

Department of Physics and Astronomy

Heidelberg University

Master thesis in Physics

submitted by

Nathalie Soybelman

born in Reutlingen, Germany

2021

Learning Analytic Optimal Observables
from Symbolic Regression
for the LHC

This Master thesis has been carried out by Nathalie Soybelman

at the

Institute for Theoretical Physics

under the supervision of

Prof. Tilman Plehn

Ermitteln analytischer optimaler Observablen mittels symbolischer Regression für den LHC:

Experimente am LHC führen zu hochdimensionalen Datensätzen, wo einzelne Prozesse durch eine große Anzahl von Parametern beschrieben werden. Relevante Informationen für die Messung eines Modellparameters zu extrahieren erschwert sich dadurch und moderne Analysemethoden werden erforderlich. In dieser Situation stellen optimale Observablen - statistisch motiviert - eine mächtige Methode bereit. Sie sind per Konstruktion die Sensibelsten für die physikalischen Parameter und hängen im Allgemeinen nicht trivial von kinematischen Observablen ab. Obwohl neuronale Netze sich durch numerisches Kodieren der optimalen Observablen als effektiv erwiesen haben, ist ein interpretierbarer Ausdruck, in Form einer mathematischen Formel, erwünscht.

In dieser Arbeit zeigen wir, wie symbolische Regression benutzt werden kann, um analytische auf dem Matrix-Element basierende Formeln zu erhalten. Unsere Methode wird genutzt, um Dimension-6-Wilson-Koeffiziente des SMEFT zu studieren, die Einfluss auf Higgs-Kopplungen nehmen. Mittels des ZH -Produktionsprozesses wird unser Einsatz vorgestellt und auch erklärt. Anschließend fokussieren wir uns auf den bekannten Fall von CP-Verletzung in Vektor-Boson-Fusion, womit unser Ergebnis sowohl mit theoretischen Vorhersagen als auch mit neuronalen Netzen verglichen wird. Beim Ermitteln der Konfidenzintervalle erweist sich symbolische Regression als konkurrenzfähig gegenüber den neuronalen Netzen.

Learning Analytic Optimal Observables from Symbolic Regression for the LHC:

Experiments at the LHC give rise to high-dimensional data sets where a single event is characterized by many parameters. This makes it challenging to identify the information needed for the measurement of a model parameter and requires modern analysis techniques. Optimal observables, being well motivated by statistics, provide a powerful method for such a study. By construction they are the most sensitive observables to the physical parameters of interest while in general showing non-trivial dependencies on the kinematic observables. While neural networks proved to be an efficient technique to numerically encode them, a more interpretable form is desired as mathematical expressions remain the language of physics.

In this thesis we show how symbolic regression can be used to obtain analytic formulas using matrix-element information. We apply our method to study dimension-6 Wilson coefficients of the SMEFT affecting Higgs couplings. We use the process of associated ZH -production to introduce and explain our approach. Later we focus on the known case of CP-violation in weak-boson-fusion Higgs production comparing our result to theory predictions as well as neural networks. From studying calculated confidence intervals, we find our symbolic regression result to be competitive with the neural nets.

Contents

1	Introduction	5
1.1	The Standard Model	5
1.2	Effective field theories	8
1.3	Optimal observable and score	10
2	Tools	13
2.1	Event simulation	13
2.2	MadMiner	15
2.3	Symbolic regression	18
3	ZH production	21
3.1	Feynman rules and matrix element	21
3.2	Kinematic observables	25
3.3	Score for f_B	27
3.4	Symbolic regression on joint score	30
3.4.1	Polynomial functions for $f_B = 0$	30
3.4.2	Rational function for $f_B = 10$	33
3.5	Photon propagator	35
3.5.1	Results for $f_B = 0$	36
3.5.2	Results for $f_B = 10$	36
3.6	Two quark flavors	39
3.6.1	Results for $f_B = 0$	41
3.6.2	Results for $f_B = 10$	44
4	Weak Boson Fusion	47
4.1	Feynman rules, matrix element, CP-observable	47
4.2	Score for $f_{W\widetilde{W}}$	50
4.3	Symbolic regression at parton level	52
4.3.1	Results for $f_{W\widetilde{W}} = 0$	52
4.3.2	Results for $f_{W\widetilde{W}} = 1$	54
4.4	Detector effects	56
4.5	Exclusion limits	57
5	Summary and Outlook	60

1 Introduction

Modern LHC physics relies on state-of-the-art statistics and analysis methods to perform precision measurements for the search of new physics. Using likelihood methods to compare experimental and simulated data we can extract information on physical parameters. The data sets usually span a high-dimensional phase-space, making it challenging to extract the relevant information needed for the analysis of a certain parameter. Machine learning techniques were introduced to resolve this issue. However due to their complicated internal structure any possibility of interpreting the results in terms of mathematical expressions is lost.

The properties of the Higgs-boson, the only scalar particle in the Standard Model, are one of the big questions in modern particle physics. The theory framework for its analysis is the SMEFT (Standard Model Effective Field Theory) [1, 2]. By introducing dimension-6 operators with corresponding Wilson coefficients, the SMEFT can be used for model independent searches for new physics beyond the Standard Model. Any analysis approach boils down to the question of how to measure a certain model parameter given a certain process. The answer is provided by *optimal observables* [3–6] which are based on the matrix-element and per definition contain all relevant information for the measurement of a specific physical parameter.

