} \log f(x_0)}_{=0}(x - x_0) + \underbrace{\frac{\partial^2}{\partial x_i x_j} \log f(x_0)}_{=F_{ij}(x_0)}(x - x_0)_i \, (x - x_0)_j + ... \tag{7.22}$$

such that the integral is approximated by

$$\int dx^n f(x) \approx f(x_0) \sqrt{\frac{(2\pi)^n}{\det F(x_0)}} \,. \tag{7.23}$$

Note that $f(x)$ is given by the exclusive likelihood, with the maximum at $f(x_0)$ kept through profiling but not through marginalization. The matrix $F(x_0)$ is the Hessian of the log-likelihood at the maximum, i.e. the Fisher information matrix in the space of the nuisance parameters. In SFitter, x_0 is extracted with an analytic expression, approximating the Poisson distribution in Eq. (7.11) with a Gaussian. The resulting error is compensated by keeping a finite first derivative in Eq. (7.22), which in turn requires us to modify Eq. (7.23) by introducing an additional term depending on the first derivative of the log-likelihood. Both the first and second derivatives can be computed numerically. All these approximations in evaluating the exclusive and marginal likelihoods have been checked by evaluating the exclusive likelihood using Markov chains.

Validation

We can validate the implementation of the Bayesian marginalization over nuisance parameters and Wilson coefficients starting from the fully exclusive likelihood using the operator basis and dataset of Ref. [178]. The SMEFT Lagrangian is given in Eq. (7.8), but without the muon Yukawa, the top-gluon coupling $\mathcal{O}_{tG}$, and the invisible branching ratio of the Higgs. For the direct comparison we construct the marginal likelihood by profiling or marginalizing over all nuisance parameters and Wilson coefficients. We then extract the posterior probability and 68% and 95% confidence intervals. Unless otherwise specified, we assume flat, wide priors for all Wilson coefficients. This choice minimizes the impact of the prior on the final result, and we have verified that our priors on the Wilson coefficients indeed fulfill this condition. In Fig. 7.2, we show the 68% and 95%CL

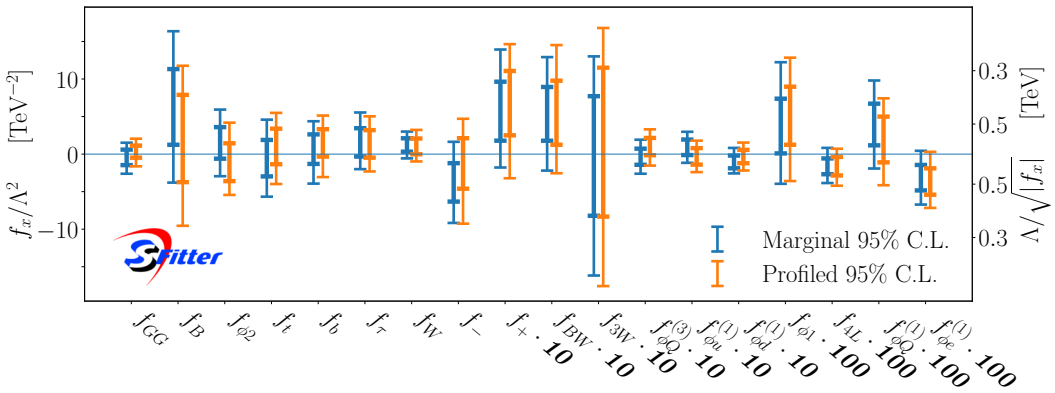


Figure 7.2: 68% and 95% confidence intervals from profile likelihoods and Bayesian marginalization. The dataset is the same as in Ref. [178].

limits from the corresponding 18-dimensional operator analysis. We see that the results of the two methods are in excellent agreement.

Going beyond confidence intervals, we can look at the distributions of the 1-dimensional profile likelihoods or marginalized probabilities. We show three examples in Fig. 7.3. Because the analysis relies on actual LHC data, the central values are not at zero Wilson coefficients. The well-measured Wilson coefficient f_W shows no difference between the profile and the marginalized results. For f_{GG}, we see a slight deviation in the central values, within one standard deviation and therefore not statistically significant. This effect points to the theory and pdf uncertainties, which we assume to be flat, and which therefore allow the central value to move freely for the profile likelihood approach, while the marginalization leads to a well-defined maximum when combining two individually flat likelihood distributions. In Fig. 7.2 we see that this difference only has a slight effect on the lower boundary when we extract 95%CL limits on f_{GG}. Finally, we see a similar effect for f_-, even though this measurement depends on several different LHC channels. According to Fig. 7.2 this is one of the largest and still not significant differences between the two methods.

The source of the differences in Fig. 7.3 can be traced back to whether the uncertainty-related nuisance parameters are marginalized or profiled. Fig. 7.4 shows that, once the uncertainty treatment is fixed, the results are independent of whether the Wilson coefficients are marginalized or profiled over.

Next, we check 2-dimensional profiled and marginalized likelihoods. Figure 7.5 shows three examples involving the same Wilson coefficients as in Fig. 7.3. First, we see that there exists an anti-correlation between f_{GG} and f_t, the modified top Yukawa also affecting the loop-induced production process $gg \to H$. This suggests that a slightly high rate measurement can be accommodated by adjusting either of the two Wilson coefficients. Because the uncertainty on this measurement includes sizable theory and pdf contributions, the same difference between the two methods can be seen for each of the two Wilson coefficients individually and for their correlation. Another instructive example is the correlation between f_W, determined from kinematic distributions, and $f_{\phi 2}$ leading to a shift in the Higgs wave function. Here the difference only appears in $f_{\phi 2}$, the parameter extracted from total rates and especially sensitive to theory uncertainties. Finally, we show the correlation between f_- and $f_{\phi Q}^{(3)}$ and observe the usual correlation from the sizable range of kinematic di-boson measurements [254].

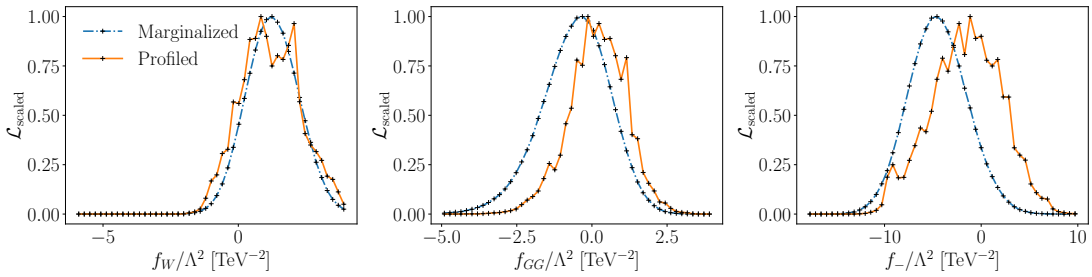


Figure 7.3: Profile likelihoods vs marginalized likelihood for a set of Wilson coefficients. The two curves are scaled such that the maximum values are at $\mathcal{L}_{\text{scaled}} = 1$.

Finally, we can check for alternative maxima in the likelihood and find that f_+ is the only Wilson coefficient exhibiting a non-trivial second mode. This can be understood from the f_+ vs f_- plane. By a numerical accident, the SMEFT corrections to all Higgs production and decay processes vanish in the SM-maximum and also close to the point $f_-/\Lambda^2 = -3$ and $f_+/\Lambda^2 = 2.7$. The only measurement which breaks this degeneracy is $H \to Z\gamma$, with limited statistical power. In the f_+ axis, the position of the maximum is fully determined by $H \to \gamma\gamma$, which is measured precisely enough to resolve the two modes, while in the f_- axis the constraints cannot distinguish the second maximum from the SM point.

Given the consistency condition of the SMEFT approach, we should not compare the two modes at face value, even though the Bayesian setup would allow for this. On the other hand, we need to confirm that this choice of modes does not affect other parameters in a significant manner once it is embedded in the 18-dimensional space. In Fig. 7.6 we show what happens if we restrict our parameter analysis to either the SM-mode or the second mode. To this end we run Markov chains mapping out both modes and then separate the samples through the condition $f_+/\Lambda^2 \lesseqgtr 2$. We see that choosing the second mode in f_+ has a small effect on f_-, pushing the best-fit closer to $f_- = -3$, but none of the other Wilson coefficients is affected. We also confirmed that both modes are of equal height by choosing a Breit-Wigner proposal function, which ensures that the Markov chains can move large distances, helping each individual chain to jump between both modes.

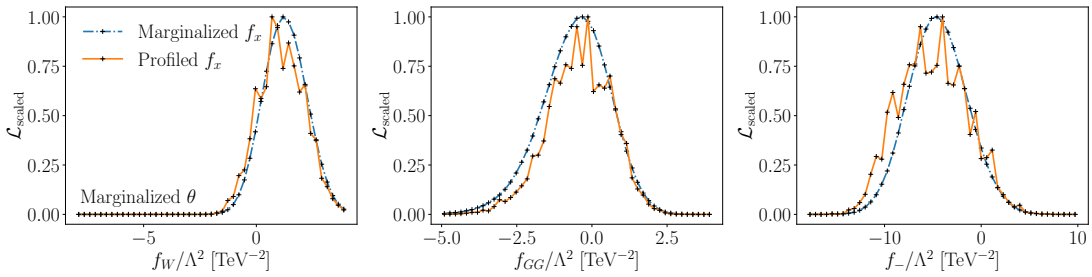


Figure 7.4: Likelihoods profiled vs marginalized over the Wilson coefficients f_x, but always marginalized over all nuisance parameters θ. We show the same Wilson coefficients as in Fig. 7.3.

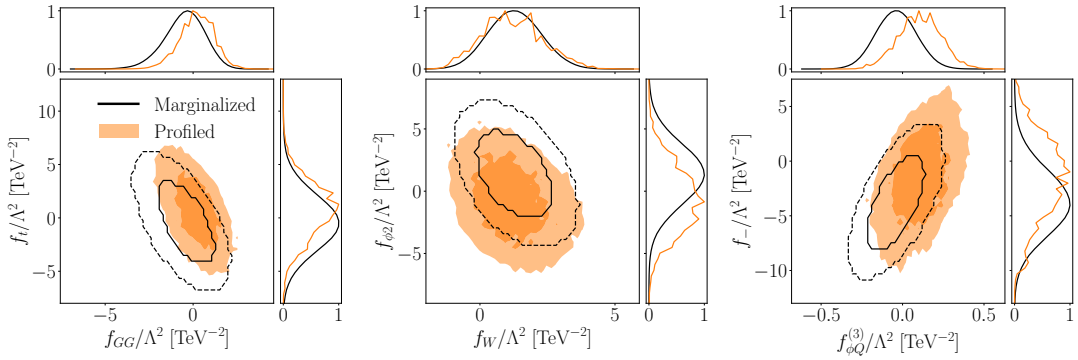


Figure 7.5: Comparison of 2-dimensional correlations of profiled and marginalized likelihoods.

Uncertainties and correlations

After confirming that the slight differences between the profile and marginalization approaches are related to the treatment of uncertainties, we can check the impact of the SFitter-specific uncertainty treatment. By default, and as explained earlier, we construct the exclusive likelihood with flat theory uncertainties and Gaussian systematics. By switching all uncertainties to Gaussian distributions we construct the completely Gaussian likelihood shown in Fig. 7.7. If we marginalize over the different uncertainties, the central limit theorem guarantees that for enough different uncertainties the results will be identical. The exact level of agreement between different uncertainty models depends on the dataset and the size of the individual uncertainties and cannot be generalized. For instance, sizable differences will appear when an outlier measurement generates a tension in the global analysis. Such a tension can be accommodated more easily using a single flat uncertainty with its reduced cost in the likelihood value.

Because the main difference between profiling and marginalizing over uncertainties appears for the flat theory uncertainties, the results from Fig. 7.7 motivate the question how relevant the theory uncertainties really are for the Run 2 dataset analyzed in Ref. [178]. We show three 1-dimensional likelihoods in Fig. 7.8 and indeed find that after marginalizing over all nuisance parameters and over all other Wilson coefficients the theory uncertainties do not play any visible role. Obviously, this statement is dependent on a given dataset, on the operators we are looking at, and on the assumed uncertainties, and it clearly does not generalize to all global Run 2 analyses.

The last effect we need to study is the impact of correlations between the different

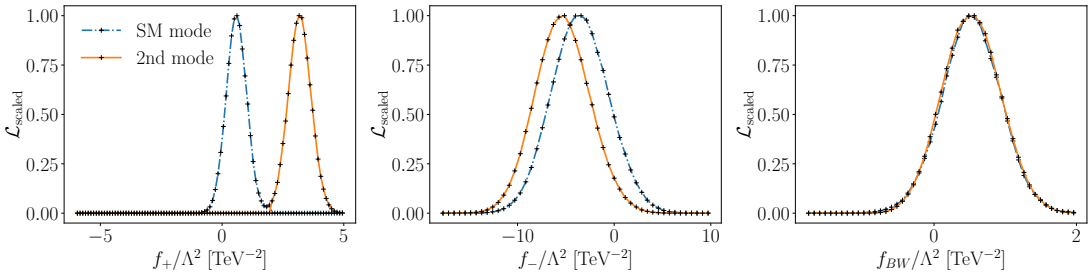


Figure 7.6: Marginalized likelihoods for the SM-like and the second mode in f_+, again for the 18-dimensional analysis.

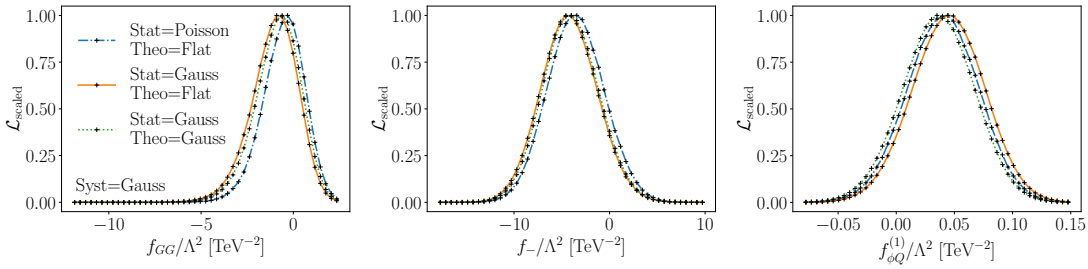


Figure 7.7: Marginalized likelihoods for different uncertainty modeling. The SFitter default is a Poisson likelihood with flat theory uncertainties and Gaussian systematics (blue dot-dashed).

uncertainties. In Fig. 7.9 we show what happens with the 1-dimensional marginalized likelihoods when we switch off all correlations between systematic uncertainties of the same kind. We see that the correlations have a much larger impact than anything else we have studied in this section. While the size of the uncertainties do not change much, the central values essentially vary freely within one standard deviation. An analogous effect was observed in Ref. [274]. We cannot emphasize enough that all statements about the validity of different approximations do not generalize to new, incoming measurements, as we will see in the following section. However, something that will not change is the key relevance of correlations as indicated by Fig. 7.9.

7.4 Updated dataset

After the detailed comparison of a profile likelihood and Bayesian SFitter approach we can, in principle, apply the numerically simpler Bayesian approach to update the SMEFT analysis of the Higgs-electroweak sector with a series of new Run 2 results. As a first step, we introduce the set of new kinematic measurements entering the updated SFitter analysis. We focus on an improved treatment of correlated uncertainties.

7.4.1 WW resonance search

Once we notice that especially boosted kinematics with large momentum transfer through Higgs interactions play a key role in SMEFT analyses [254, 306], it is clear that the reinterpretation of VH and VV resonance searches should be extremely useful for a global SMEFT analysis [178, 245]. To the best of our knowledge, SFitter is currently the only global analysis framework which includes these kinds of signatures.

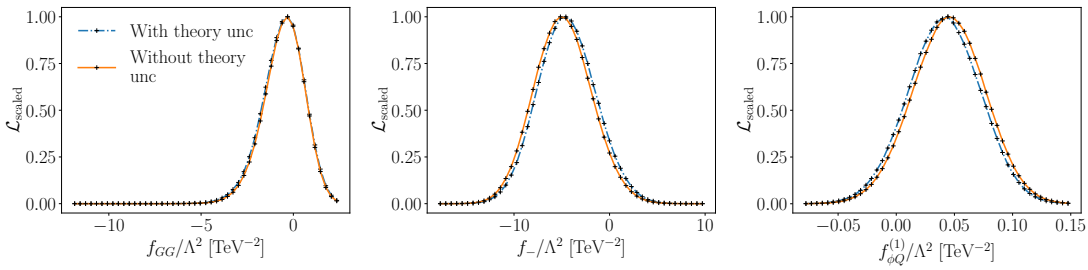


Figure 7.8: Marginalized likelihoods with and without theory uncertainties.

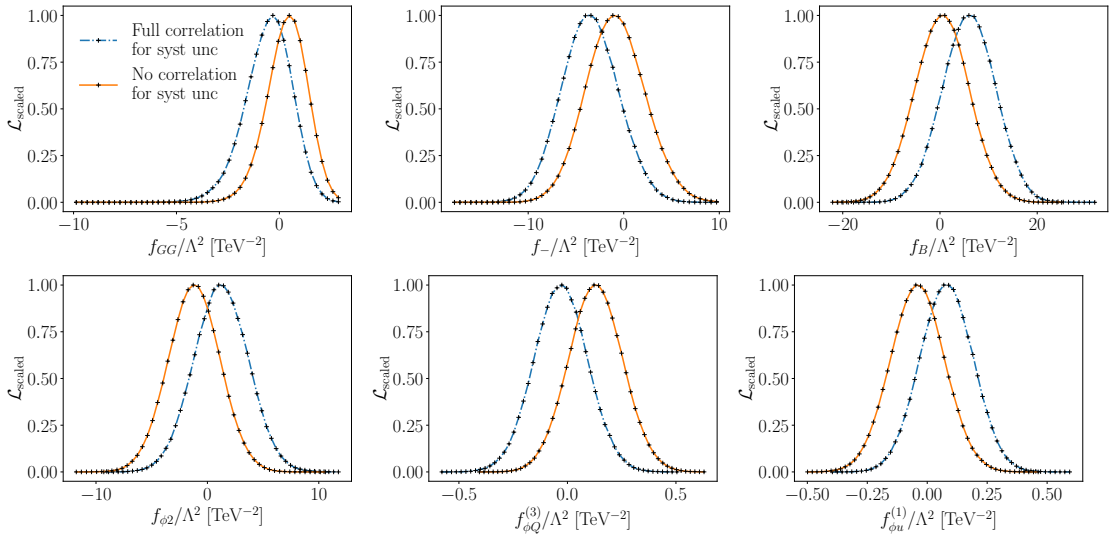


Figure 7.9: Marginalized likelihoods with and without correlations between systematic uncertainties of the same category.

First, we add the ATLAS search for resonances in the semi-leptonic VV final state [253], as briefly discussed in Ref. [3]. We only use the WW 1-lepton category in the merged Drell-Yan and gluon-fusion high-purity signal region,

$$pp \to W^+W^- \to \ell^+\nu_\ell\, jj + \ell^-\bar{\nu}_\ell\, jj\,. \tag{7.24}$$

Our signal consists of W^+W^- production modified by SMEFT operators. We neglect SMEFT effects in the leading W+jets and $t\bar{t}$ backgrounds. We include all other $W_{\ell\nu}V_{jj}$ and $Z_{\ell\ell}V_{jj}$ channels as SM-backgrounds and verified that SMEFT corrections to the other di-boson channels are sufficiently suppressed by the analysis setup.

The signal is simulated using Madgraph [161], Pythia [110], FastJet [307], and Delphes [111] with the standard ATLAS card at leading order and in the SM and requiring the lepton pair to come from an intermediate on-shell $W^\pm$. The hadronic W-decay is simulated using Pythia. Fat jets are identified using the default categorization in Delphes and ignoring the cut on the D_2 variable. The complete SM-rate is compared to the left panel of Fig. 7.10, taken from Ref. [253]. We reproduce the event selection based on the analysis cuts listed in Tab. 2 of Ref. [253]. No re-calibration of energy scales or fat-jet invariant mass windows is required, but we adjust the histogram entries by a factor 1.606 to match the ATLAS normalization of the di-boson background and accommodate efficiencies and higher-order corrections [308]. In the right panel of Fig. 7.10 we show the final m_{WW} distribution obtained with this procedure. Finally, we extract the statistical and systematic uncertainty from the ATLAS analysis, as shown in the lower panel in Fig. 7.10. Whenever backgrounds are estimated from control regions, the Gaussian systematic uncertainties are smaller than the Poisson-shaped statistical uncertainties in the signal region.

To include the VV channel in our SMEFT analysis we re-bin the original distribution such that we have a minimum of five observed events per bin. The kinematic distribution we use in SFitter is shown in the left panel of Fig. 7.11. Here all statistical uncertainties are treated as uncorrelated and added in quadrature, the same for the systematic background uncertainties linked to Monte Carlo statistics, while other systematic uncertainties are conservatively treated as fully correlated and consequently added linearly. Finally, we

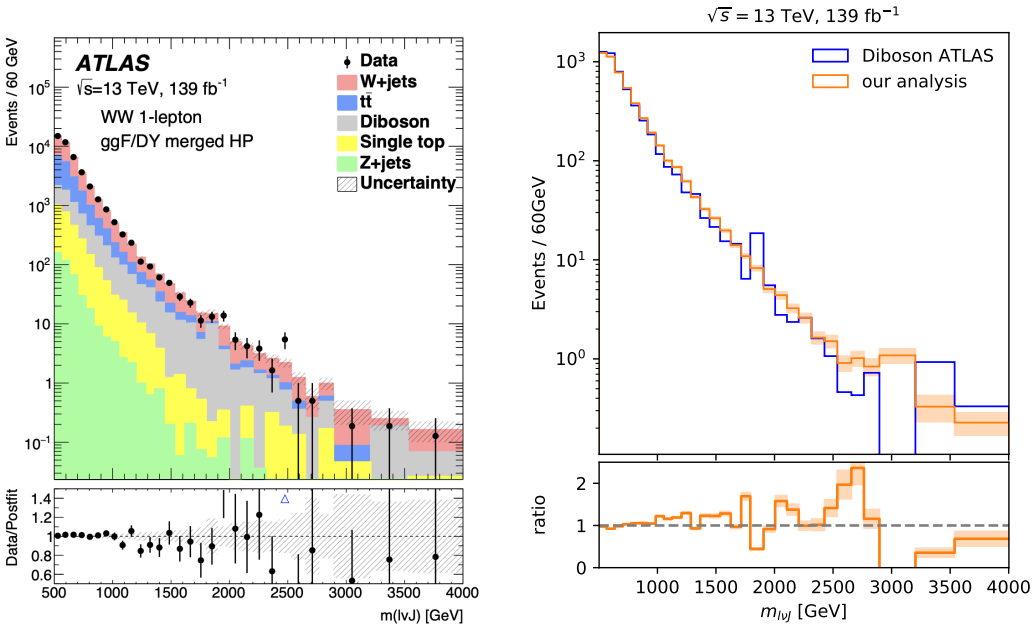


Figure 7.10: Left: measured m_{VV} distribution [253]. Right: comparison between ATLAS results and our SM background estimate. The orange band shows the statistical uncertainty from the Monte Carlo generation.

add a 80% theory uncertainty on the signal predictions in all bins and assuming no correlation among them. Of this 70% account for the uncertainties in our SMEFT Monte Carlo predictions and 10% for V+jets and single-top modeling.

In the right panel of Fig. 7.11 we show the limit in terms of the Gauss-equivalent

$$\Delta\chi^2 = \chi^2 - \chi^2_{\text{min}} = -2\log\mathcal{L} + 2\log\mathcal{L}_{\text{max}} \,, \tag{7.25}$$

extracted from different bins of the measured m_{WW} distribution. We see that the likelihood maximum slightly deviates from the SM point $f_{\phi Q}^{(3)} = 0$, and the last bin completely dominates the likelihood distribution. This is expected for momentum-enhanced operators which modify the tails of momentum distributions, as systematically analyzed in Ref. [254]. We will discuss the effect of the under-fluctuation in the last bin in more detail in Sec. 7.5.1.

7.4.2 WH resonance search

Complementing the dataset of Ref. [178] we include two new resonance searches, one described in Ref. [3] and another ATLAS analysis looking for

$$pp \to WH \to \ell\bar{\nu}_\ell\, b\bar{b} \tag{7.26}$$

at high invariant masses [309]. We focus on WH production with one b-tag, because it includes the best kinematic measurement at high m_{VH}. This analysis applies cuts on the WH topology and requires exactly one single-b-tagged fat jet. In the merged category the b-tags are part of a fat jet.

We generate di-boson events for the combined di-boson channels with lepton-hadron

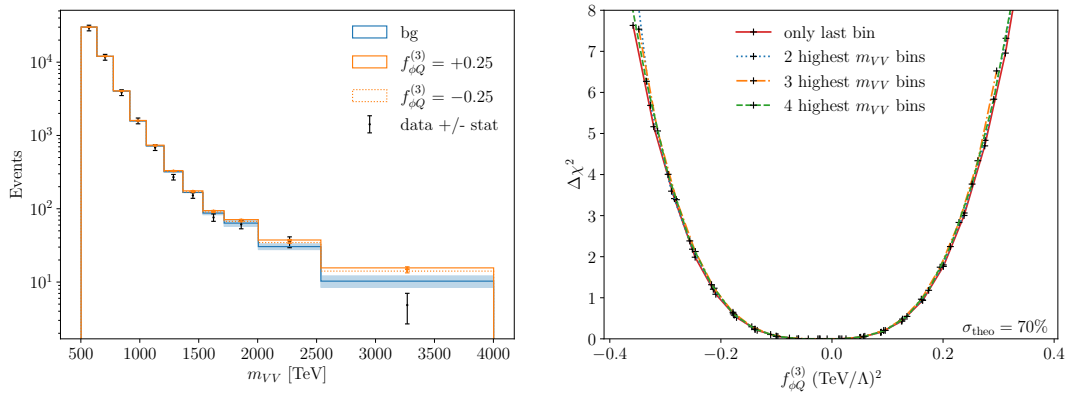


Figure 7.11: Left: re-binned m_{WW} distribution for the semi-leptonic WW analysis implemented in SFitter. We show the complete continuum background, including statistical and systematic uncertainties, and the effect of a finite Wilson coefficient $f_{\phi Q}^{(3)}$. Right: toy analysis for the same Wilson coefficient using different numbers of bins.

decays

$$pp \to W_{\ell\nu} W_{jj},\ W_{\ell\nu} Z_{jj},\ Z_{\ell\ell} W_{jj},\ Z_{\ell\ell} Z_{jj}\ , \tag{7.27}$$

again using the Madgraph-Pythia-FastJet-Delphes chain with the standard ATLAS card at leading order. They can be compared to the grey di-boson background in the left panel of Fig. 7.12, including the b-tagging and corresponding mis-tagging. After adjusting the m_{WH}-independent efficiency factor we find the agreement illustrated in the right panel of Fig. 7.12. We apply the same efficiency factor for the WH signal and then use the reweighting module in Madgraph to estimate the SMEFT rates. The W-decay to electrons or muons is included through Madgraph, while the Higgs decay to $b\bar{b}$ pairs is simulated by Pythia. We neglect SMEFT corrections to the $t\bar{t}$ and W/Z+jets backgrounds, assuming that the targeted phase space region favors the Higgs signal. Having to make this assumption is unfortunate, but we emphasize that the number of experimental measurements should prevent us from falling for SMEFT corrections canceling between the different signals and backgrounds.