While in the past, optimal observables were obtained through theory approximations at parton level, today neural networks can be used to encode the information directly at detector level [7–9]. Unfortunately, this method is not widely used as neural networks are often perceived as black boxes. Instead of using a neural network we are going to apply symbolic regression to find an interpretable expression for the optimal observable, a function of measurable kinematic quantities. We will apply this method not only on partonic processes but on data including shower and detector simulations.

We start by giving a short review of the Standard Model followed by an introduction to effective field theories and present the concept of optimal observables. After explaining the computational tools needed for our analysis in Sec. 2 we will apply our methods to learn optimal observables for Wilson coefficients of dimension-6 operators influencing the VVH -couplings. As a kickoff we look at a toy process of ZH -production in Sec. 3, then we focus on the more complicated case of weak-boson-fusion in Sec. 4. A brief summary of our results is given in Sec. 5.

1.1 The Standard Model

The Standard Model of particle physics is the most sophisticated theory for the fundamental interactions of nature. With the discovery of the Higgs boson at the LHC in 2012 the long effort of confirming the Standard Model was completed. So far it fully describes nearly all experiments performed in high energy physics.

The Standard model is a quantum field theory which is based on the following 4 aspects:

- The spontaneously broken $SU(3) \times SU(2) \times U(1)$ gauge symmetry
- The particle content consisting of 12 fermions, 4 gauge bosons and 1 scalar boson
- The 3 fundamental forces defining the electromagnetic, weak and strong interactions
- The Higgs mechanism giving rise to fermion and gauge boson masses by spontaneously breaking the $SU(2) \times U(1)$ symmetry

Lagrangian

Given this foundation one can construct a Lagrangian \mathcal{L}_{SM} that includes all possible terms that can be constructed from the particle fields as long as they are invariant under the imposed gauge symmetry as well as the Poincaré group consisting of Lorentz transformations and translations. Additionally we require the Standard Model to be a renormalizable theory, therefore only terms with dimensions up to 4 are included. We can break up the Lagrangian into 4 terms:

$$\mathcal{L}_{\text{SM}} = \mathcal{L}_{\text{kin}} + \mathcal{L}_{\psi} + \mathcal{L}_{\phi} + \mathcal{L}_{\text{Yuk}} \quad (1)$$

The first part is the kinetic term containing derivatives of the fields. It describes their propagation through spacetime and ensures that the fields are dynamical. It is given by:

$$\mathcal{L}_{\text{kin}} = -\frac{1}{4}G_a^{\mu\nu}G_{a\mu\nu} - \frac{1}{4}W_b^{\mu\nu}W_{b\mu\nu} - \frac{1}{4}B^{\mu\nu}B_{\mu\nu} - \sum_{f \in \{Q,L\}} i\bar{\psi}_f \not{D}\psi_f - (D^\mu\phi)^\dagger D_\mu\phi \quad (2)$$

We use ψ for the fermionic fields of all quarks and leptons and ϕ for the scalar field corresponding to the Higgs field. G are the gluon fields and W and B are the gauge bosons of the electroweak theory. The field strength tensors used above are defined as follows:

$$\begin{aligned} G_a^{\mu\nu} &= \partial^\mu G_a^\nu - \partial^\nu G_a^\mu - g_s f_{abc} G_b^\mu G_c^\nu \\ W_a^{\mu\nu} &= \partial^\mu W_a^\nu - \partial^\nu W_a^\mu - g \epsilon_{abc} W_b^\mu W_c^\nu \\ B^{\mu\nu} &= \partial^\mu B^\nu - \partial^\nu B^\mu \end{aligned} \quad (3)$$

where g_s is the strong and g the weak coupling constant and $f_{abc}(\epsilon_{abc})$ are the structure constants of the $SU(3)$ ($SU(2)$) group. For the $U(1)$ group the last term vanishes as it is an Abelian symmetry group.

We also introduce the covariant derivative which ensures that the Lagrangian is invariant under gauge transformations:

$$D^\mu = \partial^\mu + i\frac{g_s}{2}G_a^\mu\lambda_a + i\frac{g}{2}W_b^\mu\sigma_b + ig'YB^\mu \quad (4)$$

where λ_a are the Gell-Mann matrices, σ_b the Pauli matrices and Y the hypercharge. They are the generators of their corresponding symmetry group. Applying the covariant derivative for example on the Higgs field we obtain:

$$D^\mu\phi = \partial^\mu\phi + i\frac{g}{2}\sigma_b W_b^\mu\phi + i\frac{g'}{2}B^\mu\phi \quad (5)$$

The next term in the Lagrangian is $\mathcal{L}_{\psi} = 0$. Since fermions are charged and chiral we cannot construct mass terms that are invariant under symmetry transformations. We therefore move on to \mathcal{L}_{ϕ} , the Higgs potential. It gives rise to the Higgs mass and self-couplings. It is given by:

$$\mathcal{L}_{\phi} = -\mu^2\phi^\dagger\phi - \lambda(\phi^\dagger\phi)^2 \quad (6)$$

Last but not least the Yukawa term \mathcal{L}_{Yuk} accounts for the interactions between fermions and Higgs:

$$\mathcal{L}_{\text{Yuk}} = \sum_{f \in \{Q,L\}} Y_f \bar{\psi}_L \phi \psi_{f,R} + \text{h.c.} \quad (7)$$

Here Y_f are the 3×3 Yukawa matrices.

Higgs mechanism

The Higgs potential as defined in Eq. (6) leads to a spontaneous symmetry breaking of $SU(2) \times U(1)_Y \rightarrow U(1)_Q$. Taking $\mu^2 < 0$, the minimum of the potential is given by:

$$\langle \phi \rangle = \sqrt{-\frac{\mu^2}{2\lambda}} \equiv \frac{v}{\sqrt{2}} \quad (8)$$

This means the Higgs field acquires a non-zero vacuum expectation value (VEV) which is not invariant under the gauge transformations and therefore breaks 3 out of 4 generators of the $SU(2) \times U(1)_Y$ group.

We can view the Higgs field as the VEV plus an excitation h . From now on we refer to the Higgs field given by:

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h \end{pmatrix} \quad (9)$$

In the next step we can insert this expression for the Higgs field into the Lagrangian defined above. From the Yukawa part given in Eq. (7) we acquire fermion masses given by $Y_f v / \sqrt{2}$ as well as interactions with the Higgs boson. More relevant for the course of this work are the interactions with the weak gauge bosons as well as their masses. They arise from the kinetic term of the Higgs field given in Eq. (2). Inserting the covariant derivative given in Eq. (5) the Lagrangian for the gauge boson masses is given by:

$$\begin{aligned} \mathcal{L}_{M_V} &= \frac{1}{8} \begin{pmatrix} 0 & v \end{pmatrix} \begin{pmatrix} gW_{3\mu} + g'B_\mu & g(W_1 - iW_2)_\mu \\ g(W_1 + iW_2)_\mu & -gW_{3\mu} + g'B_\mu \end{pmatrix} \begin{pmatrix} gW_3^\mu + g'B^\mu & g(W_1 - iW_2)^\mu \\ g(W_1 + iW_2)^\mu & -gW_3^\mu + g'B^\mu \end{pmatrix} \begin{pmatrix} 0 \\ v \end{pmatrix} \\ &= \frac{v^2}{8} \left(g^2 W_1^2 + g^2 W_2^2 + (-gW_3 + g'B)^2 \right) \end{aligned} \quad (10)$$

We see that in the last term we have mixing between W_3 and B meaning their mass matrix is not diagonal. To resolve this issue we redefine the fields as follows:

$$Z_\mu = \frac{1}{\sqrt{g^2 + g'^2}} (-g'B_\mu + gW_{3\mu}) \quad A_\mu = \frac{1}{\sqrt{g^2 + g'^2}} (gB_\mu + g'W_{3\mu}) \quad (11)$$

The mixing angle is the Weinberg angle θ_W defined as:

$$\tan \theta_W \equiv \frac{g'}{g} \quad (12)$$

Even though W_1 and W_2 are already in a diagonal mass basis it is worthwhile to redefine them as well to insure a well defined electromagnetic charge:

$$W_\mu^\pm = \frac{1}{\sqrt{2}} (W_1 \mp iW_2) \quad (13)$$

Inserting Eq. (11) and (13) into Eq. (10) we obtain the masses of the weak gauge bosons W^\pm and Z^0 given by:

$$m_W = \frac{1}{4} v^2 g^2 \quad m_Z = \frac{1}{4} v^2 (g^2 + g'^2) \quad (14)$$

while the photon A remains massless. It corresponds to the unbroken $U(1)_Q$ symmetry.

In analogy to the calculation in Eq. (10) we can also obtain the interactions of the weak bosons with the Higgs. They are then given by:

$$\mathcal{L}_{V-h} = \frac{m_W^2}{v^2} W_\mu^- W^{+\mu} (2hv + h^2) + \frac{1}{2} \frac{m_Z^2}{v^2} Z_\mu Z^\mu (2hv + h^2) \quad (15)$$

This gives us the Feynman rules for example for the ZZH vertex, given by $2im_Z^2/v$, or the WWH vertex given by $4im_W^2/v$. We will use these later for the matrix element calculations in Sec. 3.1 and 4.1.

1.2 Effective field theories

While the Standard Model describes nearly all experiments so far performed in high energy physics there is a large amount of evidence for it not being a full theory of nature. For instance the Standard Model does not include dark matter or neutrino masses and does not explain the baryon asymmetry in the early universe. Additionally it suffers from fine-tuning issues like the hierarchy problem which requires a precise cancellation between the quantum corrections to the Higgs mass and the bare mass in order for the Higgs mass to be at the measured value. These and many other mysteries remain unanswered by the Standard Model.

Unfortunately, regardless of the evidence, the LHC did not discover any new physics resolving these issues. A reason for that can be that the new physics lies at energy scales we cannot reach with our experiments. Instead, we have to search for indirect signs of new physics manifesting in deviations from the Standard Model. A discovered deviation needs to be interpreted within a theoretical framework in order to understand what the new physics is. Taking a fully constructed model for such an analysis will, however, be highly inefficient given the amount of models existing in theory today as well as their large parameter spaces. Instead, a framework for model independent analysis will prove to be far more useful. This is where the idea of effective field theories becomes important. They can serve as a bridge between theory models and data analysis by matching the full theory to an EFT and perform the data analysis in the EFT framework.