To define a meaningful measurement for our global analysis we have to merge bins of the original distribution such that at least three observed events appear per bin. In Fig. 7.13, we show the actually implemented distribution for the complete SM background and including a finite Wilson coefficient $f_{\phi Q}^{(3)}$. For each bin we include a statistical uncertainty following a Poisson distribution and a Gaussian systematic uncertainty, as reported by ATLAS. In addition, we include a 13% theory uncertainty also reported by ATLAS and a theory uncertainty between 1% and 4% per bin from our SMEFT predictions, but neglecting correlation between various bins.

We can check some of our assumptions on the way we model theory uncertainties from a three-parameter analysis with $f_{\phi Q}^{(3)}$, f_W and f_{WW}. Neglecting the correlations in the theory uncertainties is justified by the left panel of Fig. 7.14. It shows the Gauss-equivalent $\Delta\chi^2$ for varying the theory uncertainties with different correlations; the orange and green lines represent a 10% and 30% theory uncertainty, fully correlated. The green line shows results without theory uncertainty, and the red line assumes our SMEFT theory uncertainty without correlations. These results are very close to each other, so we can

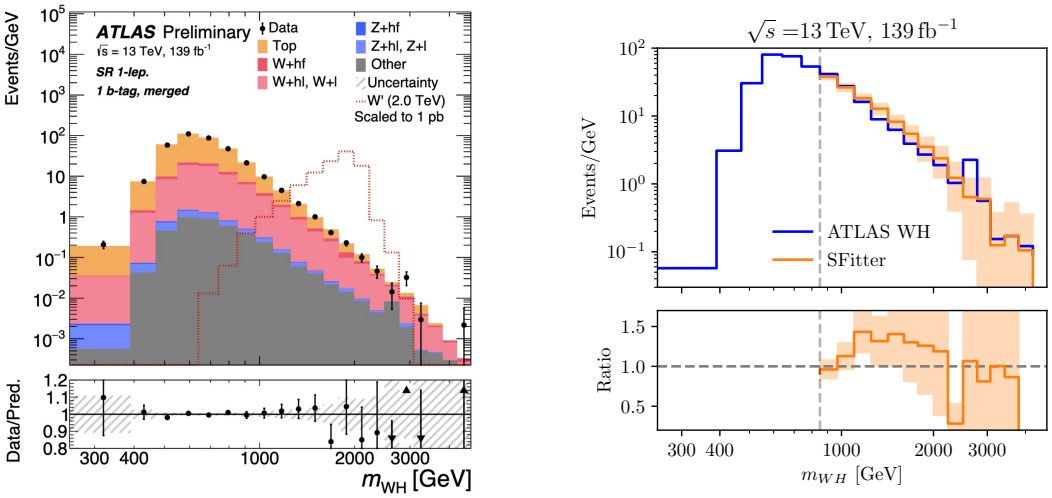


Figure 7.12: Left: measured m_{WH} distribution [309]. Right: comparison between the the ATLAS results and our SM background estimate. The orange band shows the statistical uncertainty from the Monte Carlo generation.

ignore correlations in the theory uncertainties from the EFT prediction.

The central panel compares constraints from the 3-parameter analysis from the entire m_{WH} distribution and only including one bin at a time. The limit improves sharply when the 4th and 5th bins are included. This can be understood from Fig. 7.13, where both of these bins show significant under-fluctuations. In the right panel of Fig. 7.14 we show that by removing under-fluctuations from the global analysis by setting all measured values to the number of events expected from the SM we lose constraining power. Again, demonstrating that our analysis strongly benefits from under-fluctuations.

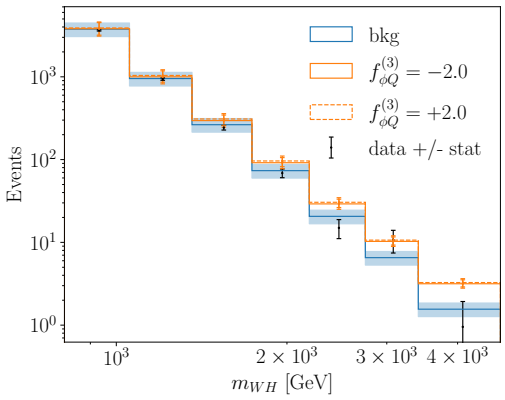


Figure 7.13: Re-binned m_{WH} distribution implemented in SFitter, including statistical and systematic uncertainties. We show the complete continuum background and the effect of a finite Wilson coefficient $f_{\phi Q}^{(3)}$.

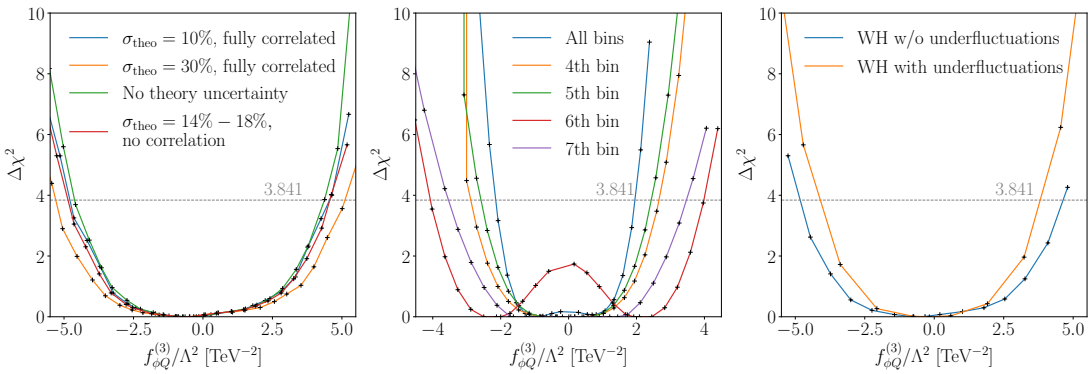


Figure 7.14: Log-likelihood for a 3-parameter analysis of the WH search as a function of $f_{\phi Q}^{(3)}$. We vary the theory uncertainties and their correlation (left), the number of bins included with uncorrelated theory uncertainties for a 1-dimensional analysis (center), and the treatment of under-fluctuations (right).

7.4.3 ZH resonance search

The second boosted VH analysis we re-interpret in terms of SMEFT is a CMS resonance search in the process [310]

$$pp \to ZH \to e^+e^- \, b\bar{b} \, . \tag{7.28}$$

We include the non-VBF category with $\leq$1 b-tags and with two b-tags. We find that the two-b category is more constraining than the $\leq 1b$ category. This can happens because the relative size of the SMEFT correction prefers this category. To determine the number of b-tags in an event, we look at the corresponding fat jet and the number of b-quarks inside the jet.

We validate our analysis simulating events for Z' peak in the heavy vector triplet model (HVT), that is used by CMS to illustrate a possible signal,

$$pp \to Z' \to Z_{\ell\ell} H_{bb} \, . \tag{7.29}$$

This signal has the advantage that it is localized in m_{ZH} and simulated at leading order using Madgraph, which means it is easier to use for calibration than a continuum background. Again, we use Madgraph, Pythia, FastJet, and Delphes with the standard CMS card at leading order. The combined sample is then compared to the HVT peak shown in Fig. 7.15. We extract the experimental efficiencies after scaling the invariant mass by the same factor 1.05 for both categories. The right panels in Fig. 7.15 show the simulated Z' signal for the two categories, compared with the quoted CMS distributions.

The SMEFT signal in the ZH channels is then computed using the same efficiencies and the reweighting module in Madgraph. The Z-decays are included in the Madgraph simulation, while the Higgs decays are simulated in Pythia. As before, we ignore SMEFT effects on the $t\bar{t}$ background.

Also for the CMS ZH channel we need to re-bin the m_{ZH} distribution to define a meaningful set of measurements, now with at least two events per bin and separately for the two categories. The results are shown in Fig. 7.16. For each bin we include the systematic and statistical uncertainties from Ref [310]. In addition, we include different theory uncertainties per bin from the SMEFT prediction and event generation in

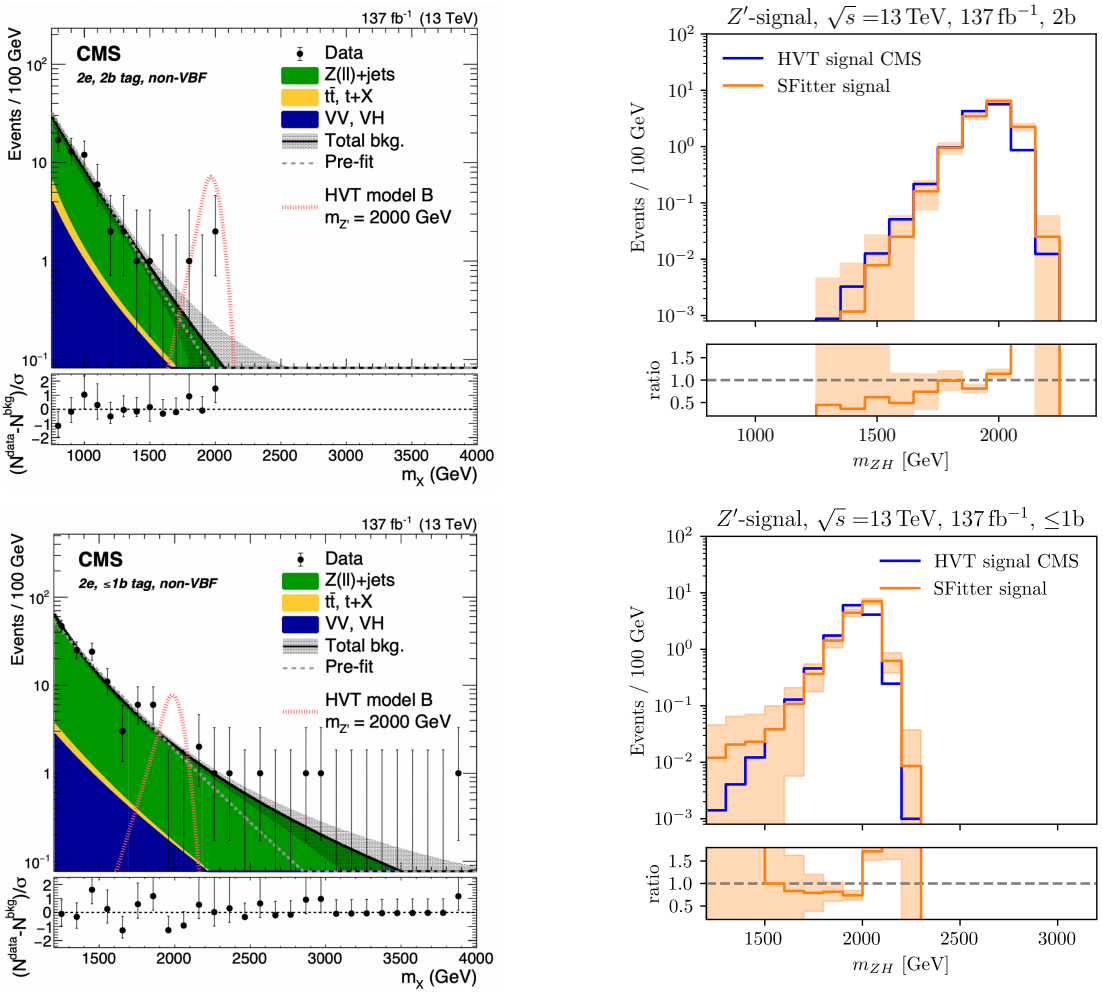


Figure 7.15: Left: measured m_{ZH} distributions for the two b-tagging categories [310]. Right: comparison between the Z' signal quoted by CMS and our estimate. The orange bands show the statistical uncertainty from the Monte Carlo generation.

Madgraph. As discussed in detail for the ATLAS WH analysis, we neglect the correlation between bins.

7.4.4 Boosted Higgs production

Boosted Higgs production, in association with one or more hard jets,

$$pp \to Hj(j) \,, \tag{7.30}$$

has been known to distinguish between a top-induced Higgs-gluon-gluon coupling and the corresponding dimension-6 operator for a long time [312, 313]. It has therefore been suggested as a channel to measure the dimension-6 Wilson coefficient f_{GG} in the presence of a modified top Yukawa coupling f_t [314–317], where it competes with channels like the off-shell Higgs production [318, 319]. In the SFitter Higgs analysis it can be added to the set of measurements to provide complementary information to the total Higgs

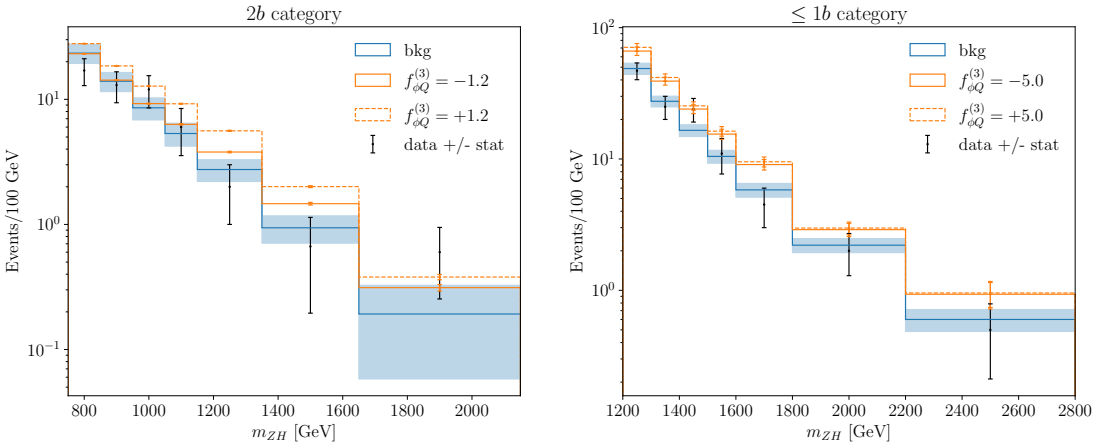


Figure 7.16: Re-binned m_{ZH} distributions for the $2b$ category (left) and the $\leq 1b$ category implemented in SFitter, including statistical and systematic uncertainties. We show the complete continuum background and the effect of a finite Wilson coefficient $f_{\phi Q}^{(3)}$.

production rate. We take the measurement of the Higgs p_T distribution in the $\gamma\gamma$ channel by ATLAS [311].

The main contribution to boosted Higgs production comes from the partonic channel $gg \to Hg$, with subleading corrections from $gg \to Hgg$. This allows us to include SMEFT corrections to $gg \to Hg$ only. They can be separated into rescalings of the top Yukawa coupling, for instance via $\mathcal{O}_{u\phi,33}$, corrections to the top-gluon coupling from $\mathcal{O}_{tG}$, and the effective Higgs-gluon interaction induced by $\mathcal{O}_{GG}$.

Because these effective vertices enter also $t\bar{t}H$ production, these operators lead to a non-trivial interplay in the global analysis. Moreover, as discussed in Sec. 7.4.5 below, f_{tG} is well-constrained by top pair production $pp \to t\bar{t}$. In fact, it constitutes the most significant contact between global top and Higgs analyses [184, 272].

We calibrate the boosted Higgs analysis simulating the SM signal for the partonic sub-channels $gg \to Hg$ and $gg \to Hgg$ using Madgraph. The gluon-initiated channels are simulated at 1-loop, while the quark-initiated one at tree level. For the one-loop simulation we use a fixed renormalization scale $\mu_R = m_H$. This setup is also used for the SMEFT simulations. Figure 7.17 shows the comparison between our simulation and the SM signal estimate provided by ATLAS. We use the same binning as in the original distribution, but omit the bins with $p_{T,\gamma\gamma} < 45$ GeV.

The simulation of SMEFT effects is tackled with different methods. The effect of a shifted top Yukawa is just a rescaling of the SM cross section, that can be easily computed analytically,

$$\frac{\sigma_{\text{SMEFT}}}{\sigma_{\text{SM}}} = \left(1 - \frac{f_t}{\sqrt{2}} \frac{v^2}{\Lambda^2}\right)^2 . \tag{7.31}$$

Second, $\mathcal{O}_{tG}$ also enters the top loops, but induces a different Lorentz structure compared to the SM amplitude. Its contributions are simulated independently using SMEFT@NLO [320] in Madgraph. In the event generation, the EFT operator is renormalized at $\mu_{\text{EFT}} = \mu_R = m_H$.

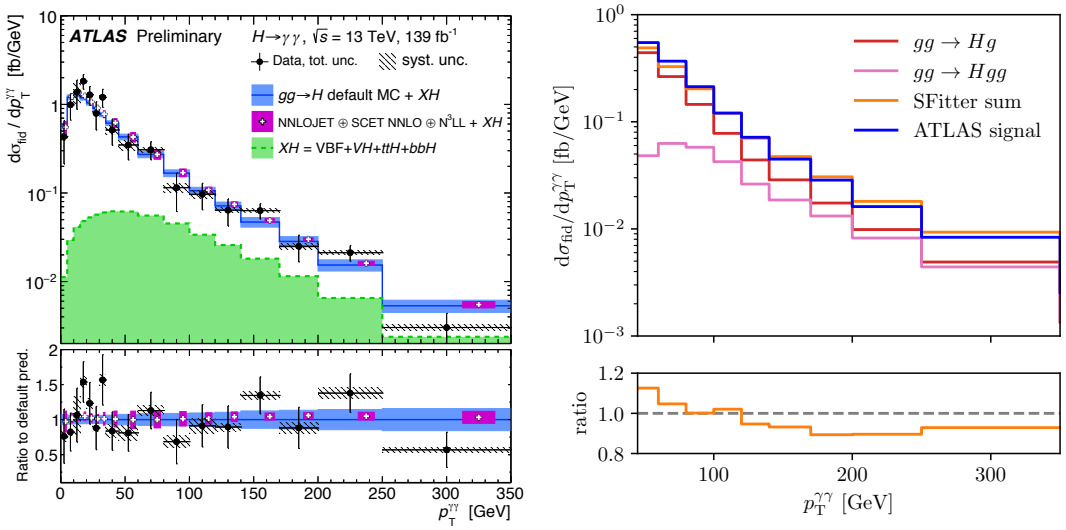


Figure 7.17: Left: measured $d\sigma_{fid}/dp_T^{\gamma\gamma}$ distribution [311]. Right: comparison between the the ATLAS distribution and our SM estimate summing contributions from $gg \to Hg$ and $gg \to Hgg$.

Finally, $\mathcal{O}_{GG}$ enters at the tree level. Because the pure interference between tree and loop diagrams cannot be generated directly in Madgraph, we choose to simulate both the linear and the squared term with a modified `loop_sm` UFO model, where the point-like Higgs-gluon vertices are mimicked by sending the bottom quark mass and Yukawa coupling to 15 TeV. We verified that any value larger than 10 TeV gives equivalent results. This way the simulation is formally at one loop for all terms. The results of this approximation were cross-checked against the analytic results in Refs. [312,313] for the interference and against the tree-level simulation for the pure square.

The mixed quadratic terms, i.e. the interferences between two operators, can be computed analytically for the combination of $\mathcal{O}_{tG}$ or $\mathcal{O}_{GG}$ with a shifted Yukawa coupling. The combination of f_{tG} and f_{GG} needs to be simulated independently, in our case using using SMEFT@NLO and the reweighting module in Madgraph.

In Fig. 7.18 we show the impact of four relevant SMEFT coefficients on the kinematic distribution we implement in SFitter. For each bin we include the systematic and statistical uncertainties from Ref [311], as well as an additional 20% theory uncertainty reflecting the scale uncertainty on the SMEFT prediction.

7.4.5 From the top

From the combined top-Higgs analyses [184,272] we know that the Higgs-gauge sector and the top sector cannot be treated completely independently. The two operators

$$\mathcal{O}_{u\phi,33} = \phi^\dagger\phi\ \bar{Q}_3\tilde{\phi}u_{R,3} \qquad \text{and} \qquad \mathcal{O}_{tG} = ig_s(\bar{Q}_3\sigma^{\mu\nu}T^A u_{R,3})\ \tilde{\phi}G^A_{\mu\nu} \tag{7.32}$$

contribute to top pair and associated $t\bar{t}H$ production and are, at the same time, crucial to interpret gluon-fusion Higgs production, together with the Higgs-related operator $\mathcal{O}_{GG}$, as discussed above. By the definition of top-sector and Higgs-sector SMEFT analyses in SFitter, $\mathcal{O}_{tG}$ is covered by the top analysis, while we keep $\mathcal{O}_{u\phi,33}$ as part of the Higgs analysis, together with a complete treatment of $t\bar{t}H$ production. This means we can

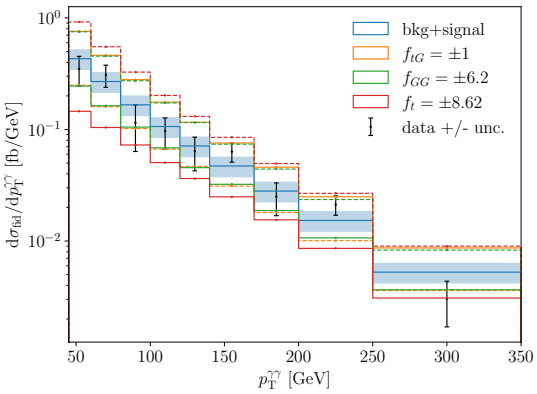


Figure 7.18: Reconstructed $p_{T,H}$ distribution implemented in SFitter, including statistical and systematic uncertainties as well as additional uncertainties on our prediction. We show the complete continuum of signal and background and the effect of three finite Wilson coefficients f_t, f_{tG} and f_{GG}. The negative values are represented by dashed lines and the positive values by solid lines.

include the limits on f_{tG} from the dedicated SFitter analysis of the top sector [246] using its 1-dimensional profile likelihood. We implement these constraints as an external measurement or prior. The corresponding profile likelihood is shown in Fig. 7.19. It consists of 100 data points which are dense enough that we can linearly interpolate between them.

We choose the range in f_{tG} to cover extremely small log-likelihoods, to avoid numerical issues in the combined analysis. Still, while it is very unlikely to occur, we also want to describe points outside of this range, so we extrapolate the log-likelihood further with two quadratic fits; one fitted to negative Wilson coefficients and one fitted to positive Wilson coefficient. A quadratic fit in this context means exponentially suppressed Gaussian tails.

7.4.6 Rates and signal strengths

In addition to the new kinematic measurements above, we update the set of Higgs rate measurements of Ref. [178], adding those listed in Tab. 7.1. The two $H \to \tau\tau$ and three out of four $H \to$ inv measurements are completely new constraints, while the

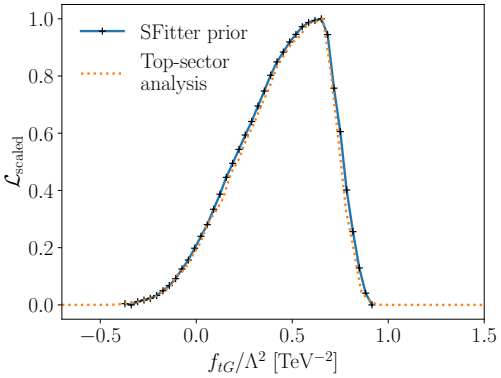


Figure 7.19: Profile likelihood for f_{tG} from the SFitter top-sector analysis [246].

others update results included in our previous analysis. The first column indicates which production channels were implemented in SFitter. We do not always use all the channels covered in a given ATLAS or CMS paper, if some of them are clearly subleading or some of them appear impossible to implement in the necessary details. Production channels in parentheses are numerically subleading, but were retained nevertheless.

The systematic and statistical uncertainties of the new measurements are typically smaller compared to the older ones. On the other hand, we attempt a more comprehensive and conservative estimate of the theory uncertainties, given the available information. In Ref. [178] we typically discarded many theory uncertainties on the signal quoted in the actual papers and replaced them with the leading uncertainty on the complete signal prediction from the HXSWG [332–334], added linearly as expected for uncorrelated flat uncertainties combined by profiling. In our new, comprehensive treatment, all theory uncertainties quoted by the analyses are retained. We include them separately and combine them. In addition, we include the uncertainties reported by the HXSWG [332–334] as the uncertainty on our SFitter prediction, again split by contribution and ready to be profiled over or marginalized.

We illustrate the implementation procedure in some more detail only for the recent Run-2 $H \to WW$ analysis by CMS [329]. Among the results presented, we implement the four signal strength measurements. Because they are reported for individual production modes (and not only in the STXS binning), they can be directly compared to the known expressions for Higgs production rates in the SMEFT, without re-deriving. These have been long implemented in SFitter for the main Higgs production channels (ggF, VBF, WH, ZH, ttH) and decays ($b\bar{b}$, WW, gg, $\tau\tau$, ZZ, $\gamma\gamma$, $Z\gamma$, $\mu\mu$). A re-derivation of the SMEFT expression can also be avoided in cases where the final results are not given for specific production channels, but the expected signal contribution from each production channel is provided.

The key ingredient to SFitter is a detailed breakdown of all uncertainties. This is crucial in order to obtain the best possible approximation of the full experimental likelihood. For Ref. [329] we consider different uncertainties for each production channel, that are reported in the paper and in the corresponding HepData entry.

Production	Decay	ATLAS	CMS
All	$H \to \gamma\gamma$	[321]	[322]
ZH	$H \to$ inv	[323]	[324]
VBF (ggF, VH)	$H \to$ inv	[325]	
VBF (ggF, ZH, $t\bar{t}H$)	$H \to$ inv		[326]
All	$H \to \tau\tau$	[327]	
VH	$H \to \tau\tau$		[328]
ggF, VBF	$H \to WW$	[309]	
ggF, VBF, VH	$H \to WW$		[329]
WH, ZH	$H \to b\bar{b}$	[330]	
ggF, VBF (VH, ttH)	$H \to \mu\mu$		[331]

Table 7.1: List of the new Run 2 Higgs measurements included in this analysis, we denote $V = W, Z$.

The statistical uncertainty is taken from the experimental paper, symmetrized and implemented as Poisson or Gaussian distribution. For experimental systematics, SFitter provides 31 predefined categories of Gaussian uncertainties, correlated across measurements and, where appropriate, across experiments. All uncertainties belonging to the same category are added in quadrature. The categories used to implement the CMS analysis cover luminosity, detector effects, lepton reconstruction, and b-tagging. Detector effects combine the jet energy scale and resolution uncertainties, as well as the missing transverse momentum scale uncertainty. Whenever the experimental papers quote significant uncertainties that do not fit any predefined category, we add them as an uncorrelated Gaussians, but this is not the case for the analysis of Ref. [329].

Theoretical uncertainties are typically implemented with flat uncorrelated likelihoods. One exception is the Monte Carlo statistics uncertainty, which we usually treat as an uncorrelated Gaussian. The CMS analysis quotes five theoretical uncertainties, that are all introduced independently. In addition, we have six theoretical uncertainties on the SFitter prediction: three on the production rate and three on the decay branching ratio, following the HXSWG prescription [332–334].

As a final step we compare the systematic uncertainties quoted on the final result with the sum of the uncertainties implemented in SFitter. If we are missing information for example on the correlations, our implementation might not be conservative, so we introduce an additional uncorrelated Gaussian uncertainty to compensate. This happens for the CMS reference analysis in the ZH channel. For this measurement we implement two uncorrelated Gaussian uncertainties, three correlated Gaussian uncertainties, plus the eleven flat uncertainties.

7.5 Global SFitter analysis

After validating the marginalization technique in SFitter and introducing a set of promising new observables, we can provide the final global analysis of the Higgs and electroweak sector after Run 2, including the leading link to the top sector. To be conservative, we will compare all our results with a profile likelihood treatment. We will find and explain differences of the two methods facing the same extended dataset.

7.5.1 Marginalization vs profiling complications

While in Sec. 7.3 we have found that for the dataset of Ref. [178] the marginalization and profiling approaches lead to, essentially, identical results, one analysis implemented in SFitter as part of Ref. [3] actually leads to significant differences. The data driving this separation of profiling and marginalization is the m_{WW} distribution measured by ATLAS [252], shown in the left panel of Fig. 7.11. It has the unique feature of a sizable under-fluctuation in the last bin.

Such an under-fluctuation is challenging to accommodate in the SMEFT. First, under-fluctuations can only be explained by operators with large interference terms, where the Wilson coefficients have to be carefully tuned to be large enough to explain a sizable effect and small enough to not be dominated by dimension-6 squared contributions. Second, a localized under-fluctuation in only one bin of one kinematic distribution requires a subtle balance of several Wilson coefficients, to control all other bins in all other di-boson and VH channels.

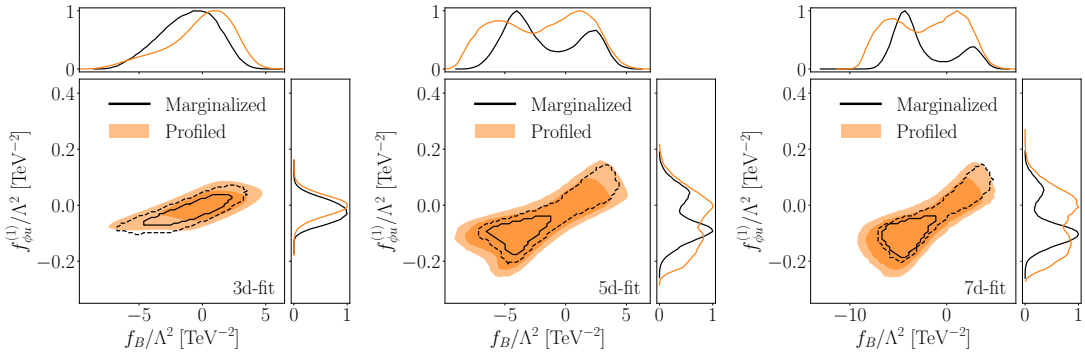


Figure 7.20: SFitter analysis with different SMEFT models describing the full Run 2 dataset, including the boosted WW production.

In Fig. 7.20 we show low-dimensional analyses of the full Run 2 dataset including the WW kinematics shown in Fig. 7.11, constraining three, five and seven Wilson coefficients. For the three parameters $\{f_B, f_{\phi u}^{(1)}, f_W\}$ we see that the maximum of the likelihood is perfectly compatible with the SM. The reason is that the SMEFT model is not flexible enough to accommodate the under-fluctuation, so we only encounter the issue when we look at the value of the likelihood in the maximum. Adding first $\{f_{\phi Q}^{(1)}, f_{\phi Q}^{(3)}\}$ and then $\{f_{\phi d}^{(1)}, f_{3W}\}$ to the SMEFT model allows us to accommodate the under-fluctuation, leading to a second likelihood maximum.

When we compare the two likelihood maxima, differences between the profiling and the marginalization appear. This is not surprising, given that the two methods ask different questions. By definition, the profile likelihood identifies the most likely parameter point, which according to Fig. 7.20 is close to the SM point, $f_B \approx 0 \approx f_{\phi u}^{(1)}$. This does not change when we increase the operator basis or expressivity of the SMEFT model. The marginalization adds volume effects in the space of Wilson coefficients, and they increasingly prefer the non-SM maximum once the SMEFT model is flexible enough to explain the under-fluctuation. Consequently, the marginalized analysis proceeds to challenge the SM in favor of an alternative SMEFT parameter point.

7.5.2 Full analysis

After identifying and understanding the issue with marginalized likelihoods for the updated dataset we now perform the full, 21-dimensional parameters analysis on all available data. The theory framework is defined by the Lagrangian in Eq. (7.8). The dataset consists of all measurements from Ref. [178], combined with the new and updated channels described in Sec. 7.4. We will discuss the standard profile likelihood results below, in a first step we focus on the marginalization. In Fig. 7.21 we show a set of 1-dimensional marginalized likelihoods. In the first row we show three Wilson coefficients affected by the under-fluctuation in m_{WW}, as discussed in the previous Sec. 7.5.1. While the marginalized likelihood for f_W follows a standard single-mode distribution, those for f_B and $f_{\phi u}^{(1)}$, for example, show two distinct modes accommodating the observed under-fluctuation.

In the second row we show the alternative maximum in f_+ we already observed for the dataset from Ref. [178] and which we discuss in Fig. 7.6 of Sec. 7.3. For the final SFitter result we will remove the second maximum as an expansion around the wrong

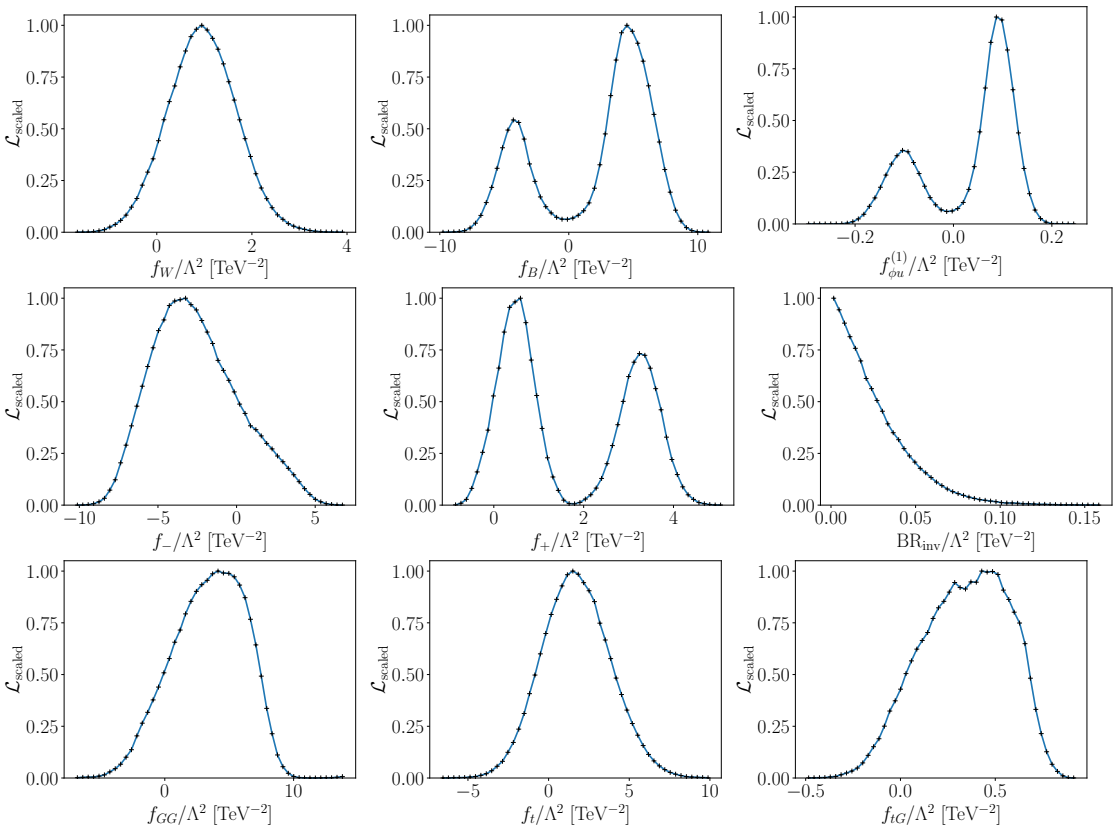


Figure 7.21: Set of marginalized likelihoods for the 21-dimensional SFitter analysis including the full set of measurements.

SMEFT limit. We also see that the invisible Higgs width is strongly constrained, even after we account for a modified Higgs production process rather than assuming SM Higgs production combined with the exotic invisible Higgs decay.

In the last row we show the effect of including $\mathcal{O}_{tG}$ in the Higgs analysis. Comparing the limit on f_{tG} to its prior in Fig. 7.19 we see that this parameter gains essentially nothing from the Higgs measurements, but it will broaden the limits on the correlated parameter

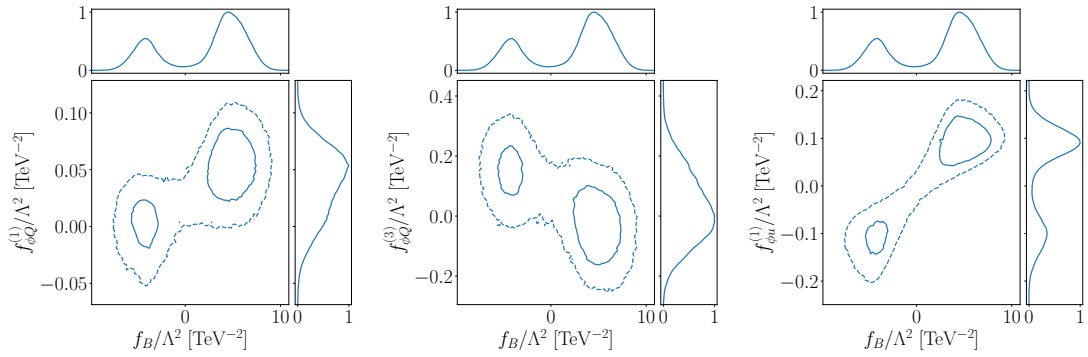


Figure 7.22: Set of marginalized correlations for the 21-dimensional SFitter analysis including the full set of measurements. The solid and dashed lines show $\Delta\chi^2 = 2$ and 7 respectively.

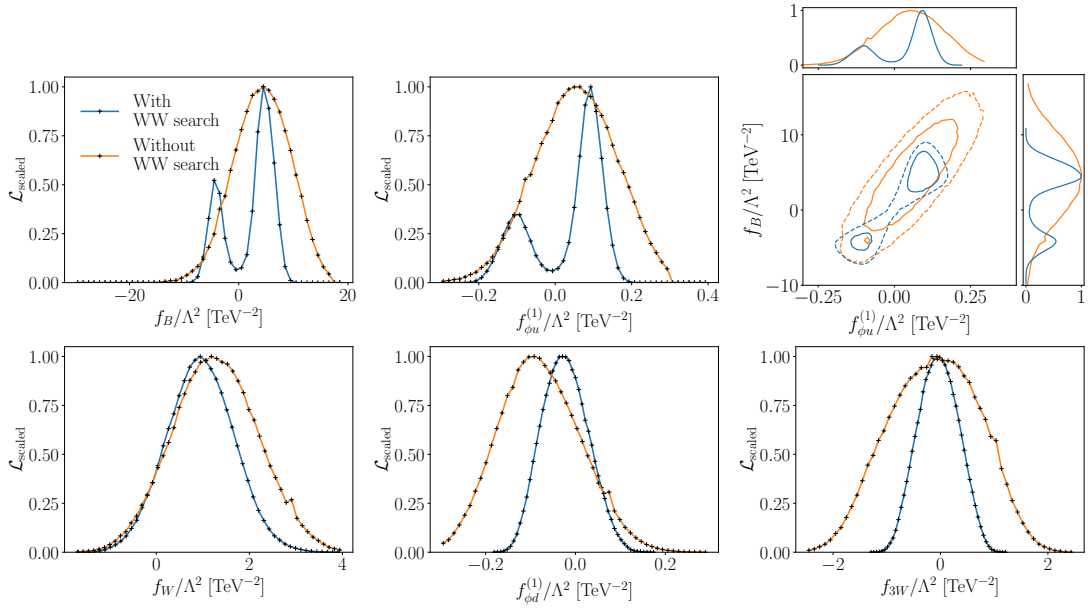


Figure 7.23: Set of marginalized likelihoods for the 21-dimensional SFitter analysis with and without the ATLAS WW resonance search altogether.

f_{GG} affecting gluon-fusion Higgs production.

To follow up on the discussion of Fig. 7.20 we show a more complete set of 2-dimensional marginalized likelihoods related to the m_{WW} under-fluctuation in Fig. 7.22. In the full analysis the correlation does not just affect $f_{\phi u}^{(1)}$, but the full range of gauge-fermion operators. This is expected from the argument that we need to carefully tune many Wilson coefficients to accommodate a deviation in a single di-boson process in a single bin of the high-invariant-mass distribution. As mentioned before, the apparent signal for physics beyond the Standard Model is an artifact of the marginalization and its volume effects, and cannot be reproduced with the profile likelihood. Note that this does not mean the marginalization is wrong or wrongly done, this difference just reflects the two methods asking different questions.

To study the impact of the critical WW-resonance analysis on our global analysis we show a set of marginalized likelihoods with and without this analysis, i.e. with and without the entire m_{WW} distribution. Obviously, removing this distribution also removes the secondary maximum structure, as we immediately see in Fig. 7.23. Removing the entire distribution replaces the marginalized likelihoods for f_B and $f_{\phi u}^{(1)}$ by their broad

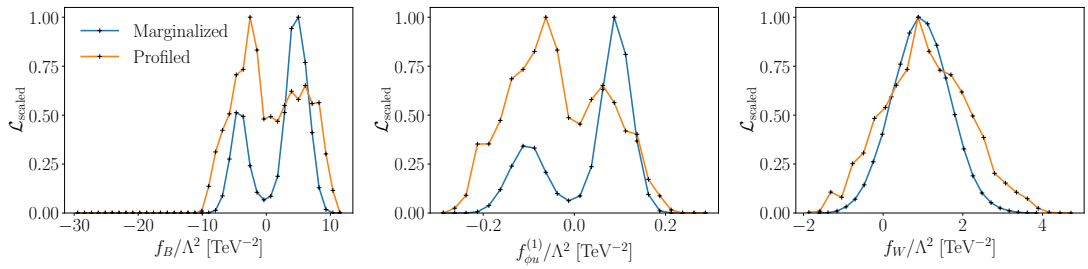


Figure 7.24: Set of marginalized and profiled likelihoods for the 21-dimensional SFitter analysis with the ATLAS WW resonance search.

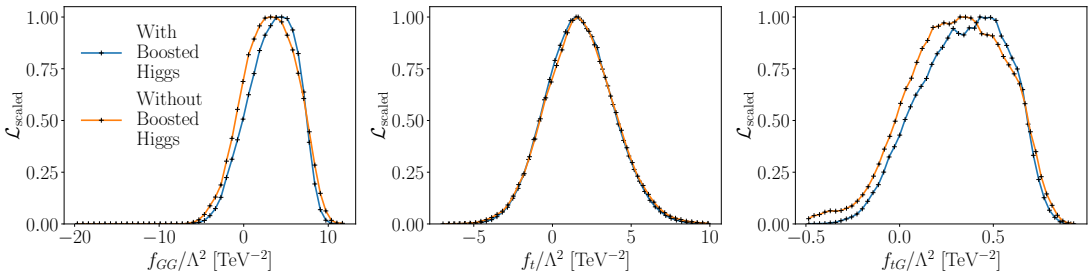


Figure 7.25: Set of marginalized likelihoods for the 21-dimensional SFitter analysis with and without the boosted Higgs analysis.

envelopes, still correlated, but without the distinctive maxima. For f_W the additional observable has limited impact, for $f_{\phi d}^{(1)}$ is leads to a smaller uncertainties combined with a shifted maximum, and for f_{3W} the non-anomalous bins of the WW-analysis provide key information.

Finally, in Fig. 7.24 we compare the 1-dimensional marginalized likelihoods with the corresponding profile likelihoods for a set of Wilson coefficients. For f_B and $f_{\phi u}^{(1)}$ we see the difference in the treatment of the secondary likelihood maximum, while f_W serves as an example for the many parameters where the two methods give the same results, as discussed in detail in Sec. 7.3 and Fig. 7.2. Indeed, the results from the two methods only disagree when the likelihoods develop secondary maxima.

Moving on with the effects observed in Fig. 7.21 we can look at the top-Higgs sector with f_{GG}, f_t, and the added f_{tG}. These three Wilson coefficients are constrained by the Higgs production in gluon fusion, associated top-Higgs production, and top pair production through the prior shown in Fig. 7.19. We have already seen that this prior is practically identical to the final outcome in Fig. 7.21. Nevertheless, we can ask what the impact of the boosted Higgs production process is, given that it should provide a second measurement of the three Wilson coefficients with different relative weights. In Fig. 7.25 we show the results of the 21-dimensional SFitter analysis with and without the new boosted Higgs measurement introduced in Sec. 7.4.4. Unfortunately, the likelihood distributions are essentially identical, corresponding to our expectation from the limited

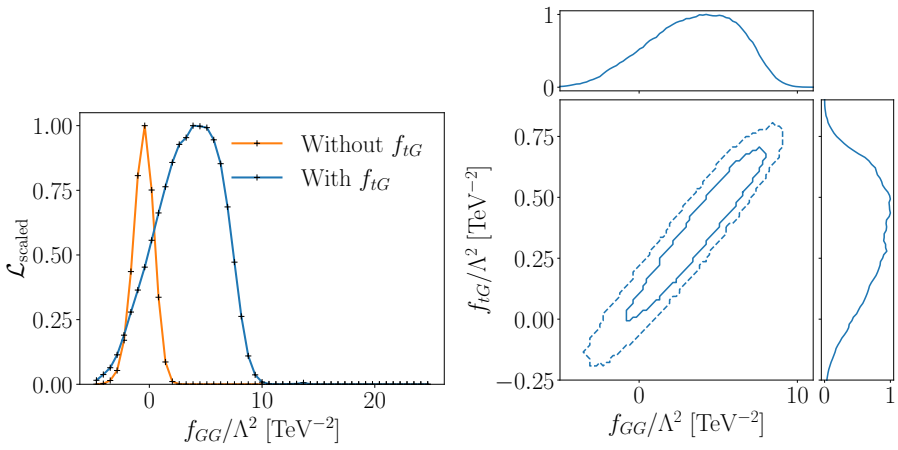


Figure 7.26: Left: marginalized likelihoods for the SFitter analysis with and without f_{tG}, using the same dataset; Right: marginalized correlation for the 21-dimensional SFitter analysis.

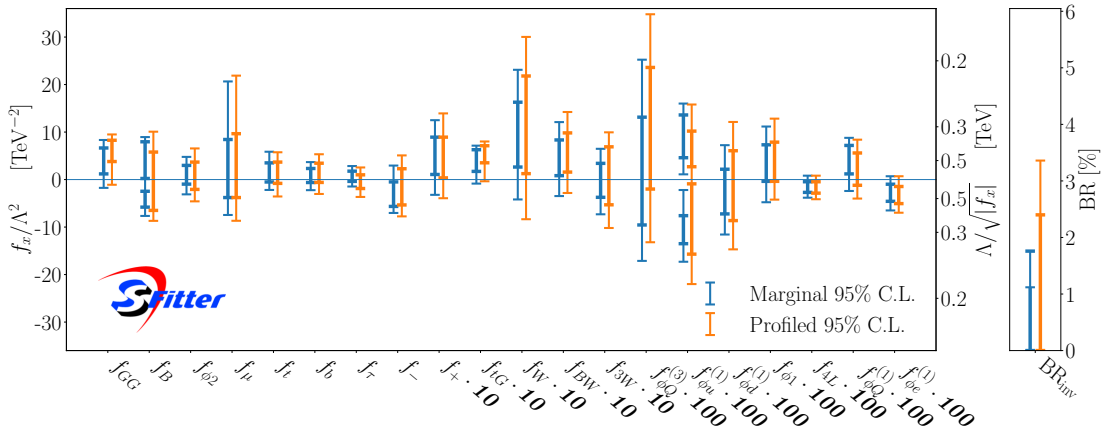


Figure 7.27: Comparison of 21- dimensional SFitter analysis with all updated measurements included. We show the 68% and 95%CL error bars from consistent marginalization and profile likelihood treatments of all nuisance parameters and Wilson coefficients. For the numerical values see Appendix B.3.

statistics of this measurements and the limited range in $p_{T,H}$, where significant differences can only be expected for $p_{T,H} > 250$ GeV [318], and even for this kinematic range it is not clear how well the measurement separates effects from f_{GG} and f_{tG}, while the f_t measurement is completely dominated by $t\bar{t}H$ production.

Even though completely justified, the only visible effect of including f_{tG} in the Higgs analysis is to wash out the limit on f_{GG}. In Fig. 7.26 we first show the change on the 1-dimensional marginalized likelihood of f_{GG} when we remove f_{tG} from the SFitter analysis. Indeed, the measurement of f_{GG} becomes much better. This is explained by the strong correlation between f_{GG} and f_{tG} shown in the right panel.

After the in-depth discussion of all features we show the 68% and 95%CL limits from the 21-dimensional SFitter analysis with the full updated dataset in Fig. 7.27. To extract these limits we start with the respective 1-dimensional marginal or profile likelihood, identify the maximum, and move outward keeping the likelihood values on the left and the right border of the integral the same. If there exists an additional peak, we compute the integral under the likelihood for the part of the curve above a given likelihood threshold. The 68% and 95%CL error bars are then defined the same way for the marginal and profile likelihood.

The profile likelihood results in Fig. 7.27 provide an update of the limits shown in Fig. 7.2 [178]. We emphasize that this update does not automatically mean an improvement of the limits, because of our more comprehensive uncertainty treatment, the added operator $\mathcal{O}_{tG}$, and the now measured Yukawa coupling f_μ. Computing the uncertainties on the Wilson coefficients which are all in agreement with the Standard Model at least for the profile likelihood approach, we remove modes around non-SM likelihood maxima. Those appear through sign flips in Yukawa couplings and in f_+ and would require order-one effects from new physics. We safely assume that new physics with this kind of effects would have been observed somewhere already.

In Fig. 7.27 we see that all results from the marginalization and profiling approach are consistent with each other. The only kind-of-significant deviation appears in f_B and the correlated gauge-fermion operators like $f_{\phi u}^{(1)}$. The reason for this discrepancy can be

traced back to an under-fluctuation in the m_{WW} measurement and actual differences between the likelihood and Bayesian approaches.

7.6 Outlook

Global SMEFT analyses are the first step into the direction of interpreting all LHC data on hard scattering process in a common framework. They allow us to combine rate and kinematic measurements from the Higgs-gauge sector, the top sector, jet production, exotics searches, even including parton densities and flavor physics. They can be considered an improved bin-wise analyses of LHC measurements, but with a consistent effective theory framework. This framework allows us to provide precision predictions matching the precision of the data we analyze, and it ensures that their result is relevant fundamental physics. Because any realistic effective theory description involves a truncation in dimensionality, SMEFT results always have to be considered in relation to the fundamental physics models they represent.

From a brief look at the analyzed data we know that our SMEFT analysis of the electroweak gauge and Higgs sector will not describe established anomalies, but serve as a consistent, global limit-setting tool. This makes it even more important to treat all uncertainties, statistical, systematic, and theory, completely and consistently. Technically, this leads us directly to the question if we want to use a profile likelihood or a Bayesian marginalization treatment. Because the two methods ask different questions, it is not at all clear that technically correct analyses following the two approaches lead to the same results. We have shown, for a first time, what the current challenges in global LHC analyses are and how the two methods do turn up slight differences.

We have started with an in-depth discussion of the current challenges in the Higgs and electroweak data and the corresponding validation of the marginalization in SFitter, in comparison to our classic profile likelihoods. Using the established dataset of Ref. [178] we have shown that the two methods give extremely similar results. We have also found that for this dataset the exact treatment of the theory uncertainties is not a leading problem, while a correct treatment of correlations of the measurements and the uncertainties is crucial.

Next, we have updated this dataset, including a set of kinematic di-boson measurements and boosted Higgs production. These measurements allow us to constrain operators with a modified Lorentz structure especially well. Kinematic distributions from di-boson resonance searches probe the largest momentum transfers of our SFitter dataset, but their interpretation in terms of SMEFT operators requires significant effort. A systematic publication of the corresponding likelihood by ATLAS and CMS would fundamentally change the appreciation for these analyses, from failed resonance searches to the most exciting SMEFT results.