The EFT approach is a fundamental concept to physics. The underlying idea is the separation of energy scales, meaning that at different energy scales we have different physics to consider. For example we do not have to think about quantum gravity when we want to calculate the speed of a falling apple hitting the ground, and when throwing a ball against a wall we rarely have to worry about it tunneling through. The physics at a high energy scale is described by an underlying theory which we often do not know, and do not need, to understand the physics at lower energies that are of interest to us. Instead of the underlying theory we can take an effective theory that includes the relevant effects for the energy scale of interest, but not those which are only relevant at higher energies. Naturally this approximation is only valid as long as the considered energies E are smaller than a certain energy scale Λ :

$$E \ll \Lambda \quad (16)$$

A classic example is the Fermi theory where the W -propagator is approximated by the W -mass resulting in an effective 4 fermion interaction:

$$\frac{1}{q^2 - M_W^2} \longrightarrow \frac{1}{M_W^2}, \quad q^2 \ll M_W^2 \quad (17)$$

Historically this approximation was used before knowing that it is an approximation, i.e. it was used without the knowledge of the full theory. Detailed explanations and more examples can be found in [10].

$\mathcal{O}_{BB} = -\frac{g'^2}{4}(\phi^\dagger\phi) B_{\mu\nu}B^{\mu\nu}$	$\mathcal{O}_{B\tilde{B}} = -\frac{g'^2}{4}(\phi^\dagger\phi) B_{\mu\nu}\tilde{B}^{\mu\nu}$
$\mathcal{O}_{WW} = -\frac{g^2}{4}(\phi^\dagger\phi) W_{\mu\nu}^a W^{\mu\nu a}$	$\mathcal{O}_{W\tilde{W}} = -\frac{g^2}{4}(\phi^\dagger\phi) W_{\mu\nu}^a \tilde{W}^{\mu\nu a}$
$\mathcal{O}_{BW} = -\frac{gg'}{4}(\phi^\dagger\sigma^a\phi) B_{\mu\nu}W^{\mu\nu a}$	$\mathcal{O}_{B\tilde{W}} = -\frac{gg'}{4}(\phi^\dagger\sigma^a\phi) B_{\mu\nu}\tilde{W}^{\mu\nu a}$
$\mathcal{O}_{GG} = (\phi^\dagger\phi) G_{\mu\nu}^a G^{\mu\nu a}$	$\mathcal{O}_{G\tilde{G}} = (\phi^\dagger\phi) G_{\mu\nu}^a \tilde{G}^{\mu\nu a}$
$\mathcal{O}_B = \frac{ig'}{2}(D^\mu\phi)^\dagger D^\nu\phi B_{\mu\nu}$	$\mathcal{O}_{\tilde{B}} = \frac{ig'}{2}(D^\mu\phi)^\dagger D^\nu\phi \tilde{B}_{\mu\nu}$
$\mathcal{O}_W = \frac{ig}{2}(D^\mu\phi)^\dagger \sigma^a D^\nu\phi W_{\mu\nu}^a$	

Table 1: Dimension 6 operators for Higgs-gauge boson interactions. Left column: CP-conserving operators. Right column: CP-violating operators

The central idea is that we can do the same thing in the context of the Standard Model: We can consider it to be a low energy effective theory of an underlying and unknown theory which is valid for energies far beyond the reach of modern and possibly future colliders. This theory framework is called SMEFT [1, 2] and can be used for analysis of possible deviations from the SM. In the previous section we discussed how the SM is constructed by renormalizable dimension 4 operators. Since it is only valid up to a certain energy scale, we must include higher order terms. To have the correct mass dimensions, these terms need to be suppressed by the high energy scale $1/\Lambda^d$. We usually consider $\Lambda = 1$ TeV. The Lagrangian can be written in the following way:

$$\mathcal{L}_{\text{EFT}} = \mathcal{L}_{\text{SM}} + \sum_i \frac{f_i}{\Lambda^{D_i-4}} \mathcal{O}_i \quad (18)$$

The operators \mathcal{O}_i consist of the SM fields and are invariant under the same symmetries as the SM. Each operator has a corresponding dimensionless coupling constant called *Wilson coefficient* f_i . Precise measurements of these Wilson coefficients are a way of quantifying the deviations from the SM and the possibility of new physics.

While dimension-5 operators are related to neutrino physics, for Higgs physics we primarily focus on dimension-6 operators. Dimension-8 operators are considered too suppressed to make an impact and are excluded from studies in most cases. Listing and discussing all possible dimension-6 operators is beyond the scope of this work. Instead we focus on the operators relevant for Higgs-gauge boson interactions. A list of these operators is given in Tab. 1 and is taken from [11]. $\tilde{V}_{\mu\nu}$ for $V = B, W, G$ are dual tensors given by:

$$\tilde{V}_{\mu\nu} = \epsilon_{\mu\nu\rho\sigma} V^{\rho\sigma} \quad (19)$$

The presence of the Levi-Civita tensor $\epsilon_{\mu\nu\rho\sigma}$ violates CP. Therefore operators with a dual tensor are CP-violating. In the electroweak theory of the SM operators $V_{\mu\nu}\tilde{V}^{\mu\nu}$ are not present in the Lagrangian since they can be written as total derivatives and therefore do not have any effect on

the physics¹. Hence, CP-violation tests in the gauge sector are very sensitive to new physics. We will discuss such a CP-violating dimension-6 term in Sec. 4.1.