Accidentally, the updated dataset also lead to differences in the marginalization and profiling treatments of the same exclusive likelihood. The measurement driving this difference is an under-fluctuation in the tail of the kinematic m_{WW} distribution. Under-fluctuations are difficult to reconcile with SMEFT analyses, because they require a balance between linear and squared operator contributions. To complicate things, a sizable number of kinematic distributions probes large momentum transfer, all consistent with the Standard Model. For a small number of Wilson coefficients one under-fluctuation will just lead to a poor log-likelihood value in the SM-like likelihood maximum. A larger

number of Wilson coefficients defines a powerful model which accommodated this deviation. For the final result, the complex correlations between Wilson coefficients lead to volume effects in the marginalization, which, expectedly, separated the final profile likelihood and marginalized results.

Chapter 8

Summary and outlook

The first part of this thesis was dedicated to BNNs and how to utilize these networks for LHC physics. In Chap. 3 we have trained a BNN on the task of jet calibration, illustrating how to understand and make use of the uncertainty estimates provided by a BNN. We have shown how one of the two uncertainties captures the limitations arising from a finite amount of training data, while the other one incorporates inherent noise of the dataset. By smearing the training data correspondingly, we have demonstrated how additional systematic uncertainties can be added. Furthermore, we went beyond a simple Gaussian likelihood by introducing a more flexible parameterization in form of a Gaussian mixture model, demonstrating how to describe non-Gaussian features.

In Chaps. 4 and 5, we discussed how to use BNNs to estimate uncertainties in the context of event generation. If classical Monte Carlo simulations based on first principles are replaced by generative models, an additional source of uncertainty is inserted. In Chap. 4, we demonstrated on simple toy examples how the uncertainty estimate of a BINN is directly related to the functional form of the true distribution by deriving and comparing to analytic expressions, demonstrating that the predicted bin-wise uncertainties are highly correlated. Furthermore, we validated the set up on the Drell-Yan process. While the process is rather simple, it can be seen as a first step in the direction of adding necessary uncertainty estimates when replacing Monte Carlo simulations with generative models.

In Chap. 5, we discussed a slightly different idea. Instead of training a generative model to sample directly from the physical distributions, we shifted the focus to only one part of the event generation chain: the computation of the matrix element. The matrix element is computationally expensive in cases of high multiplicity. The advantage of this is that the problem is turned into a much simpler regression task which has allowed us to use the same uncertainty treatment as the one discussed in Chap. 3. To utilize the predicted uncertainties further, we have introduced a feedback training by focusing the networks attention on events with large uncertainties, forcing the network to learn problematic amplitudes of divergent phase space regions.

All in all, we were able to demonstrate on these three different applications how LHC physics can benefit from machine learning in general and in particular from BNNs. While it is clear that they cannot replace the careful analysis of statistical, systematic and theory uncertainties, the predicted uncertainties of a BNN provide an additional handle to understand or improve neural network predictions – clearing the way to further integrate modern machine learning methods into current and future LHC analysis.

The second part of this thesis was dedicated to the global analysis of dimension-6 SMEFT operators including Higgs and di-boson measurements, as well as electroweak precision

observables. In Chap. 6 we investigated a gauge-triplet vector resonance model and derived the matching expressions to the SMEFT Lagrangian. We showed that the theory uncertainty due to the choice of the matching scale has significant impact on the bounds of the parameter space, revealing the importance of carefully taking all theory uncertainties into account. Furthermore, we highlighted the difference between direct searches and a SMEFT search. While the former is sensitive to a narrow resonance, a SMEFT analysis can only be used in searches sufficiently below the resonance mass. It turns out that the results obtained from the SMEFT parameterization lead to more conservative constraints compared to the direct search. This revealed that SMEFT analysis cannot replace all model-specific searches. Instead, both should be used to take full advantage of all LHC measurements.

In Chap. 7 we investigated different statistical approaches and uncertainty treatments. If we want to use the SMEFT framework to establish global limits on new physics, it is crucial to have a consistent and accurate uncertainty treatment. We found that correlations between measurements have a strong influence on our parameter bounds. Furthermore, we compared a profiled likelihood based approach with a Bayesian analysis. While both procedures answer different questions, on our established dataset of Ref. [178] we only encountered mild differences between both statistical treatments. However, after including new measurements we found a significant difference when adding the invariant mass distribution, m_{WW}, published by ATLAS in Ref. [252]. When fitting many Wilson coefficients all at once, the model is complex enough to accommodate the observed under-fluctuations. This produces a large part in the parameter space preferred by the measurements but away from the SM. When integrating over this space, the marginal likelihood collects a large amount of volume, resulting in a disfavored SM point. In contrast, the profiled likelihood does not encounter the same volume effect, leading to a significant difference between both statistical approaches.

The two studies revealed how important a careful uncertainty treatment in a global SMEFT analysis is. If we want to get accurate constraints on the space of new physics, we have to carefully think about all sources of theoretical uncertainties and model correlations between measurements as preciously as possible.

Chapter A

Appendix: Bayesian neural network studies

A.1 Jet calibration – comparison to smeared data

A slight modification of this appendix appeared in Ref. [1]. The text and figures are mostly identical to the ones presented in this publication.

To further validate the proposed approach of Chap. 3, Fig. A.1 compares the performance of the BNN approach with a more traditional smearing of the input objects. For smearing the objects we use a Bayesian neural network trained on data without smearing and evaluate this network on a test dataset with modified inputs. Each jet in the test sample is smeared once up and once down, then the difference of the two network outputs is evaluated and divided by two. We then show the average in the given $p_{T,j}$-range. The BNN prediction is in good agreement with modified inputs, giving additional confidence in uncertainty predicted by the Bayesian network.

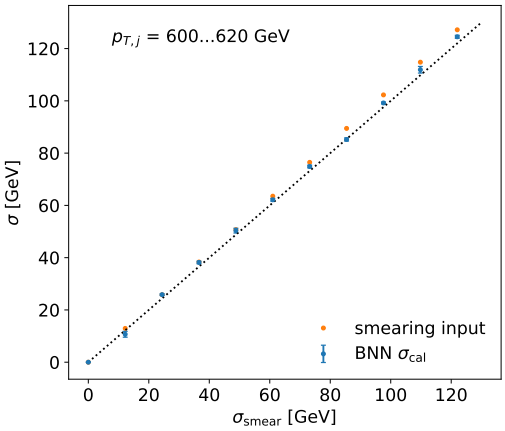


Figure A.1: Comparison of the Bayesian approach, taken from Fig. 3.9 (blue), and smearing of input data (orange). When smearing the input data, we train a Bayesian network on nominal events and test it on inputs modified up and down by σ_{smear}.

A.2 Event generation – $(2 \to 4)$-process

A slight modification of this appendix appeared in Ref. [4]. The text and figures are mostly identical to the ones presented in this publication.

To further validate the Bayesian network approach of Chap. 5, we tested our set up on the $(2 \to 4)$-process

$$gg \to \gamma\gamma gg \; . \tag{A.1}$$

We know that that increasing the number of particles in the final state leads to a significant drop in network performance [108, 121, 123, 167, 168]

For this process we use a network with seven hidden layers, 24 kinematic input dimensions, $\{32, 64, 256, 512, 128, 64, 32\}$ nodes, and two output dimensions corresponding to the amplitude and its uncertainty. This larger network has around 600k parameters. The training dataset contains around 90k amplitudes. Aside from these changes, we apply the same basic BNN training with two levels of loss boosting and process-specific performance boosting.

As for the $(2 \to 3)$-process we first show the performance of the network training in Fig. A.2. While the overall scale of the agreement has increased for the sub-percent level to the percent level, we still see that the network has learned the largest amplitudes

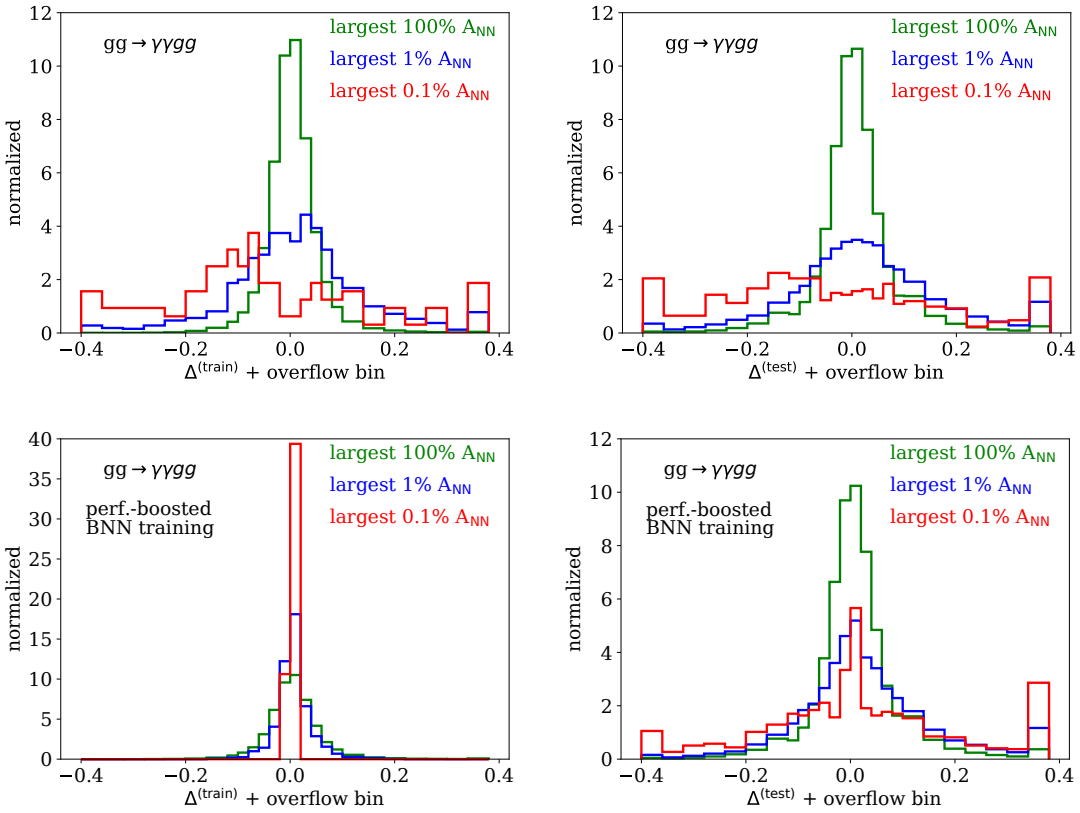


Figure A.2: Performance of the basic (upper) and performance-boosted (lower) BNNs for the $(2 \to 4)$-process in terms of the precision of the generated amplitudes, Eq. (5.3), evaluated on the training (left) and test datasets (right).

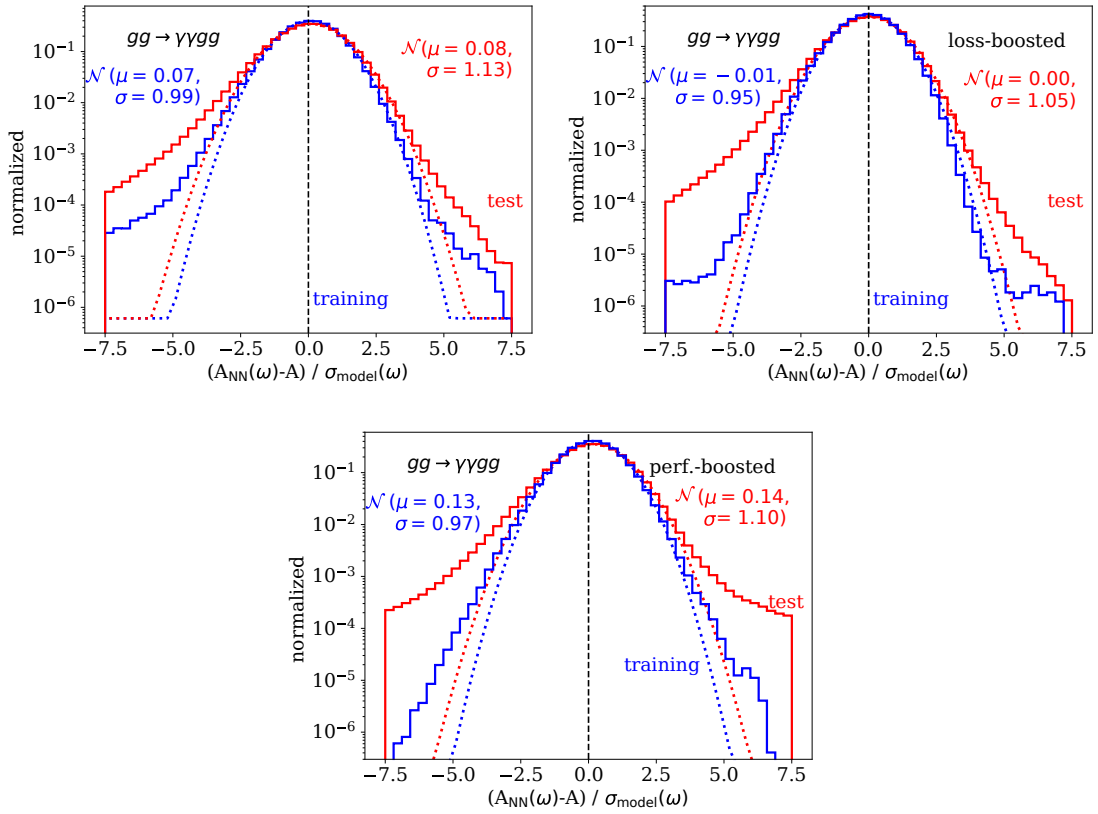


Figure A.3: Pulls of the loss-boosted (left), and performance-boosted (right) BNN for the $(2 \to 4)$-process, defined in Eq. (5.6) and evaluated on the training and test data.

extremely well after process boosting. Unlike for the standard BNN, there is a certain amount of overtraining after performance-boosting, indicating the shift from a fit-like network training to an interpolation-like training.

Next, we check the consistency of the network output by looking at the ω-dependent pull distribution defined Eq. (5.6). We see that especially for large amplitudes the loss-boosting guarantees a well-behaved, consistent network, while the additional process boosting reverses some of the beneficial effects of the loss-boosting. This effect was already observed for the $(2 \to 3)$-process.

Finally, we show the 1-dimensional kinematic distributions for the basic BNN and for the performance-boosted BNN. As for the $(2 \to 3)$-process the boosting step has a spectacular effect on the training data in the poorly learned kinematic tails. After integrating over the additional phase space directions this improvement translates well into the test dataset, but at the expense of the uncertainty estimate on the training data.

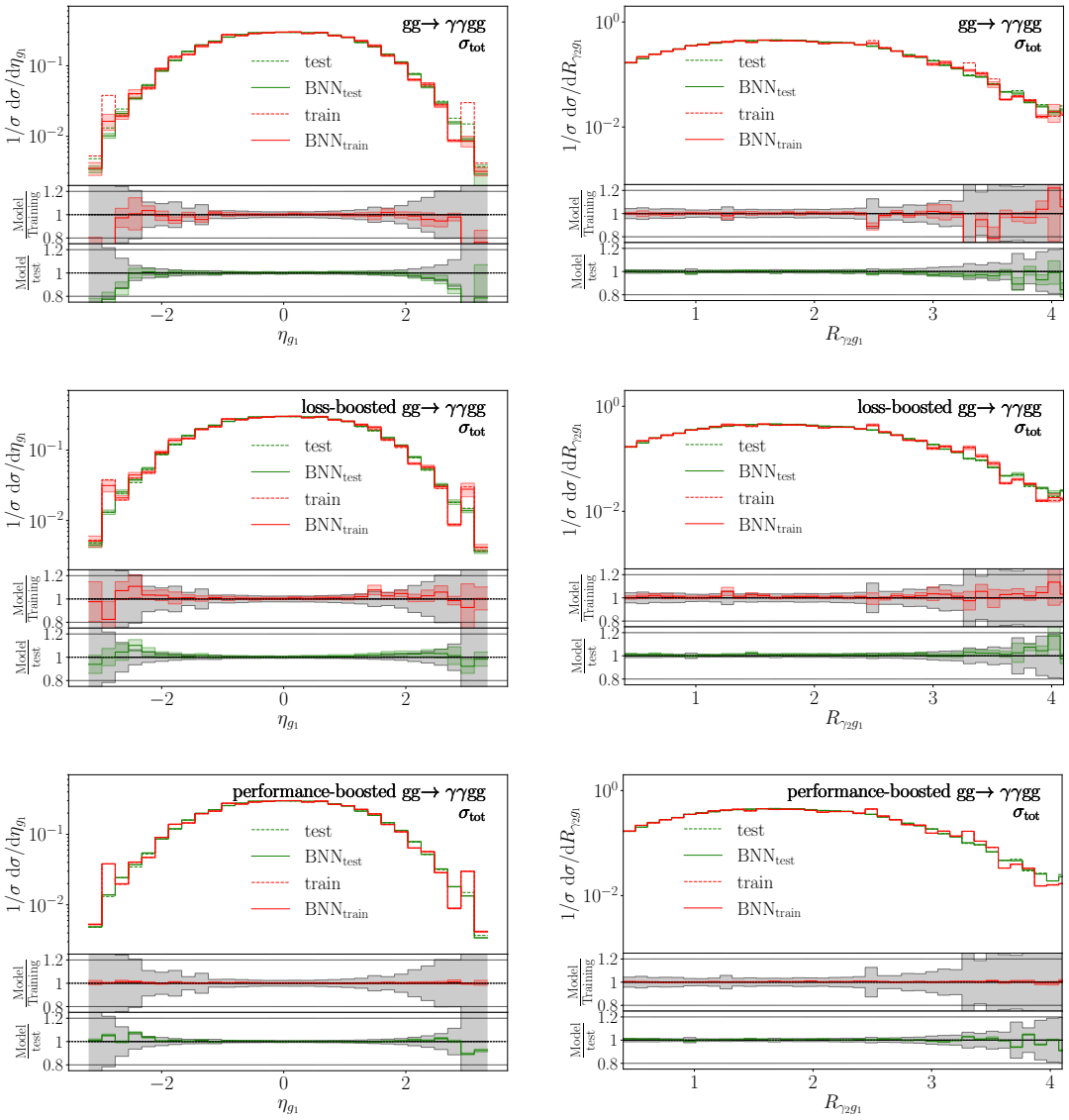


Figure A.4: Kinematic distribution for the $(2 \to 4)$-process without boosting (upper), after loss boosting (center), and after process boosting (lower). The grey error bars in the lower panels indicate the statistical limitation of the training and test data.

Chapter B

Appendix: global SMEFT analysis

B.1 Operator basis of Chapter 6

A slight modification of this appendix appeared in Ref. [3]. The text is mostly identical to the one presented in this publication.

In Chap. 6 we consider the dimension-6 SMEFT Lagrangian

$$\begin{aligned}\mathcal{L}_{\text{SMEFT}} \supset & -\frac{\alpha_s}{8\pi}\frac{f_{GG}}{\Lambda^2}\mathcal{O}_{GG} + \frac{f_{WW}}{\Lambda^2}\mathcal{O}_{WW} + \frac{f_{BB}}{\Lambda^2}\mathcal{O}_{BB} + \frac{f_{BW}}{\Lambda^2}\mathcal{O}_{BW} + \frac{f_W}{\Lambda^2}\mathcal{O}_W + \frac{f_B}{\Lambda^2}\mathcal{O}_B \\ & + \frac{f_{WWW}}{\Lambda^2}\mathcal{O}_{WWW} + \frac{f_{\phi 1}}{\Lambda^2}\mathcal{O}_{\phi 1} + \frac{f_{\phi 2}}{\Lambda^2}\mathcal{O}_{\phi 2} + \frac{f_\tau m_\tau}{v\Lambda^2}\mathcal{O}_\tau + \frac{f_b m_b}{v\Lambda^2}\mathcal{O}_b + \frac{f_t m_t}{v\Lambda^2}\mathcal{O}_t + \\ & + \frac{f_{LLLL}}{\Lambda^2}\mathcal{O}_{LLLL} + \frac{f_{\phi e}}{\Lambda^2}\mathcal{O}_{\phi e} + \frac{f_{\phi d}}{\Lambda^2}\mathcal{O}_{\phi d} + \frac{f_{\phi u}}{\Lambda^2}\mathcal{O}_{\phi u} + \frac{f^{(1)}_{\phi Q}}{\Lambda^2}\mathcal{O}^{(1)}_{\phi Q} + \frac{f^{(3)}_{\phi Q}}{\Lambda^2}\mathcal{O}^{(3)}_{\phi Q}\,,\end{aligned} \tag{B.1}$$

$\mathcal{O}_{GG} = \phi^\dagger\phi G^a_{\mu\nu}G^{a\mu\nu}$	$\mathcal{O}_{BW} = \phi^\dagger\hat{B}_{\mu\nu}\hat{W}^{\mu\nu}\phi$
$\mathcal{O}_{BB} = \phi^\dagger\hat{B}_{\mu\nu}\hat{B}^{\mu\nu}\phi$	$\mathcal{O}_{WW} = \phi^\dagger\hat{W}_{\mu\nu}\hat{W}^{\mu\nu}\phi$
$\mathcal{O}_B = (D_\mu\phi)^\dagger\,\hat{B}^{\mu\nu}\,(D_\nu\phi)$	$\mathcal{O}_W = (D_\mu\phi)^\dagger\,\hat{W}^{\mu\nu}\,(D_\nu\phi)$
$\mathcal{O}_{WWW} = \mathrm{Tr}\left(\hat{W}_{\mu\nu}\hat{W}^{\nu\rho}\hat{W}_\rho{}^\mu\right)$	
$\mathcal{O}_{\phi 1} = (D_\mu\phi)^\dagger\,\phi\phi^\dagger\,(D^\mu\phi)$	$\mathcal{O}_{\phi 2} = \frac{1}{2}\partial^\mu\left(\phi^\dagger\phi\right)\partial_\mu\left(\phi^\dagger\phi\right)$
$\mathcal{O}_b = (\phi^\dagger\phi)\,\bar{q}_3\phi d_3$	$\mathcal{O}_\tau = (\phi^\dagger\phi)\,\bar{l}_3\phi e_3$
$\mathcal{O}_t = (\phi^\dagger\phi)\,\bar{q}_3\widetilde{\phi}u_3$	
$\mathcal{O}_{LLLL} = \left(\bar{l}_1\gamma_\mu l_2\right)\left(\bar{l}_2\gamma^\mu l_1\right)$	$\mathcal{O}_{\phi e} = (\phi^\dagger i\overleftrightarrow{D}_\mu\phi)\,(\bar{e}_i\gamma^\mu e_j)\,\delta^{ij}$
$\mathcal{O}_{\phi d} = (\phi^\dagger i\overleftrightarrow{D}_\mu\phi)\left(\bar{d}_i\gamma^\mu d_j\right)\delta^{ij}$	$\mathcal{O}_{\phi u} = (\phi^\dagger i\overleftrightarrow{D}_\mu\phi)\,(\bar{u}_i\gamma^\mu u_j)\,\delta^{ij}$
$\mathcal{O}^{(1)}_{\phi Q} = (\phi^\dagger i\overleftrightarrow{D}_\mu\phi)\,(\bar{q}_i\gamma^\mu q_j)\,\delta^{ij}$	$\mathcal{O}^{(3)}_{\phi Q} = (\phi^\dagger i\overleftrightarrow{D^A_\mu}\phi)\left(\bar{q}_i\gamma^\mu t^A q_j\right)\delta^{ij}$

Table B.1: Basis of dimension-6 SMEFT operators adopted in our global analysis. Flavor indices are denoted by i, j and are implicitly contracted when repeated.

where the Wilson coefficients are denoted by f_i. We use the dimension-6 operator basis of Ref. [178], which is based on the HISZ set [249] and defined in Tab. B.1. We adopt the "+" convention for the covariant derivatives, e.g. $D_\mu\phi = (\partial_\mu + ig'B_\mu/2 + igt^A W_\mu^A)\phi$, where $t^A = \sigma^A/2$ are the $SU(2)$ generators and σ^A the Pauli matrices. We have also defined $(\phi^\dagger i\overleftrightarrow{D}_\mu\phi) = i\phi^\dagger(D_\mu\phi) - i(D_\mu\phi^\dagger)\phi$, $(\phi^\dagger i\overleftrightarrow{D}^I_\mu\phi) = i\phi^\dagger t^A(D_\mu\phi) - i(D_\mu\phi^\dagger)t^A\phi$ and the dual Higgs field $\widetilde{\phi} = i\sigma^2\phi^\star$. The field strengths are normalized as $\hat{B}_{\mu\nu} = ig'B_{\mu\nu}/2$ and $\hat{W}_{\mu\nu} = igt^A W^A_{\mu\nu}$. Finally, the operators $\mathcal{O}^{(1),(3)}_{\phi Q}, \mathcal{O}_{\phi u}, \mathcal{O}_{\phi d}, \mathcal{O}_{\phi e}$ are defined in a $U(3)^5$-invariant flavor structure, while for $\mathcal{O}_{LLLL}$ we only retain the (1221) contraction, that is relevant for the definition of the Fermi constant, and for $\mathcal{O}_b, \mathcal{O}_t, \mathcal{O}_\tau$ we only consider the 3rd fermion generation. The latter choice is justified considering that, in a $U(3)^5$-symmetric scenario, these operators are weighted by a Yukawa coupling insertion, that acts as a suppression for the first two families.