1.3 Optimal observable and score

So far we established that measuring Wilson coefficients in the EFT framework is useful to quantify deviations from the SM. Now we devote ourselves to the question of *how* to measure these parameters in the most proficient way.

Optimal observable

Optimal observables are an efficient method for the analysis of a certain physical parameter. For simple processes this can be an easy to interpret observable like a jet p_T that can be obtained through an educated guess. However, for high-dimensional phase-spaces the optimal observable can be a complicated and non-trivial function of the kinematic observables, which means that a more sophisticated way is needed to determine it.

We follow the derivations of [7, 12] to argue that for any parameter given a certain process, there is a single, best-suited observable for extracting the relevant information from a data set. We start with the likelihood function for a single event given by:

$$p(x|\theta) = \frac{1}{\sigma(\theta)} \frac{d\sigma(x|\theta)}{dx} \quad (20)$$

where $\sigma(\theta)$ is the total cross section as a function of the model parameter and $d\sigma(x|\theta)$ is the differential cross section characterized by a vector of observables x that consists of reconstructed 4-momenta of the final state particles. We make the approximation that the parameter θ is close to our reference point, usually the SM value, $|\theta - \theta_{\text{ref}}| \ll 1$. This approximation is not limiting us since the SMEFT parameters are supposed to be very small. Otherwise, large deviations from the SM would have already been discovered. This allows us to perform a Taylor expansion of the log-likelihood around the reference point:

$$\log p(x|\theta) = \log p(x|\theta_{\text{ref}}) + \nabla_{\theta} \log p(x|\theta) \Big|_{\theta=\theta_{\text{ref}}} |\theta - \theta_{\text{ref}}| + \mathcal{O}(|\theta - \theta_{\text{ref}}|^2) \quad (21)$$

Or equivalently:

$$p(x|\theta) \approx p(x|\theta_{\text{ref}}) \exp\left(t(x|\theta_{\text{ref}}) |\theta - \theta_{\text{ref}}|\right) \quad (22)$$

where we define the quantity we call *score* as:

$$t(x|\theta) = \nabla_{\theta} \log p(x|\theta) \quad (23)$$

Such a probability distribution is a special case of the *exponential family*, a well-studied family of probability distributions. To be precise, $t(x|\theta)$ represents the sufficient statistics of the probability distribution: it was shown that for the purpose of inferring θ , measuring $t(x|\theta)$ is exactly as useful as the full observation of x [13]. Therefore the score is precisely our optimal observable.

¹Such a term can be written for the gluon fields as they have non-vanishing boundary conditions. This is related to the strong CP-problem.

The likelihood function

Now that we saw that there is an optimal observable, we need to examine it more closely. We start by focusing on the likelihood function it is based on. Unfortunately, it is impossible to calculate the likelihood function due to the complicated Monte-Carlo simulation chain needed for event generation described in detail in Sec. 2.1. The outcome of an event depends on many parameters that for example come from the shower or the detector response and cannot be measured. We call these parameters latent variables z . To obtain the likelihood function we need to integrate them out which is not possible due to their large abundance. Hence, the likelihood function is *intractable*. Symbolically it can be written as follows:

$$p(x|\theta) = \int dz_d \int dz_s \int dz_p \underbrace{p(x|z_d)p(z_d|z_s)p(z_s|z_p)p(z_p|\theta)}_{p(x,z|\theta)} \quad (24)$$

Here z_p are the parton level variables which include 4-momenta, flavor, charge etc., z_s includes the entire shower information and z_d describes the interactions with the detector. We call $p(x, z|\theta)$ the *joint likelihood*. From the joint likelihood we can similarly define the *joint score*:

$$t(x, z|\theta) \equiv \nabla_\theta \log p(x, z|\theta) = \frac{p(x|z_d)p(z_d|z_s)p(z_s|z_p)\nabla_\theta p(z_p|\theta)}{p(x|z_d)p(z_d|z_s)p(z_s|z_p)p(z_p|\theta)} = \nabla_\theta \log p(z_p|\theta) \quad (25)$$

We see that for the joint score the θ -independent, intractable likelihoods cancel out, which leaves us only with the parton level variables z_p . Inserting Eq. (20) we can write the joint score in terms of the cross section:

$$\begin{aligned} t(x, z|\theta) &= \frac{\nabla_\theta p(z_p|\theta)}{p(z_p|\theta)} = \left(\frac{\sigma(\theta)\nabla_\theta d\sigma(z_p|\theta) - d\sigma(z_p|\theta)\nabla_\theta \sigma(\theta)}{\sigma^2(\theta)} \right) \left(\frac{\sigma(\theta)}{d\sigma(z_p|\theta)} \right) \\ &= \frac{\nabla_\theta d\sigma(z_p|\theta)}{d\sigma(z_p|\theta)} - \frac{\nabla_\theta \sigma(\theta)}{\sigma(\theta)} \end{aligned} \quad (26)$$

The differential cross section is given by:

$$d\sigma(z_p|\theta) = \frac{(2\pi)^4 f_1(x_1, Q^2) f_2(x_2, Q^2)}{8x_1 x_2 s} |\mathcal{M}|^2(z_p|\theta) d\Phi(z) \quad (27)$$

with the momentum fractions x_i of the partons, the squared center-of-mass energy s , the parton density functions (PDFs) $f_i(x_i, Q^2)$ with the momentum transfer Q^2 , the phase-space element $d\Phi(z_p)$ and the squared matrix element $|\mathcal{M}|^2$. Inserting Eq. (27) into Eq. (26) many terms cancel and we are left with:

$$t(x, z|\theta) = \frac{\nabla_\theta |\mathcal{M}|^2(z_p|\theta)}{|\mathcal{M}|^2(z_p|\theta)} - \frac{\nabla_\theta \sigma(\theta)}{\sigma(\theta)} \quad (28)$$

From simulated data we can extract the information on z_p to obtain the squared matrix-element and hence, the joint score can be calculated analytically.

Learning the score

Summarizing the previous discussion, we clarified that on the one hand we can define the optimal observable or score needed to extract information on a model parameter, but it is intractable. On

the other hand we have the joint score which we can calculate from parton level kinematics, but is not a function of the observable quantities from the detector.

Luckily, these two quantities are related and we show in the next step that the joint score can be used to estimate the score. We follow the derivation in [7] and start by assuming that there is a function $\hat{g}(x)$ that can approximate a function $g(x, z)$. We choose the L^2 -norm and get the following loss functional:

$$\begin{aligned} L[\hat{g}(x)] &= \int dx dz p(x, z|\theta) |g(x, z) - \hat{g}(x)|^2 \\ &= \int dx \underbrace{\left[\hat{g}^2(x) \int dz p(x, z|\theta) - 2\hat{g}(x) \int dz p(x, z|\theta)g(x, z) + \int dz p(x, z|\theta)g^2(x, z|\theta) \right]}_{F(x)} \end{aligned} \quad (29)$$

The desired choice for $\hat{g}(x)$ has to minimize the loss or equivalently $F(x)$. Applying calculus of variation yields:

$$0 = \frac{\delta F}{\delta \hat{g}} = 2\hat{g}(x) \underbrace{\int dz p(x, z|\theta)}_{p(x|\theta)} - 2 \int dz p(x, z|\theta)g(x, z) \quad (30)$$

Therefore the function $\hat{g}(x)$ that extremizes the loss is given by:

$$\hat{g}(x) = \frac{1}{p(x|\theta)} \int dz p(x, z|\theta)g(x, z) \quad (31)$$

Now we replace $g(x, z)$ with the joint score from Eq. 25 and obtain:

$$\hat{g}(x) = \frac{1}{p(x|\theta)} \int dz p(x, z|\theta) \frac{\nabla_{\theta} p(x, z|\theta)}{p(x, z|\theta)} = \frac{1}{p(x|\theta)} \int dz \nabla_{\theta} p(x, z|\theta) = \frac{\nabla_{\theta} p(x|\theta)}{p(x|\theta)} = t(x|\theta) \quad (32)$$

where in the last step we simply interchanged integral and derivative. We therefore showed that indeed the joint score can be used to estimate the optimal observable. Such a function can be learned through a neural network or a symbolic regression fit by minimizing the squared loss which for a sample of N events is given by

$$L[t(x|\theta)] = \frac{1}{N} \sum_{i=1}^N |t(x, z|\theta) - t(x|\theta)|^2 \quad (33)$$

In the following we will refer to it as the *mean squared error* (MSE).

2 Tools

In this section we introduce the relevant tools we need for our analysis. The general procedure is to generate events, compute the optimal observable introduced in Sec. 1.3 and apply symbolic regression on it to obtain an analytic expression for it. For each of these steps we need a different computational tool that will be explained in the following.

2.1 Event simulation

We are interested in simulated data of LHC collisions for our analysis. The event generation is performed through a simulation chain consisting of three parts: the parton level process depending on the underlying physical model of interest, the shower and the detector response. In many cases we perform our analysis already after the first step to obtain a better understanding of the physics. To be able to compare results to real experimental data it is however important to make sure our analysis tools work also on realistic data.

Parton level event generation

We use MADGRAPH5 [14, 15], a general purpose matrix-based event generator, to generate the data we need for our analysis. First of all we provide MADGRAPH5 with the relevant model. The model is essentially the Lagrangian of the theory with the corresponding Feynman rules including all physical parameters. The required UFO (Universal FeynRules Output) format of the model can be obtained by FEYNRULES. Further, we specify the process of interest for which MADGRAPH5 will automatically generate the corresponding Feynman diagrams as well as the computer code needed to evaluate the matrix element at a given phase-space point.

For the generation of the Feynman diagrams, at first, all possible combinations of the external particles are considered and if vertices exist that can connect the particles for a given combination the diagrams are saved in the output. More details on this algorithm as well as an intuitive example can be found in [15].

The squared matrix element is evaluated by the ALOHA [16] package. It uses helicity amplitudes methods which are faster and less involved than trace techniques. For the generation of amplitudes first a set of wavefunctions for the external particles is initialized using their helicities and momenta. Depending on the particle interactions of the corresponding Feynman diagram, the wavefunctions of the propagators are obtained. Taking all the obtained wavefunctions the amplitude for the diagram can be computed. The amplitudes from all diagrams can then be summed and squared yielding the final result.