The matching to the UV models described in Sec. 6.2.1 is automated for the Warsaw basis of SMEFT operators [281], in the general flavor case. The results obtained are provided on github at [250] and we give explicit expressions for the tree-level matching in Appendix B.1. In order to interface them to SFitter, the matching results are mapped onto the basis of Tab. B.1. In the following we denote the operators in the Warsaw basis, defined as in Ref. [281], by Q_k and the associated Wilson coefficients by C_k, such that the SMEFT Lagrangian in this basis has the form

$$\mathcal{L}_{\text{Warsaw}} \supset \frac{1}{\Lambda^2}\sum_k \sum_{ij} C_{k,ij}\, Q_{k,ij}\,, \tag{B.2}$$

where k runs over the operators labels and i, j are flavor indices, that are present for fermionic operators. The relations between the two operator bases are

$$\begin{aligned}
&\mathcal{O}_{GG} = Q_{\phi G}\,, && \mathcal{O}_{WWW} = \frac{g^3}{4}Q_W\,, && \\
&\mathcal{O}_{BB} = -\frac{g'^2}{4}Q_{\phi B}\,, && \mathcal{O}_{WW} = -\frac{g^2}{4}Q_{\phi W}\,, && \mathcal{O}_{BW} = -\frac{gg'}{4}Q_{\phi WB}\,,\\
&\mathcal{O}_{\phi 1} = Q_{\phi D}\,, && \mathcal{O}_{\phi 2} = -\frac{1}{2}Q_{\phi\Box}\,, && \mathcal{O}_{\phi} = Q_{\phi}\,,\\
&\mathcal{O}_{\tau} = Q_{e\phi,33}\,, && \mathcal{O}_{t} = Q_{u\phi,33}\,, && \mathcal{O}_{b} = Q_{d\phi,33}\,,\\
&\mathcal{O}_{\phi e} = Q_{\phi e,ij}\,\delta^{ij}\,, && \mathcal{O}_{\phi u} = Q_{\phi u,ij}\,\delta^{ij}\,, && \mathcal{O}_{\phi d} = Q_{\phi d,ij}\,\delta^{ij}\,,\\
&\mathcal{O}^{(1)}_{\phi Q} = Q^{(1)}_{\phi q,ij}\,\delta^{ij}\,, && \mathcal{O}^{(3)}_{\phi Q} = \frac{1}{4}Q^{(3)}_{\phi q,ij}\,\delta^{ij}\,, && \mathcal{O}_{LLLL} = Q_{ll,1221}\,,
\end{aligned} \tag{B.3}$$

and

$$\begin{aligned}
\mathcal{O}_W &= \frac{g^2}{8}Q_{\phi W} + \frac{g'g}{8}Q_{\phi WB} - \frac{3g^2}{8}Q_{\phi\Box} + \frac{g^2 m_h^2}{4}(\phi^\dagger\phi)^2 - \frac{g^2\lambda}{2}Q_\phi\\
&\quad - \frac{g^2}{4}\left[(Y_e)_{ij}Q_{e\phi,ij} + (Y_u)_{ij}Q_{u\phi,ij} + (Y_d)_{ij}Q_{d\phi,ij} + \text{h.c.}\right] - \frac{g^2}{8}\left(Q^{(3)}_{\phi q,ij} + Q^{(3)}_{\phi l,ij}\right)\delta^{ij}\\
\mathcal{O}_B &= \frac{g'^2}{8}Q_{\phi B} + \frac{gg'}{8}Q_{\phi WB} - \frac{g'^2}{2}Q_{\phi D} - \frac{g'^2}{8}Q_{\phi\Box}\\
&\quad - \frac{g'^2}{4}\left(\frac{1}{6}Q^{(1)}_{\phi q,ij} - \frac{1}{2}Q^{(1)}_{\phi l,ij} + \frac{2}{3}Q_{\phi u,ij} - \frac{1}{3}Q_{\phi d,ij} - Q_{\phi e,ij}\right)\delta^{ij}\,,
\end{aligned} \tag{B.4}$$

where all repeated flavor indices are implicitly summed over, and λ is the quartic coupling

in the Higgs potential, normalized such that

$$V(\phi) = -\frac{m_h^2}{2}\phi^\dagger\phi + \frac{\lambda}{2}(\phi^\dagger\phi)^2 \,. \tag{B.5}$$

As the vector triplet model we are interested in is defined in a flavor-symmetric limit, after the matching procedure the Wilson coefficients of the Warsaw basis operators $Q_{\phi e,\phi u,\phi d}$ and $Q^{(1),(3)}_{\phi l,\phi q}$ will have the form

$$C_{\phi\psi,ij} = \bar{C}_{\phi\psi}\,\delta_{ij} \,, \tag{B.6}$$

while

$$C_{ll,ijkl} = \bar{C}_{ll}\delta_{ij}\delta_{kl} + \bar{C}'_{ll}\delta_{il}\delta_{kj} \,. \tag{B.7}$$

Using this notation, the mapping in terms of Wilson coefficients is

$$\begin{aligned}
f_B &= \frac{8}{g'^2}\bar{C}^{(1)}_{\phi l} & -\frac{\alpha_s}{8\pi}f_{GG} &= C_{\phi G} \\
f_W &= -\frac{8}{g^2}\bar{C}^{(3)}_{\phi l} & f_{WWW} &= \frac{4}{g^3}C_W \\
f_{BB} &= -\frac{4}{g'^2}\left[C_{\phi B} - \bar{C}^{(1)}_{\phi l}\right] & f_{\phi 1} &= C_{\phi D} + 4\bar{C}^{(1)}_{\phi l} \\
f_{WW} &= -\frac{4}{g^2}\left[C_{\phi W} + \bar{C}^{(3)}_{\phi l}\right] & f_{\phi 2} &= -2C_{\phi\Box} - 2\bar{C}^{(1)}_{\phi l} + 6\bar{C}^{(3)}_{\phi l} \\
f_{BW} &= 4\left[-\frac{C_{\phi WB}}{gg'} - \frac{\bar{C}^{(3)}_{\phi l}}{g^2} + \frac{\bar{C}^{(1)}_{\phi l}}{g'^2}\right] & f_\phi &= C_\phi - 4\lambda\bar{C}^{(3)}_{\phi l}
\end{aligned} \tag{B.8}$$

and for the fermionic ones

$$\begin{aligned}
\frac{m_\tau}{v}f_\tau &= C_{e\phi,33} - 2(Y_e)_{33}\bar{C}^{(3)}_{\phi l} & f_{\phi e} &= \bar{C}_{\phi e} - 2\bar{C}^{(1)}_{\phi l} \\
\frac{m_t}{v}f_t &= C_{u\phi,33} - 2(Y_u)_{33}\bar{C}^{(3)}_{\phi l} & f_{\phi u} &= \bar{C}_{\phi u} + \frac{4}{3}\bar{C}^{(1)}_{\phi l} \\
\frac{m_b}{v}f_b &= C_{d\phi,33} - 2(Y_d)_{33}\bar{C}^{(3)}_{\phi l} & f_{\phi d} &= \bar{C}_{\phi d} - \frac{2}{3}\bar{C}^{(1)}_{\phi l} \\
f^{(1)}_{\phi Q} &= \bar{C}^{(1)}_{\phi q} + \frac{1}{3}\bar{C}^{(1)}_{\phi l} & f^{(3)}_{\phi Q} &= 4\left[\bar{C}^{(3)}_{\phi q} - \bar{C}^{(3)}_{\phi l}\right] \\
f_{LLLL} &= \bar{C}'_{ll} \,.
\end{aligned} \tag{B.9}$$

In addition, the Higgs quartic coupling gets redefined as

$$\lambda_{\text{HISZ}} = \lambda_{\text{Warsaw}} + \frac{4m_h^2}{\Lambda^2}\,\bar{C}^{(3)}_{\phi l} \,. \tag{B.10}$$

This translates into corrections to the cubic and quartic Higgs self-couplings, which do not contribute to any of the observables in our fit.

Matching expressions at tree-level

Matching the heavy vector triplet model defined in Sec. 6.2.3 at tree level onto the Warsaw basis, we obtain

$$\begin{aligned}
C_{\phi^2} &= -\frac{3}{8}\frac{(\tilde{g}_H + g_2\tilde{g}_M)^2}{\tilde{m}_V^2} \\
C^{(3)}_{\phi l,ij} = \bar{C}^{(3)}_{\phi l}\delta_{ij} &= -\frac{1}{4}\frac{(\tilde{g}_l + g_2\tilde{g}_M)(\tilde{g}_H + g_2\tilde{g}_M)}{\tilde{m}_V^2}\delta_{ij} \\
C^{(3)}_{\phi Q,ij} = \bar{C}^{(3)}_{\phi q}\delta_{ij} &= -\frac{1}{4}\frac{(\tilde{g}_q + g_2\tilde{g}_M)(\tilde{g}_H + g_2\tilde{g}_M)}{\tilde{m}_V^2}\delta_{ij} \\
C_{ll,ijkl} = \bar{C}_{ll}\delta_{ij}\delta_{kl} + \bar{C}'_{ll}\delta_{il}\delta_{kj} &= \frac{1}{8}\frac{(\tilde{g}_l + g_2\tilde{g}_M)^2}{\tilde{m}_V^2}\left(\delta_{ij}\delta_{kl} - 2\delta_{il}\delta_{kj}\right) \\
C_{f\phi,ij} &= -\frac{(Y_f)_{ij}}{4}\frac{(\tilde{g}_H + g_2\tilde{g}_M)^2}{\tilde{m}_V^2} \qquad (f = e, u, d).
\end{aligned} \tag{B.11}$$

These results were also derived e.g. in Refs. [185, 195, 198, 335]. The full expressions for the 1-loop matching are derived here for the first time and are provided at Ref. [250].

B.2 Collection of analytic expressions for marginal likelihood

In Chap. 7, we introduced the Bayesian SFitter framework. To speed up the computations for the marginal likelihood, we derived several analytic expressions which are listed in the following subsections.

Poisson-Gamma model

The Poisson-Gamma model arises if we model the measurement in the signal region as a Poisson distribution and assume a Poisson-like prior for the number of expected background events b:

$$p(b) \sim \text{Pois}(b_{\text{CR}}|b\,k)\,, \tag{B.12}$$

where b_{CR} is the measured value in the CR and k is an interpolation factor between CR and signal region.

If we introduce the notation:

$$b_{\text{CR}} = b_{\text{SR}}\,k\,, \tag{B.13}$$

the marginal likelihood is given by:

$$\begin{aligned}
P_{pp}(d|s, b_{\text{SR}}, k) &= \int d(bk)\,\text{Pois}(d|s+b)\,\text{Pois}(b_{\text{SR}}k|bk) \\
&= \int db\,\frac{k}{\Gamma(d+1)\Gamma(b_{\text{SR}}k+1)}e^{-(s+b)}(s+b)^d e^{-kb}(bk)^{b_{\text{SR}}k}
\end{aligned} \tag{B.14}$$

for which an analytic expression can be derived by using the binomial theorem and identifying one of the integral representation of the gamma function. The final result can

be expressed as a weighted sum of Poisson distributions:

$$P_{pp}(d|s, b_{\mathrm{SR}}, k) = \sum_{i=0}^{d} C(i, k, b_{\mathrm{SR}}) \,\mathrm{Pois}(d-i|s)$$
$$\text{with } C(i, k, b_{\mathrm{SR}}) = \frac{\Gamma(i + k b_{\mathrm{SR}} + 1)}{\Gamma(i+1)\Gamma(b_{\mathrm{SR}} k + 1)} \left(\frac{k}{1+k}\right)^{k b_{\mathrm{SR}}} \left(\frac{1}{1+k}\right)^{i+1} k \tag{B.15}$$

Flat-Flat distribution

If we have more than one flat theory uncertainty, the marginal likelihood contains a convolution of two flat distributions. We can write the probability density of a flat distributions as:

$$F(x|\mu, \sigma) = \frac{1}{2\sigma} \Theta[x - (\mu - \sigma)] \Theta[(\mu + \sigma) - x] \tag{B.16}$$

where σ is the half-width of the flat distribution and μ its central value.

The convolution of two flat distributions is then given by the integral expression:

$$\begin{aligned} p_2(x|p, \sigma_1, \sigma_2) &= \int dx_1 \, F(x_1|p, \sigma_1) \, F(x|x_1, \sigma_2) \\ &= \int_{p-\sigma_1}^{p+\sigma_1} dx_1 \, F(x - x_1|0, \sigma_2) \,. \end{aligned} \tag{B.17}$$

Performing the integration for the case of $\sigma_1 > \sigma_2$, leads to the expression:

$$p_2(x) = \frac{1}{4\sigma_1\sigma_2} \begin{cases} x - p + \sigma_1 + \sigma_2 & \text{for } p - \sigma_1 - \sigma_2 < x < p - \sigma_1 + \sigma_2 \\ 2\,\sigma_2 & \text{for } p - \sigma_1 + \sigma_2 < x < p + \sigma_1 - \sigma_2 \\ -x + p + \sigma_1 + \sigma_2 & \text{for } p + \sigma_1 - \sigma_2 < x < p + \sigma_1 + \sigma_2 \\ 0 & \text{otherwise} \end{cases} . \tag{B.18}$$

The result for $\sigma_2 > \sigma_1$ is given by the same formula but σ_1 is exchanged with σ_2.

Flat-Flat-Flat distribution

The convolution of three flat distributions is given by:

$$\begin{aligned} p_3(x) &= \int dx_1 \, dx_2 \, F(x_1|p, \sigma_1) \, F(x_2 - x_1|0, \sigma_2) \, F(x - x_2|0, \sigma_3) \\ &= \int dx_2 \, p_2(x_2) \, F(x - x_2|0, \sigma_3) \,, \end{aligned} \tag{B.19}$$

where p_2 is defined in Eq. (B.18).

We can think of the integral given in Eq. (B.19) as moving a box with the width $2\sigma_3$ over the distribution p_2. The integration corresponds to the area below p_2 inside of the box. Depending on the parameters of the flat distributions and x, the integral is given by the area of a triangle, a trapezoid, rectangle or a combination of these shapes. Let's assume $\sigma_1 > \sigma_2 > \sigma_3$. There are a two distinct cases:

$$(1) : \sigma_1 - \sigma_2 \geq \sigma_3$$

$$(2) : \sigma_1 - \sigma_2 < \sigma_3 \tag{B.20}$$

$\sigma_1 - \sigma_2$ is the length of central flat region of p_2. Depending whether this length is larger than σ_3 or smaller alters the result. The integration splits up into several regions. The boundaries of these regions are given by:

$$\begin{aligned}
b_1 &= p - \sigma_1 - \sigma_2 - \sigma_3 \\
b_2 &= p - \sigma_1 - \sigma_2 + \sigma_3 \\
b_3 &= p - \sigma_1 + \sigma_2 - \sigma_3 \\
b_4 &= p - \sigma_1 + \sigma_2 + \sigma_3 \\
b_5 &= p + \sigma_1 - \sigma_2 - \sigma_3 \\
b_6 &= p + \sigma_1 - \sigma_2 + \sigma_3 \\
b_7 &= p + \sigma_1 + \sigma_2 - \sigma_3 \\
b_8 &= p + \sigma_1 + \sigma_2 + \sigma_3 \,.
\end{aligned} \tag{B.21}$$

Performing the computation for case (1), leads to the expression:

$$p_3(x) = \begin{cases}
(x - b_1)^2 / 2 & \text{for } b_1 < x \leq b_2 \\
2\,\sigma_3\ (x - (b_1 + b_2)/2) & \text{for } b_2 < x \leq b_3 \\
2s_2\,(x - b_3) + (b_4 - x)\,((b_3 + x)/2 - (b_1 + b_2)/2) & \text{for } b_3 < x \leq b_4 \\
4\sigma_2\sigma_3 & \text{for } b_4 < x \leq b_5 \\
2s_2(b_6 - x) + (x - b_5)\,((b_7 + b_8)/2 - (x + b_6)/2) & \text{for } b_5 < x \leq b_6 \\
2\sigma_3\,((b_7 + b_8)/2 - x) & \text{for } b_6 < x \leq b_7 \\
(b_8 - x)^2 / 2 & \text{for } b_7 < x \leq b_8
\end{cases} . \tag{B.22}$$

For case (2), we can derive a similar expression. The key difference is the order of the boundaries b_4 and b_5 and the central expression:

$$p_3(x) = \begin{cases}
(x - b_1)^2 / 2 & \text{for } b_1 < x \leq b_2 \\
2\,\sigma_3\ (x - (b_1 + b_2)/2) & \text{for } b_2 < x \leq b_3 \\
2s_2\,(x - b_3) + (b_4 - x)\,((b_3 + x)/2 - (b_1 + b_2)/2) & \text{for } b_3 < x \leq b_5 \\
(b_4 - x)\,((b_3 + x)/2 - (b_1 + b_2)/2) & \\
\qquad + (x - b_5)\,((b_7 + b_8)/2 - (x + b_6)/2) & \\
\qquad + 4\sigma_2(\sigma_1 - \sigma_2) & \text{for } b_5 < x \leq b_4 \\
2s_2(b_6 - x) + (x - b_5)\,((b_7 + b_8)/2 - (x + b_6)/2) & \text{for } b_4 < x \leq b_6 \\
2\sigma_3\,((b_7 + b_8)/2 - x) & \text{for } b_6 < x \leq b_7 \\
(b_8 - x)^2 / 2 & \text{for } b_7 < x \leq b_8
\end{cases} . \tag{B.23}$$

Flat-Gauss distribution

The convolution of a flat distribution and a Gaussian is given by:

$$\begin{aligned}
GF(d|p, \sigma_{\text{Gauss},}, \sigma_{\text{Flat}}) &= \int dp\, G(d|\tilde{p}, \sigma_{\text{Gauss}})\, F(\tilde{p}|p, \sigma_{\text{Flat}}) \\
&= \frac{1}{2\sigma_{\text{Flat}}} \int_{p-\sigma_{\text{Flat}}}^{p+\sigma_{\text{Flat}}} d\tilde{p}\, G(d|\tilde{p}, \sigma_{\text{Gauss}})
\end{aligned}$$

$$= \frac{1}{4\sigma_{\text{Flat}}} \left(\text{erf}\left(\frac{p + \sigma_{\text{Flat}} - d}{\sqrt{2}\sigma_{\text{Gauss}}} \right) - \text{erf}\left(\frac{p - \sigma_{\text{Flat}} - d}{\sqrt{2}\sigma_{\text{Gauss}}} \right) \right) , \quad \text{(B.24)}$$

where σ_{Gauss} is the width of the Gaussian distribution, σ_{Flat} is the half-width of the flat distribution, p is the central value of the resulting distribution and d is its argument.

Flat-Flat-Gauss distribution

The convolution of two flat distributions and one Gaussian distribution is given by:

$$\text{GFF}(d|p, \sigma_{\text{Flat},1}, \sigma_{\text{Flat},2}, \sigma_{\text{Gauss}}) = \int d\tilde{p}\, G(d|\tilde{p}, \sigma_{\text{Gauss}})\, p_2(\tilde{p}|p, \sigma_{\text{Flat},1}, \sigma_{\text{Flat},2}) , \quad \text{(B.25)}$$

where p_2 is defined in Eq. (B.18). Because p_2 contains several distinct regions, it makes sense to define the following boundaries:

$$\begin{aligned} b_1 &= p - \sigma_{\text{Flat},1} - \sigma_{\text{Flat},2} \\ b_2 &= p - \sigma_{\text{Flat},1} + \sigma_{\text{Flat},2} \\ b_3 &= p + \sigma_{\text{Flat},1} - \sigma_{\text{Flat},2} \\ b_4 &= p + \sigma_{\text{Flat},1} + \sigma_{\text{Flat},2} , \end{aligned} \quad \text{(B.26)}$$

where we assumed $\sigma_{\text{Flat},1} > \sigma_{\text{Flat},2}$. The final expression becomes:

$$\begin{aligned} \text{GFF}(d|p, \sigma_{\text{Flat},1}, \sigma_{\text{Flat},2}, \sigma_{\text{Gauss}}) = \frac{1}{4\sigma_{\text{Flat},1}\sigma_{\text{Flat},2}} \Big(& F_2(b_1, b_2) - b_1\, F_1(b_1, b_2) \\ & + 2\sigma_{\text{Flat},2}\, F_1(b_2, b_3) \; + b_4\, F_1(b_3, b_4) - F_2(b_3, b_4) \Big) , \end{aligned} \quad \text{(B.27)}$$

where F_1 and F_2 are defined as:

$$\begin{aligned} F_1(a, b; \mu, \sigma) &= \int_a^b dx\, G(x|\mu, \sigma) \\ &= \frac{1}{2} \left(\text{erf}\left(\frac{b - \mu}{\sqrt{2}\sigma} \right) - \text{erf}\left(\frac{a - \mu}{\sqrt{2}\sigma} \right) \right) , \end{aligned} \quad \text{(B.28)}$$

and

$$\begin{aligned} F_2(a, b; \mu, \sigma) &= \int_a^b dx\, x\, G(x|\mu, \sigma) \\ &= \sigma^2 \, [G(a|\mu, \sigma) - G(b|\mu, \sigma)] + \mu\, F_1(a, b; \mu, \sigma) . \end{aligned} \quad \text{(B.29)}$$

In Eq. (B.27), we suppressed the dependency on μ and σ.

B.3 Numerical results

Table B.2 reports the numerical values of the boundaries of the 68% and 95% CL intervals shown in Fig. 7.27. It was taken from the appendix of Ref. [5].

	Marginalised		Profiled	
Coefficient	68% CL	95% CL	68% CL	95% CL
f_{GG}	[1.19, 6.65]	[-1.75, 8.33]	[3.79, 8.28]	[-1.09, 9.50]
f_B	[-5.81, -2.47]	[-7.66, 8.94]	[-6.49, 5.79]	[-8.69, 10.08]
	[0.23, 7.96]			
$f_{\phi 2}$	[-0.96, 2.98]	[-3.11, 4.77]	[-2.07, 3.68]	[-4.59, 6.55]
f_μ	[-3.79, 8.43]	[-7.45, 20.66]	[-3.79, 9.66]	[-8.68, 21.88]
f_t	[-0.51, 3.51]	[-2.19, 5.86]	[-0.80, 3.68]	[-3.56, 5.75]
f_b	[-0.62, 2.33]	[-2.23, 3.68]	[-0.60, 3.44]	[-3.03, 5.33]
f_τ	[-0.38, 1.77]	[-1.46, 2.84]	[-1.88, 1.00]	[-3.66, 2.55]
f_-	[-5.65, -0.49]	[-7.03, 2.95]	[-5.34, 2.28]	[-7.75, 5.09]
$f_+ \times 10$	[1.07, 8.93]	[-3.21, 12.50]	[0.36, 8.93]	[-3.93, 13.93]
$f_{tG} \times 10$	[1.72, 6.28]	[-0.85, 7.14]	[3.53, 7.12]	[-0.36, 8.02]
$f_W \times 10$	[2.64, 16.28]	[-4.19, 23.11]	[1.25, 21.80]	[-8.35, 30.03]
$f_{BW} \times 10$	[0.82, 8.34]	[-3.47, 12.10]	[1.57, 9.82]	[-2.83, 14.22]
$f_{3W} \times 10$	[-3.74, 3.41]	[-7.31, 6.48]	[-5.31, 6.89]	[-10.19, 9.94]
$f_{\phi Q}^{(3)} \times 100$	[-9.56, 13.13]	[-17.12, 25.24]	[-1.99, 23.60]	[-13.19, 34.80]
$f_{\phi u}^{(1)} \times 100$	[-13.5, -7.6]	[-17.23, -2.20]	[-15.70, -0.90]	[-22.00, 15.80]
	[4.6, 13.60]	[1.10 16.00]	[2.70, 10.20]	
$f_{\phi d}^{(1)} \times 100$	[-7.23, 2.17]	[-11.57, 7.24]	[-8.64, 6.07]	[-14.69, 12.13]
$f_{\phi 1} \times 100$	[-0.38, 7.32]	[-4.77, 11.17]	[-0.37, 7.88]	[-4.22, 12.83]
$f_{4L} \times 100$	[-2.70, -0.46]	[-3.82, 0.82]	[-2.86, -0.46]	[-4.14, 0.82]
$f_{\phi Q}^{(1)} \times 100$	[1.19, 7.19]	[-2.40, 8.78]	[-1.20, 5.60]	[-4.00, 8.40]
$f_{\phi e}^{(1)} \times 100$	[-4.58, -0.98]	[-6.50, 0.70]	[-5.06, -1.46]	[-6.98, 0.67]
BR_{inv}	[0, 1.12]	[0, 1.76]	[0, 2.40]	[0, 3.36]

Table B.2: Numerical values for the results shown in Fig. 7.27.

Acknowledgments

First of all, I am very grateful to my supervisor Tilman Plehn for supporting me on my journey for the last four years, giving me the opportunity to work on interesting projects, teaching me about physics and encouraging me to explore the field of deep learning. I would also like to thank Monica Dunford for refereeing my thesis and Joerg Jaeckel and Björn Malte Schäfer for being additional members of my defense committee.

Furthermore, I gratefully acknowledge the Heidelberg University and the research training group "Particle physics beyond the Standard Model" (DFG GRK 1940) for funding me and making all of this possible.

Moreover, I would like to thank all of my collaborators, Gregor Kasieczka, Florian Otterpohl, Manuel Haußmann, Ilaria Brivio, Sebastian Bruggisser, Emma Geoffray, Wolfgang Kilian, Michael Krämer, Benjamin Summ, Simon Badger, Anja Butter, Sebastian Pitz and Nina Elmer. None of the research presented in this thesis would have been possible without them.

Special thanks goes to my proof readers Lennert Thormählen, Ramon Winterhalder, Marco Bellagente and Emma Geoffray for doing their best to make this thesis at least readable. I would especially like to thank the LaTeX master mind, Lorenz Vogel, who transformed this pile of garbage into a nice looking pile of garbage. I hope that he will continue using his magic to also help other PhD students in need.

I would like to thank all of the current and former members of the pheno research group for great discussions and nice coffee breaks. In particular, I would like to thank my PhD sisters and brothers: Theo Heimel, the god master of coding, for sharing his seemingly infinite amount of knowledge with me, Emma Geoffray for going with me through years of SFittering (the "S" stands for pain), Ramon Winterhalder for being my machine learning and drinking mentor, Lennert Thormählen, my PhD cousin, for still sticking with us even in times of post-Jexit, Peter Reimitz, the director of the IPR, for making all of the afternoon coffee breaks very enjoyable, Luigi Favaro for making me the best home made Carbonara, Lennart Röver for being a creative mastermind, our Sharkov chains will revolutionise the field of statistics and marine biology, Marco Bellagente for teaching me how to endure frustration and pain, Anastasia Filimonova for making every coffee break interesting by bringing the weirdest and most exotic sweets, and Bob, the most important and beloved member of the pheno research group, for not murdering me.

Special thanks also goes to all the former and current post docs: Tanmoy Modak for inviting me so many times for dinner, Barry Dillon for great discussions about machine learning, Ilaria Brivio for teaching me everything I know about SMEFT (which is not a lot), and Anja Butter for supporting us from a distance.

Furthermore, I would like to thank Nina Elmer for supporting me in my stress ball throwing career, Benedikt Schosser for teaching me that there is more to coffee than

simply being a drug, Jörg Holsten for many interesting conversations about basically everything, and Natalie Soybelman for inventing the darts and beer night tradition.

Finally, I would like to thank my family and friends for their support through all these years. None of this would have been possible without them.