The total cross section is obtained through Monte-Carlo integrations of the squared matrix element as the integrals can not be calculated analytically. Therefore a phase-space generator generates phase-space points according to some distribution function which is not the true distribution. The weight of the event corresponds to its contribution to the real distribution. For example, if the amount of generated events is twice as large as it would be according to the true distribution their event weights would be 0.5. Hence we obtain a weighted sample that is used for the integration. The output of MADGRAPH5 is however a set of unweighted events. To obtain unweighted events from a weighted sample the standard rejection/acceptance method is applied. For each event a random number is thrown and the event is kept if its weight is below the random number. The resulting events are distributed according to the true distribution and have equal weights.

An important feature of MADGRAPH5 that will be of use to us for the calculation of the optimal observable is the reweighting option. As already mentioned above, for the event generation we provide MADGRAPH5 with our theory model which consists of the Feynman rules and all physical parameters including the Wilson coefficients of interest. Given the definition of the optimal observable in Eq. (23) we are interested in the derivative of the squared matrix element with respect to the parameter, i.e. we want to see how the change of the parameter influences the result. Therefore we can additionally provide MADGRAPH5 with different benchmarks for a certain Wilson coefficient, meaning other numerical values for the parameter of interest than given in the model files. After the event generation for the original parameter values the reweighting process will calculate a new weight for each benchmark value for every generated event. The new weight is given by:

$$w_{\text{new}} = \frac{|\mathcal{M}|_{\text{new}}^2}{|\mathcal{M}|_{\text{old}}^2} w_{\text{old}} \quad (34)$$

where the old weight is the weight after the unweighting procedure which was the same for all events. These weights will be later used by MADMINER in Sec. 2.2.

Shower and hadronisation

From the event generation above we get the 4-momenta of the incoming and outgoing particles of the defined process. But the momenta of the outgoing particles are not what we are able to directly measure at the LHC or any other collider experiment. This is due to the *shower* induced by the outgoing partons and the subsequent *hadronisation* process. The simulation of this development can be performed by PYTHIA8 [17], a general purpose shower and hadronisation generator, that can be implemented into MADGRAPH5 and automatically follows the event generation.

What happens after the collision is that free partons (quarks and gluons) coming from the hard process, described by the Feynman diagram, split into two. For example a quark can radiate of a gluon, a gluon can split into a quark anti-quark pair or a pair of gluons. The resulting partons keep splitting as long as their energy is sufficiently large, forming a shower. The shower evolution is based on the DGLAP splitting kernels $\mathcal{P}(z)$ which are derived in [18]. For a single parton they give the probability that it will split into 2 partons with energy fraction z .

Once reaching the low energy non-perturbative regime the hadronisation process starts. Here the previously individual partons group together forming hadrons. The simulated hadronisation is based on the Lund string framework [19]. The basic principle behind is color confinement, meaning that there cannot exist free quarks, only color neutral objects. Taking for example a quark and an anti-quark that would form together a color neutral state, the color confinement can be modeled with a color flux tube connecting the two partons. As the distance between the partons grows, the force between them grows linearly and the energy stored in the tube increases. At some point the tube breaks producing a new quark anti-quark pair. The remaining two string pieces can break further if their energy is large enough until only on-shell hadrons can be produced.

In the final step, unstable hadrons decay into stable particles mostly being pions. In this step the decay of other particles like Z , W or Higgs is also implemented according to the available decay channels.

In the output file the full information on the shower evolution is saved. For each particle its 4-momenta, mother and daughter particles, status (at which point the particle was produced), color and particle ID is recorded.

Detector response

The process described above happens already within the beam pipe so that only the final stable constituents of the showers enter the detector system. The full modeling of particle propagation in the detector, calorimeter response and particle reconstruction can be a very challenging and time consuming task. Full detector event simulations including interactions with detector material can take up to several minutes per event. For phenomenological studies like ours it is sufficient to use simplified detector simulation software like DELPHES [20] which adds smearing to the final state object kinematics to mimic realistic detector smearing effects.

DELPHES provides a simulation framework for a multipurpose detector consisting of an inner tracking system, electromagnetic and hadronic calorimeters, and a muon system. The process consists of detector response simulation and object reconstruction. In the first part, the final state particles from PYTHIA8 propagate through the magnetic field of the inner tracker. The tracks that are left by charged particles can be used to reconstruct their momenta. During propagation a perfect angular resolution is assumed while on the norm of the transverse momentum vector (orthogonal to the beam pipe) a smearing is applied. Tracking efficiency, energy and momentum resolutions can be specified. After passing the tracking system particles deposit energy in the calorimeter cells. The resolution is determined by the granularity of the cells. With a simplified particle-flow algorithm, tracks of charged particles are matched with their corresponding cells to separate the cell signal for charged and neutral particles. Finally the objects are reconstructed using the cell and track information. Most importantly jet reconstruction is implemented through the FASTJET package. The algorithm aims to combine all shower remnants originating from a parton to recover its original 4-momenta. The reconstructed particles and their 4-momenta are saved in the output and can be analyzed.

2.2 MadMiner

In this section we are going to introduce MADMINER [12], a Python package for machine learning-based inference techniques. It includes a large set of functions documented in [21] that can be used for typical analysis in high energy physics, out of which we will examine the ones relevant for the course of this work.