Bibliography

[1] G. Kasieczka, M. Luchmann, F. Otterpohl and T. Plehn, *Per-Object Systematics using Deep-Learned Calibration*, SciPost Phys. **9**, 089 (2020), doi:10.21468/SciPostPhys.9.6.089, arXiv:2003.11099.

[2] M. Bellagente, M. Haußmann, M. Luchmann and T. Plehn, *Understanding Event-Generation Networks via Uncertainties*, SciPost Physics **13** (2022), doi:10.21468/scipostphys.13.1.003, arXiv:2104.04543.

[3] I. Brivio, S. Bruggisser, E. Geoffray, W. Killian, M. Krämer, M. Luchmann, T. Plehn and B. Summ, *From models to SMEFT and back?*, SciPost Phys. **12**, 036 (2022), doi:10.21468/SciPostPhys.12.1.036, arXiv:2108.01094.

[4] S. Badger, A. Butter, M. Luchmann, S. Pitz and T. Plehn, *Loop Amplitudes from Precision Networks* (2022), arXiv:2206.14831.

[5] I. Brivio, S. Bruggisser, N. Elmer, E. Geoffray, M. Luchmann and T. Plehn, *To Profile or To Marginalize – A SMEFT Case Study* (2022), arXiv:2208.08454.

[6] F. Englert and R. Brout, *Broken symmetry and the mass of gauge vector mesons*, Phys. Rev. Lett. **13**, 321 (1964), doi:10.1103/PhysRevLett.13.321, `https://link.aps.org/doi/10.1103/PhysRevLett.13.321`.

[7] P. W. Higgs, *Broken symmetries, massless particles and gauge fields*, Physics Letters **12**, 132 (1964), doi:https://doi.org/10.1016/0031-9163(64)91136-9, `https://www.sciencedirect.com/science/article/pii/0031916364911369`.

[8] P. W. Higgs, *Broken symmetries and the masses of gauge bosons*, Phys. Rev. Lett. **13**, 508 (1964), doi:10.1103/PhysRevLett.13.508, `https://link.aps.org/doi/10.1103/PhysRevLett.13.508`.

[9] G. Aad *et al.*, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. B **716**, 1 (2012), doi:10.1016/j.physletb.2012.08.020, arXiv:1207.7214.

[10] S. Chatrchyan *et al.*, *Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC*, Phys. Lett. B **716**, 30 (2012), doi:10.1016/j.physletb.2012.08.021, arXiv:1207.7235.

[11] G. Aad *et al.*, *Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments*, Phys. Rev. Lett. **114**, 191803 (2015), doi:10.1103/PhysRevLett.114.191803, `https://link.aps.org/doi/10.1103/PhysRevLett.114.191803`.

[12] CMS collaboration, *CERN Document Server - CMS Papers*, `http://cdsweb.cern.ch/collection/CMS%20Papers?ln=en` (2022).

[13] ATLAS collaboration, *ATLAS Experiment - Public Results*, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/Publications (2022).

[14] LHCb collaboration, *The LHCb Publix results - LHCb publications*, https://lhcbproject.web.cern.ch/Publications/LHCbProjectPublic/Summary_all.html (2022).

[15] M. Bauer and T. Plehn, *Yet Another Introduction to Dark Matter* (2017).

[16] D. Bödeker and W. Buchmüller, *Baryogenesis from the weak scale to the grand unification scale*, Reviews of Modern Physics **93** (2021), doi:10.1103/revmodphys.93.035004, https://doi.org/10.1103%2Frevmodphys.93.035004.

[17] A. Hook, *TASI Lectures on the Strong CP Problem and Axions*, doi:10.48550/ARXIV.1812.02669 (2018).

[18] C. Rovelli, *Quantum Gravity*, Cambridge Monographs on Mathematical Physics. Cambridge University Press, doi:10.1017/CBO9780511755804 (2004).

[19] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, MIT Press, http://www.deeplearningbook.org (2016).

[20] J. Gallicchio and M. D. Schwartz, *Quark and Gluon Jet Substructure*, JHEP **1304**, 090 (2013), doi:10.1007/JHEP04(2013)090, arXiv:1211.7038.

[21] P. T. Komiske, E. M. Metodiev and M. D. Schwartz, *Deep learning in color: towards automated quark/gluon jet discrimination*, JHEP **1701**, 110 (2017), doi:10.1007/JHEP01(2017)110, arXiv:1612.01551.

[22] T. Cheng, *Recursive Neural Networks in Quark/Gluon Tagging*, Comput. Softw. Big Sci. **2** (2018), doi:10.1007/s41781-018-0007-y, arXiv:1711.02633.

[23] P. T. Komiske, E. M. Metodiev and J. Thaler, *Energy Flow Networks: Deep Sets for Particle Jets*, JHEP **01**, 121 (2019), doi:10.1007/JHEP01(2019)121, arXiv:1810.05165.

[24] S. Bright-Thonney and B. Nachman, *Investigating the Topology Dependence of Quark and Gluon Jets*, JHEP **03**, 098 (2019), doi:10.1007/JHEP03(2019)098, arXiv:1810.05653.

[25] G. Kasieczka, N. Kiefer, T. Plehn and J. M. Thompson, *Quark-Gluon Tagging: Machine Learning vs Detector*, SciPost Phys. **6**, 069 (2019), doi:10.21468/SciPostPhys.6.6.069, arXiv:1812.09223.

[26] S. H. Lim and M. M. Nojiri, *Spectral Analysis of Jet Substructure with Neural Networks: Boosted Higgs Case*, JHEP **10**, 181 (2018), doi:10.1007/JHEP10(2018)181, arXiv:1807.03312.

[27] J. Lin, M. Freytsis, I. Moult and B. Nachman, *Boosting $H \to b\bar{b}$ with Machine Learning*, JHEP **10**, 101 (2018), doi:10.1007/JHEP10(2018)101, arXiv:1807.10768.

[28] A. Butter *et al.*, *The Machine Learning landscape of top taggers*, SciPost Phys. **7**, 014 (2019), doi:10.21468/SciPostPhys.7.1.014, arXiv:1902.09914.

[29] M. Feickert and B. Nachman, *A living review of machine learning for particle physics*, doi:10.48550/ARXIV.2102.02770, https://iml-wg.github.io/HEPML-LivingReview/ (2021), arXiv:2102.02770.

[30] T. Heimel, G. Kasieczka, T. Plehn and J. M. Thompson, *QCD or What?*, SciPost Phys. **6**, 030 (2019), doi:10.21468/SciPostPhys.6.3.030, arXiv:1808.08979.

[31] M. Farina, Y. Nakai and D. Shih, *Searching for New Physics with Deep Autoencoders* (2018), arXiv:1808.08992.

[32] A. Andreassen, P. T. Komiske, E. M. Metodiev, B. Nachman and J. Thaler, *OmniFold: A Method to Simultaneously Unfold All Observables*, Phys. Rev. Lett. **124**, 182001 (2020), doi:10.1103/PhysRevLett.124.182001, arXiv:1911.09107.

[33] M. Bellagente, A. Butter, G. Kasieczka, T. Plehn and R. Winterhalder, *How to GAN away Detector Effects*, SciPost Phys. **8**, 070 (2020), doi:10.21468/SciPostPhys.8.4.070, arXiv:1912.00477.

[34] A. Butter and T. Plehn, *Generative Networks for LHC events* (2020), arXiv:2008.08558.

[35] P. C. Bhat, *Multivariate analysis methods in particle physics*, Annual Review of Nuclear and Particle Science **61**, 281 (2011), doi:10.1146/annurev.nucl.012809.104427, `https://doi.org/10.1146/annurev.nucl.012809.104427`.

[36] R. M. Neal, *Bayesian learning for neural networks*, vol. 118, Springer Science & Business Media, doi:10.1007/978-1-4612-0745-0 (1995).

[37] D. MacKay, *Probable Networks and Plausible Predictions – A Review of Practical Bayesian Methods for Supervised Neural Networks*, Comp. in Neural Systems **6**, 4679 (1995), `http://www.inference.org.uk/mackay/network.pdf`.

[38] C. Blundell, J. Cornebise, K. Kavukcuoglu and D. Wierstra, *Weight uncertainty in neural networks* (2015), `http://arxiv.org/abs/1505.05424`, Cite arxiv:1505.05424Comment: In Proceedings of the 32nd International Conference on Machine Learning (ICML 2015).

[39] Y. Gal, *Uncertainty in Deep Learning*, Ph.D. thesis, Cambridge (2016).

[40] B. Allanach, *Beyond the standard model*, CERN Yellow Reports: School Proceedings pp. Vol 5 (2017): Proceedings of the 2016 European School of High–Energy Physics (2017), doi:10.23730/CYRSP-2017-005.123, `https://e-publishing.cern.ch/index.php/CYRSP/article/view/353`.

[41] M. McCullough, *Lectures on Physics Beyond the Standard Model.*, In *6th Tri-Institute Summer School on Elementary Particles* (2018).

[42] S. Weinberg, *Phenomenological Lagrangians*, Physica A **96**, 327 (1979), doi:10.1016/0378-4371(79)90223-1.

[43] A. Krizhevsky, I. Sutskever and G. E. Hinton, *Imagenet classification with deep convolutional neural networks* (2012), doi:10.1145/3065386, This victory in the ImageNet competition is often seen as the beginning of the Deep Learning revolution.

[44] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, MIT Press, `http://www.deeplearningbook.org` (2016).

[45] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher and D. J. Schwab, *A high-bias, low-variance introduction to Machine Learning for physicists*, Physics Reports **810**, 1 (2019), doi:https://doi.org/10.1016/j.physrep.2019.03.001, `https://www.sciencedirect.com/science/article/pii/S0370157319300766`.

[46] X. Glorot, A. Bordes and Y. Bengio, *Deep sparse rectifier neural networks*, In G. Gordon, D. Dunson and M. Dudík, eds., *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, vol. 15 of *Proceedings of Machine Learning Research*, pp. 315–323. PMLR, Fort Lauderdale, FL, USA (2011).

[47] G. Kasieczka, T. Plehn, M. Russell and T. Schell, *Deep-learning Top Taggers or The End of QCD?*, JHEP **05**, 006 (2017), doi:10.1007/JHEP05(2017)006, arXiv:1701.08784.

[48] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li and M. Sun, *Graph neural networks: A review of methods and applications*, AI Open **1**, 57 (2020), doi:https://doi.org/10.1016/j.aiopen.2021.01.001, `https://www.sciencedirect.com/science/article/pii/S2666651021000012`.

[49] J. Shlomi, P. Battaglia and J.-R. Vlimant, *Graph neural networks in particle physics*, Machine Learning: Science and Technology **2**, 021001 (2021), doi:10.1088/2632-2153/abbf9a, `https://doi.org/10.1088/2632-2153/abbf9a`.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, *Attention is all you need*, In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 6000–6010. Curran Associates Inc., Red Hook, NY, USA, ISBN 9781510860964 (2017).

[51] H. Qu, C. Li and S. Qian, *Particle transformer for jet tagging*, doi:10.48550/ARXIV.2202.03772, Example of transformers used in particle physics (2022).

[52] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Mathematics of Control, Signals and Systems (1989), doi:10.1007/BF02551274.

[53] S. Ruder, *An overview of gradient descent optimization algorithms*, arXiv preprint arXiv:1609.04747 (2016).

[54] S. Kullback and R. A. Leibler, *On Information and Sufficiency*, The Annals of Mathematical Statistics **22**, 79 (1951), doi:10.1214/aoms/1177729694, `https://doi.org/10.1214/aoms/1177729694`.

[55] D. M. Blei, A. Kucukelbir and J. D. McAuliffe, *Variational inference: A review for statisticians*, Journal of the American statistical Association **112**, 859 (2017).

[56] A. Y. Ng, *Feature selection, l1 vs. l2 regularization, and rotational invariance*, In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, p. 78. Association for Computing Machinery, New York, NY, USA, ISBN 1581138385, doi:10.1145/1015330.1015435 (2004).

[57] S. Farquhar, L. Smith and Y. Gal, *Liberty or depth: Deep bayesian neural nets do not need complex weight posterior approximations*, In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Curran Associates Inc., Red Hook, NY, USA, ISBN 9781713829546 (2020).

[58] M. Patacchiola, *Evidence, KL-divergence, and ELBO*, `https://mpatacchiola.github.io/blog/2021/01/25/intro-variational-inference.html` (2021).

[59] A. Kendall and Y. Gal, *What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?*, Proc. NIPS (2017), arXiv:1703.04977.

[60] S. Bollweg, M. Haußmann, G. Kasieczka, M. Luchmann, T. Plehn and J. Thompson, *Deep-Learning Jets with Uncertainties and More*, SciPost Phys. **8**, 006 (2020), doi:10.21468/SciPostPhys.8.1.006, arXiv:1904.10004.

[61] Y. Kwon, J.-H. Won, B. J. Kim and M. C. Paik, *Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation*, Computational Statistics and Data Analysis **142**, 106816 (2020), doi:https://doi.org/10.1016/j.csda.2019.106816, `https://www.sciencedirect.com/science/article/pii/S016794731930163X`.

[62] D. P. Kingma and M. Welling, *Auto-encoding variational bayes* (2014), arXiv:1312.6114.

[63] D. P. Kingma, T. Salimans and M. Welling, *Variational dropout and the local reparameterization trick*, In *Advances in neural information processing systems*, pp. 2575–2583 (2015).

[64] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf *et al.*, *Pytorch: An imperative style, high-performance deep learning library*, In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox and R. Garnett, eds., *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc. (2019), arXiv:1912.01703.

[65] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga and A. Lerer, *Automatic differentiation in pytorch*, In *NIPS 2017 Workshop on Autodiff* (2017).

[66] D. Rezende and S. Mohamed, *Variational inference with normalizing flows*, In F. Bach and D. Blei, eds., *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538. PMLR, Lille, France (2015), arXiv:1505.05770.

[67] L. Dinh, J. Sohl-Dickstein and S. Bengio, *Density estimation using real nvp* (2016), arXiv:1605.08803.

[68] D. P. Kingma and P. Dhariwal, *Glow: Generative flow with invertible 1x1 convolutions* (2018), arXiv:1807.03039.

[69] I. Kobyzev, S. Prince and M. Brubaker, *Normalizing flows: An introduction and review of current methods*, IEEE Transactions on Pattern Analysis and Machine Intelligence p. 1–1 (2020), doi:10.1109/tpami.2020.2992934, `http://dx.doi.org/10.1109/TPAMI.2020.2992934`.

[70] D. P. Kingma and M. Welling, *An introduction to variational autoencoders*, Foundations and Trends® in Machine Learning **12**, 307–392 (2019), doi:10.1561/2200000056, `http://dx.doi.org/10.1561/2200000056`.

[71] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, *Generative adversarial networks* (2014), arXiv:1406.2661.

[72] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta and A. A. Bharath, *Generative adversarial networks: An overview*, IEEE Signal Processing Magazine **35**, 53–65 (2018), doi:10.1109/msp.2017.2765202, `http://dx.doi.org/10.1109/MSP.2017.2765202`.

[73] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan and S. Ganguli, *Deep unsupervised learning using nonequilibrium thermodynamics*, doi:10.48550/ARXIV.1503.03585 (2015).

[74] J. Ho, A. Jain and P. Abbeel, *Denoising diffusion probabilistic models*, doi:10.48550/ARXIV.2006.11239 (2020).

[75] L. Deng, *The mnist database of handwritten digit images for machine learning research*, IEEE Signal Processing Magazine **29**, 141 (2012).

[76] L. Lonnblad, C. Peterson and T. Rognvaldsson, *Using neural networks to identify jets*, Nucl. Phys. **B349**, 675 (1991), doi:10.1016/0550-3213(91)90392-B.

[77] J. Cogan, M. Kagan, E. Strauss and A. Schwartzman, *Jet-Images: Computer Vision Inspired Techniques for Jet Tagging*, JHEP **1502**, 118 (2015), doi:10.1007/JHEP02(2015)118, arXiv:1407.5675.

[78] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman and A. Schwartzman, *Jet-images — deep learning edition*, JHEP **1607**, 069 (2016), doi:10.1007/JHEP07(2016)069, arXiv:1511.05190.

[79] H. Qu and L. Gouskos, *ParticleNet: Jet Tagging via Particle Clouds* (2019), arXiv:1902.08570.

[80] CMS Collaboration, *Performance of the DeepJet b tagging algorithm using 41.9/fb of data from proton-proton collisions at 13TeV with Phase 1 CMS detector*, CMS-DP-2018-058 (2018), `http://cds.cern.ch/record/2646773`.

[81] P. Baldi, K. Bauer, C. Eng, P. Sadowski and D. Whiteson, *Jet substructure classification in high-energy physics with deep neural networks*, Phys. Rev. D **93**, 094034 (2016), doi:10.1103/PhysRevD.93.094034, arXiv:1603.09349.

[82] G. Louppe, K. Cho, C. Becot and K. Cranmer, *QCD-Aware Recursive Neural Networks for Jet Physics*, JHEP **1901**, 057 (2019), doi:10.1007/JHEP01(2019)057, arXiv:1702.00748.

[83] CMS Collaboration, *Machine learning-based identification of highly Lorentz-boosted hadronically decaying particles at the CMS experiment*, CMS-PAS-JME-18-002 (2019), `https://cds.cern.ch/record/2683870`.

[84] ATLAS Collaboration, *Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC*, Eur. Phys. J. **C79**, 375 (2019), doi:10.1140/epjc/s10052-019-6847-8, arXiv:1808.07858.

[85] S. Macaluso and D. Shih, *Pulling Out All the Tops with Computer Vision and Deep Learning*, JHEP **1810**, 121 (2018), doi:10.1007/JHEP10(2018)121, arXiv:1803.00107.

[86] A. Butter, G. Kasieczka, T. Plehn and M. Russell, *Deep-learned Top Tagging with a Lorentz Layer*, SciPost Phys. **5**, 028 (2018), doi:10.21468/SciPostPhys.5.3.028, arXiv:1707.08966.

[87] L. G. Almeida, M. Backovic, M. Cliche, S. J. Lee and M. Perelstein, *Playing Tag with ANN: Boosted Top Identification with Pattern Recognition*, JHEP **1507**, 086 (2015), doi:10.1007/JHEP07(2015)086, arXiv:1501.05968.

[88] J. Pearkes, W. Fedorko, A. Lister and C. Gay, *Jet Constituents for Deep Neural Network Based Top Quark Tagging* (2017), arXiv:1704.02124.

[89] S. Egan, W. Fedorko, A. Lister, J. Pearkes and C. Gay, *Long Short-Term Memory (LSTM) networks with jet constituents for boosted top tagging at the LHC* (2017), arXiv:1711.09059.

[90] S. Choi, S. J. Lee and M. Perelstein, *Infrared Safety of a Neural-Net Top Tagging Algorithm*, JHEP **02**, 132 (2019), doi:10.1007/JHEP02(2019)132, arXiv:1806.01263.

[91] L. Moore, K. Nordström, S. Varma and M. Fairbairn, *Reports of My Demise Are Greatly Exaggerated: N-subjettiness Taggers Take On Jet Images*, SciPost Phys. **7**, 036 (2019), doi:10.21468/SciPostPhys.7.3.036, arXiv:1807.04769.

[92] A. J. Larkoski, I. Moult and B. Nachman, *Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning*, Phys. Rept. **841**, 1 (2020), doi:10.1016/j.physrep.2019.11.001, For a review see, arXiv:1709.04464.

[93] B. Nachman, *A guide for deploying Deep Learning in LHC searches: How to achieve optimality and account for uncertainty* (2019), arXiv:1909.03081.

[94] C. Englert, P. Galler, P. Harris and M. Spannowsky, *Machine Learning Uncertainties with Adversarial Neural Networks*, Eur. Phys. J. **C79** (2019), doi:10.1140/epjc/s10052-018-6511-8, arXiv:1807.08763.

[95] J. Barnard, E. N. Dawe, M. J. Dolan and N. Rajcic, *Parton Shower Uncertainties in Jet Substructure Analyses with Deep Neural Networks*, Phys. Rev. D **95**, 014018 (2017), doi:10.1103/PhysRevD.95.014018, arXiv:1609.00607.

[96] S. Diefenbacher, H. Frost, G. Kasieczka, T. Plehn and J. M. Thompson, *CapsNets Continuing the Convolutional Quest*, SciPost Phys. **8**, 023 (2020), doi:10.21468/SciPostPhys.8.2.023, arXiv:1906.11265.

[97] K. Datta and A. Larkoski, *How Much Information is in a Jet?*, JHEP **1706**, 073 (2017), doi:10.1007/JHEP06(2017)073, arXiv:1704.08249.

[98] S. Chang, T. Cohen and B. Ostdiek, *What is the Machine Learning?*, Phys. Rev. D **97**, 056009 (2018), doi:10.1103/PhysRevD.97.056009, arXiv:1709.10106.

[99] K. Datta and A. J. Larkoski, *Novel Jet Observables from Machine Learning*, JHEP **1803**, 086 (2018), doi:10.1007/JHEP03(2018)086, arXiv:1710.01305.

[100] P. T. Komiske, E. M. Metodiev and J. Thaler, *Energy flow polynomials: A complete linear basis for jet substructure*, JHEP **1804**, 013 (2018), doi:10.1007/JHEP04(2018)013, arXiv:1712.07124.

[101] T. Roxlo and M. Reece, *Opening the black box of neural nets: case studies in stop/top discrimination* arXiv:1804.09278.

[102] L. M. Dery, B. Nachman, F. Rubbo and A. Schwartzman, *Weakly Supervised Classification in High Energy Physics*, JHEP **1705**, 145 (2017), doi:10.1007/JHEP05(2017)145, arXiv:1702.00414.

[103] T. Cohen, M. Freytsis and B. Ostdiek, *(Machine) Learning to Do More with Less*, JHEP **1802**, 034 (2018), doi:10.1007/JHEP02(2018)034, arXiv:1706.09451.

[104] E. M. Metodiev, B. Nachman and J. Thaler, *Classification without labels: Learning from mixed samples in high energy physics*, JHEP **1710**, 174 (2017), doi:10.1007/JHEP10(2017)174, arXiv:1708.02949.

[105] P. T. Komiske, E. M. Metodiev, B. Nachman and M. D. Schwartz, *Learning to classify from impure samples with high-dimensional data*, Phys. Rev. D **98**, 011502 (2018), doi:10.1103/PhysRevD.98.011502, arXiv:1801.10158.

[106] A. Andreassen, I. Feige, C. Frye and M. D. Schwartz, *JUNIPR: a Framework for Unsupervised Machine Learning in Particle Physics*, Eur. Phys. J. **C79**, 102 (2019), doi:10.1140/epjc/s10052-019-6607-9, arXiv:1804.09720.

[107] O. Cerri, T. Q. Nguyen, M. Pierini, M. Spiropulu and J.-R. Vlimant, *Variational Autoencoders for New Physics Mining at the Large Hadron Collider*, JHEP **05**, 036 (2019), doi:10.1007/JHEP05(2019)036, arXiv:1811.10276.

[108] S. Badger and J. Bullock, *Using neural networks for efficient evaluation of high multiplicity scattering amplitudes* (2020), arXiv:2002.07516.

[109] T. Plehn, M. Spannowsky, M. Takeuchi and D. Zerwas, *Stop Reconstruction with Tagged Tops*, JHEP **10**, 078 (2010), doi:10.1007/JHEP10(2010)078, arXiv:1006.2833.

[110] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen and P. Z. Skands, *An introduction to PYTHIA 8.2*, Comput. Phys. Commun. **191**, 159 (2015), doi:10.1016/j.cpc.2015.01.024, arXiv:1410.3012.

[111] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens and M. Selvaggi, *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, JHEP **02**, 057 (2014), doi:10.1007/JHEP02(2014)057, arXiv:1307.6346.

[112] M. Abadi and et al., *TensorFlow: A System for Large-Scale Machine Learning*, OSDI **16**, 265 (2016), `https://www.tensorflow.org/probability`.

[113] Y. Wen, P. Vicol, J. Ba, D. Tran and R. Grosse, *Method for calculating gradients*, Proc. NIPS (2018), arXiv:1803.04386.

[114] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, In Y. Bengio and Y. LeCun, eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015).

[115] A. Butter, S. Diefenbacher, G. Kasieczka, B. Nachman and T. Plehn, *GANplifying event samples*, SciPost Phys. **10**, 139 (2021), doi:10.21468/SciPostPhys.10.6.139, arXiv:2008.06545.

[116] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed and B. Lakshminarayanan, *Normalizing flows for probabilistic modeling and inference* (2019), arXiv:1912.02762.

[117] T. Müller, B. McWilliams, F. Rousselle, M. Gross and J. Novák, *Neural importance sampling* (2018), arXiv:1808.03856.

[118] L. Ardizzone, J. Kruse, S. Wirkert, D. Rahner, E. W. Pellegrini, R. S. Klessen, L. Maier-Hein, C. Rother and U. Köthe, *Analyzing inverse problems with invertible neural networks* (2018), arXiv:1808.04730.

[119] M. D. Klimek and M. Perelstein, *Neural Network-Based Approach to Phase Space Integration* (2018), arXiv:1810.11509.

[120] I.-K. Chen, M. D. Klimek and M. Perelstein, *Improved Neural Network Monte Carlo Simulation*, SciPost Phys. **10**, 023 (2021), doi:10.21468/SciPostPhys.10.1.023, arXiv:2009.07819.

[121] E. Bothmann, T. Janßen, M. Knobbe, T. Schmale and S. Schumann, *Exploring phase space with Neural Importance Sampling*, SciPost Phys. **8**, 069 (2020), doi:10.21468/SciPostPhys.8.4.069, arXiv:2001.05478.

[122] C. Gao, J. Isaacson and C. Krause, *i-flow: High-dimensional Integration and Sampling with Normalizing Flows* (2020), arXiv:2001.05486.

[123] C. Gao, S. Höche, J. Isaacson, C. Krause and H. Schulz, *Event Generation with Normalizing Flows*, Phys. Rev. D **101**, 076002 (2020), doi:10.1103/PhysRevD.101.076002, arXiv:2001.10028.

[124] F. Bishara and M. Montull, *(Machine) Learning Amplitudes for Faster Event Generation* (2019), arXiv:1912.11055.

[125] S. Otten *et al.*, *Event Generation and Statistical Sampling with Deep Generative Models and a Density Information Buffer* (2019), arXiv:1901.00875.

[126] B. Hashemi, N. Amin, K. Datta, D. Olivito and M. Pierini, *LHC analysis-specific datasets with Generative Adversarial Networks* (2019), arXiv:1901.05282.

[127] R. Di Sipio, M. Faucci Giannelli, S. Ketabchi Haghighat and S. Palazzo, *DijetGAN: A Generative-Adversarial Network Approach for the Simulation of QCD Dijet Events at the LHC* (2019), arXiv:1903.02433.

[128] A. Butter, T. Plehn and R. Winterhalder, *How to GAN LHC Events*, SciPost Phys. **7**, 075 (2019), doi:10.21468/SciPostPhys.7.6.075, arXiv:1907.03764.

[129] Y. Alanazi, N. Sato, T. Liu, W. Melnitchouk, M. P. Kuchera, E. Pritchard, M. Robertson, R. Strauss, L. Velasco and Y. Li, *Simulation of electron-proton scattering events by a Feature-Augmented and Transformed Generative Adversarial Network (FAT-GAN)* (2020), arXiv:2001.11103.

[130] A. Butter, T. Plehn and R. Winterhalder, *How to GAN Event Subtraction* (2019), doi:10.21468/SciPostPhysCore.3.2.009, arXiv:1912.08824.

[131] B. Stienen and R. Verheyen, *Phase Space Sampling and Inference from Weighted Events with Autoregressive Flows*, SciPost Phys. **10**, 038 (2021), doi:10.21468/SciPostPhys.10.2.038, arXiv:2011.13445.

[132] M. Backes, A. Butter, T. Plehn and R. Winterhalder, *How to GAN Event Unweighting*, SciPost Phys. **10**, 089 (2021), doi:10.21468/SciPostPhys.10.4.089, arXiv:2012.07873.

[133] M. Paganini, L. de Oliveira and B. Nachman, *Accelerating Science with Generative Adversarial Networks: An Application to 3D Particle Showers in Multilayer Calorimeters*, Phys. Rev. Lett. **120**, 042003 (2018), doi:10.1103/PhysRevLett.120.042003, arXiv:1705.02355.

[134] M. Paganini, L. de Oliveira and B. Nachman, *CaloGAN : Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks*, Phys. Rev. **D97**, 014021 (2018), doi:10.1103/PhysRevD.97.014021, arXiv:1712.10321.

[135] P. Musella and F. Pandolfi, *Fast and Accurate Simulation of Particle Detectors Using Generative Adversarial Networks*, Comput. Softw. Big Sci. **2**, 8 (2018), doi:10.1007/s41781-018-0015-y, arXiv:1805.00850.

[136] M. Erdmann, L. Geiger, J. Glombitza and D. Schmidt, *Generating and refining particle detector simulations using the Wasserstein distance in adversarial networks*, Comput. Softw. Big Sci. **2**, 4 (2018), doi:10.1007/s41781-018-0008-x, arXiv:1802.03325.

[137] M. Erdmann, J. Glombitza and T. Quast, *Precise simulation of electromagnetic calorimeter showers using a Wasserstein Generative Adversarial Network*, Comput. Softw. Big Sci. **3**, 4 (2019), doi:10.1007/s41781-018-0019-7, arXiv:1807.01954.

[138] ATLAS Collaboration, *Deep generative models for fast shower simulation in ATLAS*, ATL-SOFT-PUB-2018-001, Http://cds.cern.ch/record/2630433 (2018).

[139] ATLAS Collaboration, *Energy resolution with a GAN for Fast Shower Simulation in ATLAS*, ATLAS-SIM-2019-004, Https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2019-004/ (2019).

[140] D. Belayneh *et al.*, *Calorimetry with Deep Learning: Particle Simulation and Reconstruction for Collider Physics*, Eur. Phys. J. C **80**, 688 (2020), doi:10.1140/epjc/s10052-020-8251-9, arXiv:1912.06794.

[141] E. Buhmann, S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, A. Korol and K. Krüger, *Getting High: High Fidelity Simulation of High Granularity Calorimeters with High Speed* (2020), arXiv:2005.05334.

[142] E. Buhmann, S. Diefenbacher, E. Eren, F. Gaede, G. Kasieczka, A. Korol and K. Krüger, *Decoding Photons: Physics in the Latent Space of a BIB-AE Generative Network* (2021), arXiv:2102.12491.

[143] E. Bothmann and L. Debbio, *Reweighting a parton shower using a neural network: the final-state case*, JHEP **01**, 033 (2019), doi:10.1007/JHEP01(2019)033, arXiv:1808.07802.

[144] L. de Oliveira, M. Paganini and B. Nachman, *Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis*, Comput. Softw. Big Sci. **1**, 4 (2017), doi:10.1007/s41781-017-0004-6, arXiv:1701.05927.

[145] J. W. Monk, *Deep Learning as a Parton Shower*, JHEP **12**, 021 (2018), doi:10.1007/JHEP12(2018)021, arXiv:1807.03685.

[146] K. Dohi, *Variational Autoencoders for Jet Simulation* (2020), arXiv:2009.04842.

[147] J. Lin, W. Bhimji and B. Nachman, *Machine Learning Templates for QCD Factorization in the Search for Physics Beyond the Standard Model*, JHEP **05**, 181 (2019), doi:10.1007/JHEP05(2019)181, arXiv:1903.02556.

[148] B. Nachman and D. Shih, *Anomaly Detection with Density Estimation*, Phys. Rev. D **101**, 075042 (2020), doi:10.1103/PhysRevD.101.075042, arXiv:2001.04990.

[149] O. Knapp, G. Dissertori, O. Cerri, T. Q. Nguyen, J.-R. Vlimant and M. Pierini, *Adversarially Learned Anomaly Detection on CMS Open Data: re-discovering the top quark* (2020), arXiv:2005.01598.

[150] F. A. Di Bello, S. Ganguly, E. Gross, M. Kado, M. Pitt, L. Santi and J. Shlomi, *Towards a Computer Vision Particle Flow*, Eur. Phys. J. C **81**, 107 (2021), doi:10.1140/epjc/s10052-021-08897-0, arXiv:2003.08863.

[151] P. Baldi, L. Blecher, A. Butter, J. Collado, J. N. Howard, F. Keilbach, T. Plehn, G. Kasieczka and D. Whiteson, *How to GAN Higher Jet Resolution* (2020), arXiv:2012.11944.

[152] J. Brehmer and K. Cranmer, *Flows for simultaneous manifold learning and density estimation* (2020), arXiv:2003.13913.

[153] S. T. Radev, U. K. Mertens, A. Voss, L. Ardizzone and U. Köthe, *Bayesflow: Learning complex stochastic models with invertible neural networks* (2020), arXiv:2003.06281.

[154] S. Bieringer, A. Butter, T. Heimel, S. Höche, U. Köthe, T. Plehn and S. T. Radev, *Measuring QCD Splittings with Invertible Networks*, SciPost Phys. **10**, 126 (2021), doi:10.21468/SciPostPhys.10.6.126, arXiv:2012.09873.

[155] K. Datta, D. Kar and D. Roy, *Unfolding with Generative Adversarial Networks* (2018), arXiv:1806.00433.

[156] M. Bellagente, A. Butter, G. Kasieczka, T. Plehn, A. Rousselot, R. Winterhalder, L. Ardizzone and U. Köthe, *Invertible Networks or Partons to Detector and Back Again*, SciPost Phys. **9**, 074 (2020), doi:10.21468/SciPostPhys.9.5.074, arXiv:2006.06685.

[157] P. C. Bhat and H. B. Prosper, *Bayesian neural networks*, Conf. Proc. C **050912**, 151 (2005), `http://www.physics.ox.ac.uk/phystat05/proceedings/files//bhat_prosper_phystat05.pdf`.

[158] S. R. Saucedo, *Bayesian Neural Networks for Classification*, Master's thesis, Florida State University (2007).

[159] Y. Xu, W. Xu, Y. Meng, K. Zhu and W. Xu, *Applying Bayesian Neural Networks to Event Reconstruction in Reactor Neutrino Experiments*, Nucl. Instrum. Meth. A **592**, 451 (2008), doi:10.1016/j.nima.2008.04.006, arXiv:0712.4042.

[160] J. Y. Araz and M. Spannowsky, *Combine and Conquer: Event Reconstruction with Bayesian Ensemble Neural Networks* (2021), arXiv:2102.01078.

[161] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli and M. Zaro, *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07**, 079 (2014), doi:10.1007/JHEP07(2014)079, arXiv:1405.0301.

[162] R. D. Ball, V. Bertone, S. Carrazza, L. Del Debbio, S. Forte, A. Guffanti, N. P. Hartland and J. Rojo, *Parton distributions with QED corrections*, Nucl. Phys. **B877**, 290 (2013), doi:10.1016/j.nuclphysb.2013.10.010, arXiv:1308.0598.

[163] S. Badger *et al.*, *Machine Learning and LHC Event Generation* (2022), arXiv:2203.07460.

[164] M. Czakon, *Tops from Light Quarks: Full Mass Dependence at Two-Loops in QCD*, Phys. Lett. B **664**, 307 (2008), doi:10.1016/j.physletb.2008.05.028, arXiv:0803.1400.

[165] S. Borowka, N. Greiner, G. Heinrich, S. P. Jones, M. Kerner, J. Schlenk, U. Schubert and T. Zirke, *Higgs Boson Pair Production in Gluon Fusion at Next-to-Leading Order with Full Top-Quark Mass Dependence*, Phys. Rev. Lett. **117**, 012001 (2016), doi:10.1103/PhysRevLett.117.079901, [Erratum: Phys.Rev.Lett. 117, 079901 (2016)], arXiv:1604.06447.

[166] K. Danziger, T. Janßen, S. Schumann and F. Siegert, *Accelerating Monte Carlo event generation – rejection sampling using neural network event-weight estimates* (2021), arXiv:2109.11964.

[167] D. Maître and H. Truong, *A factorisation-aware Matrix element emulator*, JHEP **11**, 066 (2021), doi:10.1007/JHEP11(2021)066, arXiv:2107.06625.

[168] J. Aylett-Bullock, S. Badger and R. Moodie, *Optimising simulations for diphoton production at hadron colliders using amplitude neural networks*, JHEP **08**, 066 (2021), doi:10.1007/JHEP08(2021)066, arXiv:2106.09474.

[169] A. Butter, T. Heimel, S. Hummerich, T. Krebs, T. Plehn, A. Rousselot and S. Vent, *Generative Networks for Precision Enthusiasts* (2021), arXiv:2110.13632.

[170] I. Chahrour and J. D. Wells, *Comparing Machine Learning and Interpolation Methods for Loop-Level Calculations*, SciPost Phys. **12**, 187 (2022), doi:10.21468/SciPostPhys.12.6.187, arXiv:2111.14788.

[171] E. Bothmann *et al.*, *Event Generation with Sherpa 2.2*, SciPost Phys. **7**, 034 (2019), doi:10.21468/SciPostPhys.7.3.034, arXiv:1905.09127.

[172] S. Badger, B. Biedermann, P. Uwer and V. Yundin, *Numerical evaluation of virtual corrections to multi-jet production in massless QCD*, Comput. Phys. Commun. **184**, 1981 (2013), doi:10.1016/j.cpc.2013.03.018, arXiv:1209.0100.

[173] R. Kleiss, W. J. Stirling and S. D. Ellis, *A New Monte Carlo Treatment of Multiparticle Phase Space at High-energies*, Comput. Phys. Commun. **40**, 359 (1986), doi:10.1016/0010-4655(86)90119-0.

[174] S. Dawson, C. Englert and T. Plehn, *Higgs Physics: It ain't over till it's over*, Phys. Rept. **816**, 1 (2019), doi:10.1016/j.physrep.2019.05.001, arXiv:1808.01324.

[175] I. Brivio and M. Trott, *The Standard Model as an Effective Field Theory*, Phys. Rept. **793**, 1 (2019), doi:10.1016/j.physrep.2018.11.002, arXiv:1706.08945.

[176] G. Aad *et al.*, *Combined measurements of Higgs boson production and decay using up to* 80 fb^{-1} *of proton-proton collision data at* $\sqrt{s} = 13$ *TeV collected with the ATLAS experiment*, Phys. Rev. D **101**, 012002 (2020), doi:10.1103/PhysRevD.101.012002, arXiv:1909.02845.

[177] A. M. Sirunyan *et al.*, *Combined measurements of Higgs boson couplings in proton–proton collisions at* $\sqrt{s} = 13\,TeV$, Eur. Phys. J. C **79**, 421 (2019), doi:10.1140/epjc/s10052-019-6909-y, arXiv:1809.10733.

[178] A. Biekoetter, T. Corbett and T. Plehn, *The Gauge-Higgs Legacy of the LHC Run II*, SciPost Phys. **6**, 064 (2019), doi:10.21468/SciPostPhys.6.6.064, arXiv:1812.07587.

[179] J. Ellis, C. W. Murphy, V. Sanz and T. You, *Updated Global SMEFT Fit to Higgs, Diboson and Electroweak Data*, JHEP **06**, 146 (2018), doi:10.1007/JHEP06(2018)146, arXiv:1803.03252.

[180] E. da Silva Almeida, A. Alves, N. Rosa Agostinho, O. J. P. Éboli and M. C. Gonzalez-Garcia, *Electroweak Sector Under Scrutiny: A Combined Analysis of LHC and Electroweak Precision Data*, Phys. Rev. D **99**, 033001 (2019), doi:10.1103/PhysRevD.99.033001, arXiv:1812.01009.

[181] S. Kraml, T. Q. Loc, D. T. Nhung and L. D. Ninh, *Constraining new physics from Higgs measurements with Lilith: update to LHC Run 2 results*, SciPost Phys. **7**, 052 (2019), doi:10.21468/SciPostPhys.7.4.052, arXiv:1908.03952.

[182] S. van Beek, E. R. Nocera, J. Rojo and E. Slade, *Constraining the SMEFT with Bayesian reweighting*, SciPost Phys. **7**, 070 (2019), doi:10.21468/SciPostPhys.7.5.070, arXiv:1906.05296.

[183] S. Dawson, S. Homiller and S. D. Lane, *Putting standard model EFT fits to work*, Phys. Rev. D **102**, 055012 (2020), doi:10.1103/PhysRevD.102.055012, arXiv:2007.01296.

[184] J. Ellis, M. Madigan, K. Mimasu, V. Sanz and T. You, *Top, Higgs, Diboson and Electroweak Fit to the Standard Model Effective Field Theory*, JHEP **04**, 279 (2021), doi:10.1007/JHEP04(2021)279, arXiv:2012.02779.

[185] J. Brehmer, A. Freitas, D. Lopez-Val and T. Plehn, *Pushing Higgs Effective Theory to its Limits*, Phys. Rev. D **93**, 075014 (2016), doi:10.1103/PhysRevD.93.075014, arXiv:1510.03443.

[186] A. Biekötter, J. Brehmer and T. Plehn, *Extending the limits of Higgs effective theory*, Phys. Rev. **D94**, 055032 (2016), doi:10.1103/PhysRevD.94.055032, arXiv:1602.05202.

[187] Anisha, S. Das Bakshi, J. Chakrabortty and S. K. Patra, *Connecting electroweak-scale observables to BSM physics through EFT and Bayesian statistics*, Phys. Rev. D **103**, 076007 (2021), doi:10.1103/PhysRevD.103.076007, arXiv:2010.04088.

[188] S. Das Bakshi, J. Chakrabortty and M. Spannowsky, *Classifying Standard Model Extensions Effectively with Precision Observables*, Phys. Rev. D **103**, 056019 (2021), doi:10.1103/PhysRevD.103.056019, arXiv:2012.03839.

[189] C. Englert, R. Kogler, H. Schulz and M. Spannowsky, *Higgs coupling measurements at the lhc*, The European Physical Journal C **76** (2016), doi:10.1140/epjc/s10052-016-4227-1, `http://dx.doi.org/10.1140/epjc/s10052-016-4227-1`.

[190] S. Das Bakshi, J. Chakrabortty, C. Englert, M. Spannowsky and P. Stylianou, *Cp violation at atlas in effective field theory*, Physical Review D **103** (2021), doi:10.1103/physrevd.103.055008, `http://dx.doi.org/10.1103/PhysRevD.103.055008`.

[191] S. Das Bakshi, J. Chakrabortty, S. Prakash, S. U. Rahaman and M. Spannowsky, *EFT diagrammatica: UV roots of the CP-conserving SMEFT*, JHEP **06**, 033 (2021), doi:10.1007/JHEP06(2021)033, arXiv:2103.11593.

[192] Anisha, S. D. Bakshi, S. Banerjee, A. Biekötter, J. Chakrabortty, S. K. Patra and M. Spannowsky, *Effective limits on single scalar extensions in the light of recent lhc data* (2021), arXiv:2111.05876.

[193] W. Kilian, T. Ohl, J. Reuter and M. Sekulla, *Resonances at the LHC beyond the Higgs boson: The scalar/tensor case*, Phys. Rev. D **93**, 036004 (2016), doi:10.1103/PhysRevD.93.036004, arXiv:1511.00022.

[194] A. Freitas, D. López-Val and T. Plehn, *When matching matters: Loop effects in Higgs effective theory*, Phys. Rev. D **94**, 095007 (2016), doi:10.1103/PhysRevD.94.095007, arXiv:1607.08251.

[195] I. Low, R. Rattazzi and A. Vichi, *Theoretical Constraints on the Higgs Effective Couplings*, JHEP **04**, 126 (2010), doi:10.1007/JHEP04(2010)126, arXiv:0907.5413.

[196] F. del Aguila, J. de Blas and M. Perez-Victoria, *Electroweak Limits on General New Vector Bosons*, JHEP **09**, 033 (2010), doi:10.1007/JHEP09(2010)033, arXiv:1005.3998.

[197] D. Pappadopulo, A. Thamm, R. Torre and A. Wulzer, *Heavy Vector Triplets: Bridging Theory and Data*, JHEP **09**, 060 (2014), doi:10.1007/JHEP09(2014)060, arXiv:1402.4431.

[198] A. Biekötter, A. Knochel, M. Krämer, D. Liu and F. Riva, *Vices and virtues of Higgs effective field theories at large energy*, Phys. Rev. D **91**, 055029 (2015), doi:10.1103/PhysRevD.91.055029, arXiv:1406.7320.

[199] V. D. Barger, W.-Y. Keung and E. Ma, *Doubling of Weak Gauge Bosons in an Extension of the Standard Model*, Phys. Rev. Lett. **44**, 1169 (1980), doi:10.1103/PhysRevLett.44.1169.

[200] N. Arkani-Hamed, A. G. Cohen and H. Georgi, *Electroweak symmetry breaking from dimensional deconstruction*, Phys. Lett. B **513**, 232 (2001), doi:10.1016/S0370-2693(01)00741-9, arXiv:hep-ph/0105239.

[201] T. Appelquist and J. Carazzone, *Infrared Singularities and Massive Fields*, Phys. Rev. D **11**, 2856 (1975), doi:10.1103/PhysRevD.11.2856.

[202] S. Weinberg, *Implications of Dynamical Symmetry Breaking*, Phys. Rev. D **13**, 974 (1976), doi:10.1103/PhysRevD.19.1277, [Addendum: Phys.Rev.D 19, 1277–1280 (1979)].

[203] R. D. C. Miller, *Effective Field Theory and Weak Nonleptonic Interactions*, Phys. Rept. **106**, 169 (1984), doi:10.1016/0370-1573(84)90035-8.

[204] M. K. Gaillard, *The Effective One Loop Lagrangian With Derivative Couplings*, Nucl. Phys. **B 268**, 669 (1986), doi:10.1016/0550-3213(86)90264-6.

[205] L.-H. Chan, *Derivative Expansion for the One Loop Effective Actions With Internal Symmetry*, Phys. Rev. Lett. **57**, 1199 (1986), doi:10.1103/PhysRevLett.57.1199.

[206] O. Cheyette, *Effective Action for the Standard Model With Large Higgs Mass*, Nucl. Phys. **B 297**, 183 (1988), doi:10.1016/0550-3213(88)90205-2.

[207] R. D. Ball, *Chiral Gauge Theory*, Phys. Rept. **182**, 1 (1989), doi:10.1016/0370-1573(89)90027-6.

[208] B. Henning, X. Lu and H. Murayama, *How to use the Standard Model effective field theory*, JHEP **01**, 023 (2016), doi:10.1007/JHEP01(2016)023, arXiv:1412.1837.

[209] B. Henning, X. Lu and H. Murayama, *One-loop Matching and Running with Covariant Derivative Expansion*, JHEP **01**, 123 (2018), doi:10.1007/JHEP01(2018)123, arXiv:1604.01019.

[210] A. Drozd, J. Ellis, J. Quevillon and T. You, *The Universal One-Loop Effective Action*, JHEP **03**, 180 (2016), doi:10.1007/JHEP03(2016)180, arXiv:1512.03003.

[211] S. A. R. Ellis, J. Quevillon, T. You and Z. Zhang, *Extending the Universal One-Loop Effective Action: Heavy-Light Coefficients*, JHEP **08**, 054 (2017), doi:10.1007/JHEP08(2017)054, arXiv:1706.07765.

[212] J. Fuentes-Martin, J. Portoles and P. Ruiz-Femenia, *Integrating out heavy particles with functional methods: a simplified framework*, JHEP **09**, 156 (2016), doi:10.1007/JHEP09(2016)156, arXiv:1607.02142.

[213] Z. Zhang, *Covariant diagrams for one-loop matching*, JHEP **05**, 152 (2017), doi:10.1007/JHEP05(2017)152, arXiv:1610.00710.

[214] S. Das Bakshi, J. Chakrabortty and S. K. Patra, *CoDEx: Wilson coefficient calculator connecting SMEFT to UV theory*, Eur. Phys. J. C **79**, 21 (2019), doi:10.1140/epjc/s10052-018-6444-2, arXiv:1808.04403.

[215] M. Krämer, B. Summ and A. Voigt, *Completing the scalar and fermionic Universal One-Loop Effective Action*, JHEP **01**, 079 (2020), doi:10.1007/JHEP01(2020)079, arXiv:1908.04798.

[216] S. A. R. Ellis, J. Quevillon, P. N. H. Vuong, T. You and Z. Zhang, *The fermionic universal one-loop effective action*, Journal of High Energy Physics **2020** (2020), doi:10.1007/jhep11(2020)078, http://dx.doi.org/10.1007/JHEP11(2020)078, arXiv:2006.16260.

[217] A. Angelescu and P. Huang, *Integrating out new fermions at one loop*, Journal of High Energy Physics **2021** (2021), doi:10.1007/jhep01(2021)049, http://dx.doi.org/10.1007/JHEP01(2021)049, arXiv:2006.16532.

[218] T. Cohen, X. Lu and Z. Zhang, *STrEAMlining EFT Matching*, SciPost Phys. **10**, 098 (2021), doi:10.21468/SciPostPhys.10.5.098, arXiv:2012.07851.

[219] J. Fuentes-Martin, M. König, J. Pagès, A. E. Thomsen and F. Wilsch, *SuperTracer: A Calculator of Functional Supertraces for One-Loop EFT Matching*, JHEP **04**, 281 (2021), doi:10.1007/JHEP04(2021)281, arXiv:2012.08506.

[220] S. Dittmaier, S. Schuhmacher and M. Stahlhofen, *Integrating out heavy fields in the path integral using the background-field method: general formalism*, The European Physical Journal C **81** (2021), doi:10.1140/epjc/s10052-021-09587-7, http://dx.doi.org/10.1140/epjc/s10052-021-09587-7, arXiv:2102.12020.

[221] D. Zhang and S. Zhou, *Complete one-loop matching of the type-i seesaw model onto the standard model effective field theory*, Journal of High Energy Physics **2021** (2021), doi:10.1007/jhep09(2021)163, http://dx.doi.org/10.1007/JHEP09(2021)163, arXiv:2107.12133.

[222] B. S. DeWitt, *Quantum Theory of Gravity. 2. The Manifestly Covariant Theory*, Phys. Rev. **162**, 1195 (1967), doi:10.1103/PhysRev.162.1195.

[223] W. A. Bardeen, *Anomalous Ward identities in spinor field theories*, Phys. Rev. **184**, 1848 (1969), doi:10.1103/PhysRev.184.1848.

[224] G. 't Hooft, *The Background Field Method in Gauge Field Theories*, In *12th Annual Winter School of Theoretical Physics*, pp. 345–369 (1975).

[225] B. S. DeWitt, *A gauge invariant effective action*, In *Oxford Conference on Quantum Gravity*, pp. 449–487 (1980).

[226] D. G. Boulware, *Gauge Dependence of the Effective Action*, Phys. Rev. **D 23**, 389 (1981), doi:10.1103/PhysRevD.23.389.

[227] L. F. Abbott, *The Background Field Method Beyond One Loop*, Nucl. Phys. **B 185**, 189 (1981), doi:10.1016/0550-3213(81)90371-0.

[228] L. F. Abbott, *Introduction to the Background Field Method*, Acta Phys. Polon. **B 13**, 33 (1982).

[229] A. Denner, G. Weiglein and S. Dittmaier, *Application of the background field method to the electroweak standard model*, Nucl. Phys. B **440**, 95 (1995), doi:10.1016/0550-3213(95)00037-S, arXiv:hep-ph/9410338.

[230] A. Denner and S. Dittmaier, *Electroweak Radiative Corrections for Collider Physics*, Phys. Rept. **864**, 1 (2020), doi:10.1016/j.physrep.2020.04.001, arXiv:1912.06823.

[231] S. Dittmaier and C. Grosse-Knetter, *Deriving nondecoupling effects of heavy fields from the path integral: A Heavy Higgs field in an SU(2) gauge theory*, Phys. Rev. D **52**, 7276 (1995), doi:10.1103/PhysRevD.52.7276, arXiv:hep-ph/9501285.

[232] S. Dittmaier and C. Grosse-Knetter, *Integrating out the standard Higgs field in the path integral*, Nucl. Phys. B **459**, 497 (1996), doi:10.1016/0550-3213(95)00551-X, arXiv:hep-ph/9505266.

[233] J. Preskill, *Gauge anomalies in an effective field theory*, Annals Phys. **210**, 323 (1991), doi:10.1016/0003-4916(91)90046-B.

[234] O. Cata, W. Kilian and N. Kreher, *Gauge anomalies in the Standard-Model Effective Field Theory* (2020), arXiv:2011.09976.

[235] M. Beneke and V. A. Smirnov, *Asymptotic expansion of Feynman integrals near threshold*, Nucl. Phys. **B 522**, 321 (1998), doi:10.1016/S0550-3213(98)00138-2, arXiv:hep-ph/9711391.

[236] V. A. Smirnov, *Applied asymptotic expansions in momenta and masses*, Springer Tracts Mod. Phys. **177**, 1 (2002).

[237] B. Jantzen, *Foundation and generalization of the expansion by regions*, JHEP **12**, 076 (2011), doi:10.1007/JHEP12(2011)076, arXiv:1111.2589.

[238] B. Summ, *One Formula To Match Them All: The Bispinor Universal One-Loop Effective Action*, Ph.D. thesis, RWTH Aachen U., doi:10.18154/RWTH-2021-00828 (2020), arXiv:2103.02487.

[239] M. Bauer, P. Foldenauer and J. Jaeckel, *Hunting All the Hidden Photons*, JHEP **07**, 094 (2018), doi:10.1007/JHEP07(2018)094, arXiv:1803.05466.

[240] T. Cohen, N. Craig, X. Lu and D. Sutherland, *Is SMEFT Enough?*, JHEP **03**, 237 (2021), doi:10.1007/JHEP03(2021)237, arXiv:2008.08597.

[241] R. Lafaye, T. Plehn and D. Zerwas, *SFITTER: SUSY parameter analysis at LHC and LC* (2004), arXiv:hep-ph/0404282.

[242] M. Klute, R. Lafaye, T. Plehn, M. Rauch and D. Zerwas, *Measuring Higgs Couplings from LHC Data*, Phys. Rev. Lett. **109**, 101801 (2012), doi:10.1103/PhysRevLett.109.101801, arXiv:1205.2699.

[243] T. Corbett, O. J. P. Eboli, D. Goncalves, J. Gonzalez-Fraile, T. Plehn and M. Rauch, *The Higgs Legacy of the LHC Run I*, JHEP **08**, 156 (2015), doi:10.1007/JHEP08(2015)156, arXiv:1505.05516.

[244] T. Corbett, O. J. P. Eboli, D. Goncalves, J. Gonzalez-Fraile, T. Plehn and M. Rauch, *The Non-Linear Higgs Legacy of the LHC Run I* (2015), arXiv:1511.08188.

[245] A. Butter, O. J. P. Eboli, J. Gonzalez-Fraile, M. C. Gonzalez-Garcia, T. Plehn and M. Rauch, *The Gauge-Higgs Legacy of the LHC Run I*, JHEP **07**, 152 (2016), doi:10.1007/JHEP07(2016)152, arXiv:1604.03105.

[246] I. Brivio, S. Bruggisser, F. Maltoni, R. Moutafis, T. Plehn, E. Vryonidou, S. Westhoff and C. Zhang, *O new physics, where art thou? A global search in the top sector*, JHEP **02**, 131 (2020), doi:10.1007/JHEP02(2020)131, arXiv:1910.03606.

[247] D. López-Val, T. Plehn and M. Rauch, *Measuring extended Higgs sectors as a consistent free couplings model*, JHEP **10**, 134 (2013), doi:10.1007/JHEP10(2013)134, arXiv:1308.1979.

[248] A. Hocker, H. Lacker, S. Laplace and F. Le Diberder, *A New approach to a global fit of the CKM matrix*, Eur. Phys. J. C **21**, 225 (2001), doi:10.1007/s100520100729, arXiv:hep-ph/0104062.

[249] K. Hagiwara, S. Ishihara, R. Szalapski and D. Zeppenfeld, *Low-energy effects of new interactions in the electroweak boson sector*, Phys. Rev. D **48**, 2182 (1993), doi:10.1103/PhysRevD.48.2182.

[250] `github.com/BenjaminSumm/Triplet_Model_WCs`.

[251] M. Aaboud *et al.*, *Search for heavy resonances decaying into a W or Z boson and a Higgs boson in final states with leptons and b-jets in 36 fb^{-1} of $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector*, JHEP **03**, 174 (2018), doi:10.1007/JHEP03(2018)174, [Erratum: JHEP 11, 051 (2018)], arXiv:1712.06518.

[252] ATLAS Collaboration, *Search for resonances decaying into a weak vector boson and a Higgs boson in the fully hadronic final state produced in proton−proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Rev. D **102**, 112008 (2020), doi:10.1103/PhysRevD.102.112008, arXiv:2007.05293.

[253] G. Aad *et al.*, *Search for heavy diboson resonances in semileptonic final states in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Eur. Phys. J. C **80**, 1165 (2020), doi:10.1140/epjc/s10052-020-08554-y, arXiv:2004.14636.

[254] J. Brehmer, S. Dawson, S. Homiller, F. Kling and T. Plehn, *Benchmarking simplified template cross sections in WH production*, JHEP **11**, 034 (2019), doi:10.1007/JHEP11(2019)034, arXiv:1908.06980.

[255] M. Trott, *Methodology for theory uncertainties in the standard model effective field theory*, Physical Review D **104** (2021), doi:10.1103/physrevd.104.095023, `http://dx.doi.org/10.1103/PhysRevD.104.095023`, arXiv:2106.13794.

[256] J. Baglio, S. Dawson, S. Homiller, S. D. Lane and I. M. Lewis, *Validity of standard model EFT studies of VH and VV production at NLO*, Phys. Rev. D **101**, 115004 (2020), doi:10.1103/PhysRevD.101.115004, arXiv:2003.07862.

[257] R. Lafaye, T. Plehn, M. Rauch, D. Zerwas and M. Duhrssen, *Measuring the Higgs Sector*, JHEP **08**, 009 (2009), doi:10.1088/1126-6708/2009/08/009, arXiv:0904.3866.

[258] W. Kilian, T. Plehn, P. Richardson and E. Schmidt, *Split supersymmetry at colliders*, Eur. Phys. J. C **39**, 229 (2005), doi:10.1140/epjc/s2004-02046-5, arXiv:hep-ph/0408088.

[259] S. Dawson, P. P. Giardino and S. Homiller, *Uncovering the High Scale Higgs Singlet Model*, Phys. Rev. D **103**, 075016 (2021), doi:10.1103/PhysRevD.103.075016, arXiv:2102.02823.

[260] U. Haisch, M. Ruhdorfer, E. Salvioni, E. Venturini and A. Weiler, *Singlet night in Feynman-ville: one-loop matching of a real scalar*, JHEP **04**, 164 (2020), doi:10.1007/JHEP04(2020)164, [Erratum: JHEP 07, 066 (2020)], arXiv:2003.05936.

[261] Z. Zhang, *Time to Go Beyond Triple-Gauge-Boson-Coupling Interpretation of W Pair Production*, Phys. Rev. Lett. **118**, 011803 (2017), doi:10.1103/PhysRevLett.118.011803, arXiv:1610.01618.

[262] T. Corbett, O. J. P. Éboli and M. C. Gonzalez-Garcia, *Unitarity Constraints on Dimension-six Operators II: Including Fermionic Operators*, Phys. Rev. D **96**, 035006 (2017), doi:10.1103/PhysRevD.96.035006, arXiv:1705.09294.

[263] J. Baglio, S. Dawson and I. M. Lewis, *An NLO QCD effective field theory analysis of W^+W^- production at the LHC including fermionic operators*, Phys. Rev. D **96**, 073003 (2017), doi:10.1103/PhysRevD.96.073003, arXiv:1708.03332.

[264] J. Baglio, S. Dawson and I. M. Lewis, *NLO effects in EFT fits to W^+W^- production at the LHC*, Phys. Rev. D **99**, 035029 (2019), doi:10.1103/PhysRevD.99.035029, arXiv:1812.00214.

[265] A. Alves, N. Rosa-Agostinho, O. J. P. Éboli and M. C. Gonzalez-Garcia, *Effect of Fermionic Operators on the Gauge Legacy of the LHC Run I*, Phys. Rev. D **98**, 013006 (2018), doi:10.1103/PhysRevD.98.013006, arXiv:1805.11108.

[266] S. Dawson and A. Ismail, *Standard model EFT corrections to Z boson decays*, Phys. Rev. D **98**, 093003 (2018), doi:10.1103/PhysRevD.98.093003, arXiv:1808.05948.

[267] A. Falkowski, M. Gonzalez-Alonso, A. Greljo, D. Marzocca and M. Son, *Anomalous Triple Gauge Couplings in the Effective Field Theory Approach at the LHC*, JHEP **02**, 115 (2017), doi:10.1007/JHEP02(2017)115, arXiv:1609.06312.

[268] J. Lang, S. Liebler, H. Schäfer-Siebert and D. Zeppenfeld, *Effective field theory versus uv-complete model: vector boson scattering as a case study*, The European Physical Journal C **81** (2021), doi:10.1140/epjc/s10052-021-09428-7, `http://dx.doi.org/10.1140/epjc/s10052-021-09428-7`, arXiv:2103.16517.

[269] E. d. S. Almeida, A. Alves, O. J. P. Éboli and M. C. Gonzalez-Garcia, *Electroweak legacy of the LHC run II*, Phys. Rev. D **105**, 013006 (2022), doi:10.1103/PhysRevD.105.013006, arXiv:2108.04828.

[270] S. Brown, A. Buckley, C. Englert, J. Ferrando, P. Galler, D. J. Miller, L. Moore, M. Russell, C. White and N. Warrack, *TopFitter: Fitting top-quark Wilson Coefficients to Run II data*, PoS **ICHEP2018**, 293 (2019), doi:10.22323/1.340.0293, arXiv:1901.03164.

[271] N. P. Hartland, F. Maltoni, E. R. Nocera, J. Rojo, E. Slade, E. Vryonidou and C. Zhang, *A Monte Carlo global analysis of the Standard Model Effective Field Theory: the top quark sector*, JHEP **04**, 100 (2019), doi:10.1007/JHEP04(2019)100, arXiv:1901.05965.

[272] J. J. Ethier, G. Magni, F. Maltoni, L. Mantani, E. R. Nocera, J. Rojo, E. Slade, E. Vryonidou and C. Zhang, *Combined SMEFT interpretation of Higgs, diboson, and top quark data from the LHC*, JHEP **11**, 089 (2021), doi:10.1007/JHEP11(2021)089, arXiv:2105.00006.

[273] S. Iranipour and M. Ubiali, *A new generation of simultaneous fits to LHC data using deep learning*, JHEP **05**, 032 (2022), doi:10.1007/JHEP05(2022)032, arXiv:2201.07240.

[274] S. Bißmann, J. Erdmann, C. Grunwald, G. Hiller and K. Kröninger, *Correlating uncertainties in global analyses within SMEFT matters*, Phys. Rev. D **102**, 115019 (2020), doi:10.1103/PhysRevD.102.115019, arXiv:1912.06090.

[275] R. Lafaye, T. Plehn, M. Rauch and D. Zerwas, *Measuring Supersymmetry*, Eur. Phys. J. C **54**, 617 (2008), doi:10.1140/epjc/s10052-008-0548-z, arXiv:0709.3985.

[276] H. Georgi, *Weak Interactions and Modern Particle Theory*, ISBN 978-0-8053-3163-9 (1984).

[277] J. F. Donoghue, E. Golowich and B. R. Holstein, *Dynamics of the standard model*, vol. 2, CUP, doi:10.1017/CBO9780511524370 (2014).

[278] C. N. Leung, S. T. Love and S. Rao, *Low-Energy Manifestations of a New Interaction Scale: Operator Analysis*, Z. Phys. C **31**, 433 (1986), doi:10.1007/BF01588041.

[279] W. Buchmuller and D. Wyler, *Effective Lagrangian Analysis of New Interactions and Flavor Conservation*, Nucl. Phys. B **268**, 621 (1986), doi:10.1016/0550-3213(86)90262-2.

[280] M. C. Gonzalez-Garcia, *Anomalous Higgs couplings*, Int. J. Mod. Phys. A **14**, 3121 (1999), doi:10.1142/S0217751X99001494, arXiv:hep-ph/9902321.

[281] B. Grzadkowski, M. Iskrzynski, M. Misiak and J. Rosiek, *Dimension-Six Terms in the Standard Model Lagrangian*, JHEP **10**, 085 (2010), doi:10.1007/JHEP10(2010)085, arXiv:1008.4884.

[282] G. Passarino, *NLO Inspired Effective Lagrangians for Higgs Physics*, Nucl. Phys. B **868**, 416 (2013), doi:10.1016/j.nuclphysb.2012.11.018, arXiv:1209.5538.

[283] T. Corbett, O. J. P. Eboli, J. Gonzalez-Fraile and M. C. Gonzalez-Garcia, *Robust Determination of the Higgs Couplings: Power to the Data*, Phys. Rev. D **87**, 015022 (2013), doi:10.1103/PhysRevD.87.015022, arXiv:1211.4580.

[284] S. Di Vita, C. Grojean, G. Panico, M. Riembau and T. Vantalon, *A global view on the Higgs self-coupling*, JHEP **09**, 069 (2017), doi:10.1007/JHEP09(2017)069, arXiv:1704.01953.

[285] D. Gonçalves, T. Han, F. Kling, T. Plehn and M. Takeuchi, *Higgs boson pair production at future hadron colliders: From kinematics to dynamics*, Phys. Rev. D **97**, 113004 (2018), doi:10.1103/PhysRevD.97.113004, arXiv:1802.04319.

[286] J. Chang, K. Cheung, J. S. Lee, C.-T. Lu and J. Park, *Higgs-boson-pair production $H(\to b\bar{b})H(\to \gamma\gamma)$ from gluon fusion at the HL-LHC and HL-100 TeV hadron collider*, Phys. Rev. D **100**, 096001 (2019), doi:10.1103/PhysRevD.100.096001, arXiv:1804.07130.

[287] A. Biekötter, D. Gonçalves, T. Plehn, M. Takeuchi and D. Zerwas, *The global Higgs picture at 27 TeV*, SciPost Phys. **6**, 024 (2019), doi:10.21468/SciPostPhys.6.2.024, arXiv:1811.08401.

[288] S. Borowka, C. Duhr, F. Maltoni, D. Pagani, A. Shivaji and X. Zhao, *Probing the scalar potential via double Higgs boson production at hadron colliders*, JHEP **04**, 016 (2019), doi:10.1007/JHEP04(2019)016, arXiv:1811.12366.

[289] F. Krauss, S. Kuttimalai and T. Plehn, *LHC multijet events as a probe for anomalous dimension-six gluon interactions*, Phys. Rev. D **95**, 035024 (2017), doi:10.1103/PhysRevD.95.035024, arXiv:1611.00767.

[290] V. Cirigliano, J. Jenkins and M. Gonzalez-Alonso, *Semileptonic decays of light quarks beyond the Standard Model*, Nucl. Phys. B **830**, 95 (2010), doi:10.1016/j.nuclphysb.2009.12.020, arXiv:0908.1754.

[291] A. Falkowski, M. González-Alonso and K. Mimouni, *Compilation of low-energy constraints on 4-fermion operators in the SMEFT*, JHEP **08**, 123 (2017), doi:10.1007/JHEP08(2017)123, arXiv:1706.03783.

[292] S. Alioli, V. Cirigliano, W. Dekens, J. de Vries and E. Mereghetti, *Right-handed charged currents in the era of the Large Hadron Collider*, JHEP **05**, 086 (2017), doi:10.1007/JHEP05(2017)086, arXiv:1703.04751.

[293] F. Maltoni, E. Vryonidou and C. Zhang, *Higgs production in association with a top-antitop pair in the Standard Model Effective Field Theory at NLO in QCD*, JHEP **10**, 123 (2016), doi:10.1007/JHEP10(2016)123, arXiv:1607.05330.

[294] M. Grazzini, A. Ilnicka, M. Spira and M. Wiesemann, *Modeling BSM effects on the Higgs transverse-momentum spectrum in an EFT approach*, JHEP **03**, 115 (2017), doi:10.1007/JHEP03(2017)115, arXiv:1612.00283.

[295] N. Deutschmann, C. Duhr, F. Maltoni and E. Vryonidou, *Gluon-fusion Higgs production in the Standard Model Effective Field Theory*, JHEP **12**, 063 (2017), doi:10.1007/JHEP12(2017)063, [Erratum: JHEP 02, 159 (2018)], arXiv:1708.00460.

[296] R. Alonso, E. E. Jenkins, A. V. Manohar and M. Trott, *Renormalization Group Evolution of the Standard Model Dimension Six Operators III: Gauge Coupling Dependence and Phenomenology*, JHEP **04**, 159 (2014), doi:10.1007/JHEP04(2014)159, arXiv:1312.2014.

[297] S. Dawson, S. Homiller and M. Sullivan, *Impact of dimension-eight SMEFT contributions: A case study*, Phys. Rev. D **104**, 115013 (2021), doi:10.1103/PhysRevD.104.115013, arXiv:2110.06929.

[298] S. Dawson, D. Fontes, S. Homiller and M. Sullivan, *Beyond 6: the role of dimension-8 operators in an EFT for the 2HDM* (2022), arXiv:2205.01561.

[299] R. Cepedello, F. Esser, M. Hirsch and V. Sanz, *Mapping the SMEFT to discoverable models* (2022), arXiv:2207.13714.

[300] S. Brass, C. Fleper, W. Kilian, J. Reuter and M. Sekulla, *Transversal Modes and Higgs Bosons in Electroweak Vector-Boson Scattering at the LHC*, Eur. Phys. J. C **78**, 931 (2018), doi:10.1140/epjc/s10052-018-6398-4, arXiv:1807.02512.

[301] T. Cohen, X. Lu and Z. Zhang, *Functional Prescription for EFT Matching*, JHEP **02**, 228 (2021), doi:10.1007/JHEP02(2021)228, arXiv:2011.02484.

[302] B. Dumont, S. Fichet and G. von Gersdorff, *A Bayesian view of the Higgs sector with higher dimensional operators*, JHEP **07**, 065 (2013), doi:10.1007/JHEP07(2013)065, arXiv:1304.3369.

[303] S. Fichet and G. Moreau, *Anatomy of the Higgs fits: a first guide to statistical treatments of the theoretical uncertainties*, Nucl. Phys. B **905**, 391 (2016), doi:10.1016/j.nuclphysb.2016.02.019, arXiv:1509.00472.

[304] J. De Blas *et al.*, *HEPfit: a code for the combination of indirect and direct constraints on high energy physics models*, Eur. Phys. J. C **80**, 456 (2020), doi:10.1140/epjc/s10052-020-7904-z, arXiv:1910.14012.

[305] J. de Blas, M. Ciuchini, E. Franco, A. Goncalves, S. Mishima, M. Pierini, L. Reina and L. Silvestrini, *Global analysis of electroweak data in the Standard Model*, Phys. Rev. D **106**, 033003 (2022), doi:10.1103/PhysRevD.106.033003, arXiv:2112.07274.

[306] J. Brehmer, K. Cranmer, F. Kling and T. Plehn, *Better Higgs boson measurements through information geometry*, Phys. Rev. D **95**, 073002 (2017), doi:10.1103/PhysRevD.95.073002, arXiv:1612.05261.

[307] M. Cacciari, G. P. Salam and G. Soyez, *FastJet User Manual*, Eur. Phys. J. C **72**, 1896 (2012), doi:10.1140/epjc/s10052-012-1896-2, arXiv:1111.6097.

[308] J. M. Campbell, R. K. Ellis and C. Williams, *Vector boson pair production at the LHC*, JHEP **07**, 018 (2011), doi:10.1007/JHEP07(2011)018, arXiv:1105.0020.

[309] ATLAS Collaboration, *Measurements of gluon fusion and vector-boson-fusion production of the Higgs boson in $H \to WW^* \to e\nu\mu\nu$ decays using pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector* (2021), `https://cds.cern.ch/record/2759651`.

[310] CMS Collaboration, *Search for a heavy vector resonance decaying to a* Z *boson and a Higgs boson in proton-proton collisions at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **81**, 688 (2021), doi:10.1140/epjc/s10052-021-09348-6, arXiv:2102.08198.

[311] ATLAS Collaboration, *Measurements and interpretations of Higgs-boson fiducial cross sections in the diphoton decay channel using 139 $fb^{-}1$ of pp collision data at $\sqrt{s}$ = 13 TeV with the ATLAS detector* (2019).

[312] R. K. Ellis, I. Hinchliffe, M. Soldate and J. J. van der Bij, *Higgs Decay to tau+ tau-: A Possible Signature of Intermediate Mass Higgs Bosons at the SSC*, Nucl. Phys. B **297**, 221 (1988), doi:10.1016/0550-3213(88)90019-3.

[313] U. Baur and E. W. N. Glover, *Higgs Boson Production at Large Transverse Momentum in Hadronic Collisions*, Nucl. Phys. B **339**, 38 (1990), doi:10.1016/0550-3213(90)90532-I.

[314] A. Banfi, A. Martin and V. Sanz, *Probing top-partners in Higgs+jets*, JHEP **08**, 053 (2014), doi:10.1007/JHEP08(2014)053, arXiv:1308.4771.

[315] A. Azatov and A. Paul, *Probing Higgs couplings with high p_T Higgs production*, JHEP **01**, 014 (2014), doi:10.1007/JHEP01(2014)014, arXiv:1309.5273.

[316] R. V. Harlander and T. Neumann, *Probing the nature of the Higgs-gluon coupling*, Phys. Rev. D **88**, 074015 (2013), doi:10.1103/PhysRevD.88.074015, arXiv:1308.2225.

[317] C. Grojean, E. Salvioni, M. Schlaffer and A. Weiler, *Very boosted Higgs in gluon fusion*, JHEP **05**, 022 (2014), doi:10.1007/JHEP05(2014)022, arXiv:1312.3317.

[318] M. Buschmann, C. Englert, D. Goncalves, T. Plehn and M. Spannowsky, *Resolving the Higgs-Gluon Coupling with Jets*, Phys. Rev. D **90**, 013010 (2014), doi:10.1103/PhysRevD.90.013010, arXiv:1405.7651.

[319] M. Buschmann, D. Goncalves, S. Kuttimalai, M. Schonherr, F. Krauss and T. Plehn, *Mass Effects in the Higgs-Gluon Coupling: Boosted vs Off-Shell Production*, JHEP **02**, 038 (2015), doi:10.1007/JHEP02(2015)038, arXiv:1410.5806.

[320] C. Degrande, G. Durieux, F. Maltoni, K. Mimasu, E. Vryonidou and C. Zhang, *Automated one-loop computations in the standard model effective field theory*, Phys. Rev. D **103**, 096024 (2021), doi:10.1103/PhysRevD.103.096024, arXiv:2008.11743.

[321] ATLAS Collaboration, *Measurement of the properties of higgs boson production at $\sqrt{s} = 13$ tev in the $h \to \gamma\gamma$ channel using 139 fb^{-1} of pp collision data with the atlas experiment*, doi:10.48550/ARXIV.2207.00348 (2022).

[322] CMS Collaboration, *Measurements of Higgs boson production cross sections and couplings in the diphoton decay channel at $\sqrt{s} = 13$ TeV*, Journal of High Energy Physics **2021** (2021), doi:10.1007/jhep07(2021)027, `https://doi.org/10.1007%2Fjhep07%282021%29027`, arXiv:2103.06956.

[323] ATLAS Collaboration, *Search for associated production of a z boson with an invisibly decaying higgs boson or dark matter candidates at $\sqrt{s} = 13$ tev with the ATLAS detector*, Physics Letters B **829**, 137066 (2022), doi:10.1016/j.physletb.2022.137066, `https://doi.org/10.1016%2Fj.physletb.2022.137066`.

[324] CMS Collaboration, *Search for dark matter produced in association with a leptonically decaying Z boson in proton-proton collisions at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **81**, 13. 33 p (2020), doi:10.1140/epjc/s10052-020-08739-5, `https://cds.cern.ch/record/2727805`, arXiv:2008.04735.

[325] ATLAS Collaboration, *Search for invisible higgs-boson decays in events with vector-boson fusion signatures using 139 fb^{-1} of proton-proton data recorded by the atlas experiment*, doi:10.48550/ARXIV.2202.07953 (2022).

[326] CMS Collaboration, *Search for invisible decays of the higgs boson produced via vector boson fusion in proton-proton collisions at $\sqrt{s} = 13$ tev*, doi:10.48550/ARXIV.2201.11585 (2022).

[327] ATLAS Collaboration, *Measurements of higgs boson production cross-sections in the $h \to \tau^+\tau^-$ decay channel in pp collisions at $\sqrt{s} = 13$ TeV with the atlas detector*, doi:10.48550/ARXIV.2201.08269 (2022).

[328] CMS Collaboration, *Measurements of higgs boson production in the decay channel with a pair of τ leptons in proton-proton collisions at $\sqrt{s} = 13$ tev*, doi:10.48550/ARXIV.2204.12957 (2022).

[329] CMS Collaboration, *Measurements of the higgs boson production cross section and couplings in the w boson pair decay channel in proton-proton collisions at $\sqrt{s}$ = 13 tev*, doi:10.48550/ARXIV.2206.09466 (2022).

[330] ATLAS Collaboration, *Measurements of WH and ZH production in the $H \to b\bar{b}$ decay channel in pp collisions at 13 TeV with the ATLAS detector*, Eur. Phys. J. C **81**, 178. 41 p (2020), doi:10.1140/epjc/s10052-020-08677-2, `https://cds.cern.ch/record/2723187`, arXiv:2007.02873.

[331] CMS Collaboration, *Evidence for Higgs boson decay to a pair of muons*, Journal of High Energy Physics **2021** (2021), doi:10.1007/jhep01(2021)148, `https://doi.org/10.1007%2Fjhep01%282021%29148`.

[332] B. Mellado Garcia, P. Musella, M. Grazzini and R. Harlander, *CERN Report 4: Part I Standard Model Predictions* (2016), `https://cds.cern.ch/record/2150771`.

[333] LHC Higgs Working Group, *SM Higgs production cross sections at $\sqrt{s} = 13\,\text{TeV}$ (update in CERN Report 4)* (2016), `https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CERNYellowReportPageAt13TeV`.

[334] LHC Higgs Working Group, *SM Higgs Branching Ratios and Total Decay Widths (update in CERN Report 4)* (2016), `https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CERNYellowReportPageBR`.

[335] J. de Blas, J. C. Criado, M. Perez-Victoria and J. Santiago, *Effective description of general extensions of the Standard Model: the complete tree-level dictionary*, JHEP **03**, 109 (2018), doi:10.1007/JHEP03(2018)109, arXiv:1711.10391.