Obtaining the joint score

Starting an analysis with MADMINER we first choose a physical process and a set of elementary kinematic observables that summarize the observed event. For the parameters we declare different benchmarks for which MADGRAPH5 performs the reweighting from Eq. (34). MADMINER generates an automatic script that calls MADGRAPH5 and performs the event generation discussed in Sec. 2.1. In the next step we extract the relevant information from the output files of MADGRAPH5 (or DELPHES if we are interested in the full process) by defining the observables of interest, e.g. p_T of a jet or $\Delta\phi$ of 2 particles, and implementing cuts if needed. Events that satisfy the requirements induced by the cuts are selected and the specified kinematic observables are computed from the 4-momenta.

Finally we have to calculate the joint score for each event. Taking for example one EFT operator as introduced in Sec. 1.2, the matrix element of a certain process is given by the SM contribution and a new physics term proportional to the corresponding Wilson coefficient. Schematically the

squared matrix-element can therefore be written as:

$$|\mathcal{M}|^2 = p_0 + a\theta + b\theta^2 \quad (35)$$

In this notation p_0 is the contribution from the SM, b corresponds to the squared new physics term and a is the interference between both. We take the first term of the joint score in Eq. (28) and insert Eq. (35):

$$\frac{\nabla_{\theta} d\sigma(z_p|\theta)}{d\sigma(z_p|\theta)} = \frac{\nabla_{\theta} |\mathcal{M}|^2(z_p|\theta)}{|\mathcal{M}|^2(z_p|\theta)} = \frac{a + 2b\theta}{p_0 + a\theta + b\theta^2} \quad (36)$$

To obtain the joint score we thus need to extract the parameters p_0 , a and b . Analytically we will do this by hand in Sec. 3.1 when calculating the squared matrix element. Within MADMINER a so called *morphing* technique [7, 22] is used where the joint score is calculated event by event. For a given phase-space point p_0 , a and b are essentially constants and the squared matrix element is a parabola in θ . Now we make use of the weights from the reweighting procedure. The weights are essentially equivalent to the squared matrix element apart from normalization which cancels out. The constants can be extracted by simply solving a linear system of equations. Having 3 unknown we need the weights for 3 benchmark values². The set of benchmarks is called *morphing basis*.

For the second term of the joint score in Eq. (26) we have to perform the phase-space integration to obtain the total cross section. We can interchange derivation and integration and write:

$$\frac{\nabla_{\theta} \sigma(\theta)}{\sigma(\theta)} = \frac{\int \frac{1}{x_1 x_2} f_1(x_1, Q^2) f_2(x_2, Q^2) (a + 2b\theta) d\Phi(z_p)}{\int \frac{1}{x_1 x_2} f_1(x_1, Q^2) f_2(x_2, Q^2) (p_0 + a\theta + b\theta^2) d\Phi(z_p)} \quad (37)$$

Since the PDFs depend on the phase-space there is no way of further simplifying this expression. It is not possible for us to calculate this term analytically, yet MADMINER can use the the same morphing technique as before. Instead of applying it on the weights, it is applied on the total cross-section from MADGRAPH5 Monte-Carlo integration. For fixed θ it is a constant that does not depend on phase-space and needs to be calculated just once.

Putting everything together, MADMINER saves the event observables and the joint score as NUMPY [23] arrays.

Machine learning

As discussed in Sec. 1.3, the joint score can be used to estimate the optimal observable. One way to do this, is to apply machine learning techniques and obtain a neural net for the score. This can be done by the SALLY (Score Approximates Likelihood Locally) estimator implemented in MADMINER. SALLY was proven to be more effective in extracting information on parameters than conventional histogram methods [7–9].

The architecture is a fully connected neural net. It is possible to choose the observables we want to train on. During training the loss as defined in Eq. (33) is minimized by the AMSGRAD [24] optimizer.

The goal of this work is to replace the SALLY algorithm with a method that approximates the score with an analytic function. Such a function is easier to interpret than a neural net. However, we will use SALLY as a benchmark to compare its performance to our method in Sec. 4.5.

²If we have more than one parameter we would have to specify more benchmarks.

Setting limits

The objective of any analysis in high energy physics is to obtain the famous confidence intervals on the parameters of interest. Especially in the era of beyond SM searches precision measurements are required to narrow the error bars on the SM parameters so that possible deviations from the SM would lie outside the uncertainty region and can be detected.

We discussed in the previous section how to obtain the optimal observable through a neural net. We established in Sec. 1.3 that it is per definition the best observable for the parameter measurement but we do not know yet how well it can measure said parameter, i.e. what are its uncertainties. This is essential for the comparison of the goodness of different observables. The smaller the uncertainty on the parameter, the better the observable. The statistical methods [25] needed for this analysis are implemented in MADMINER and are explained in the following.

For the analysis we consider two hypothesis of which the first is the SM hypothesis H_0 with $\theta = 0$ and the second is the new physics hypothesis H_1 with a finite θ . According to the Neyman-Pearson lemma the best way of discriminating two hypothesis is to consider their log-likelihood ratio. From there we want to calculate p -values for different θ , i.e. the probability of observing data at least as extreme as predicted by H_0 under assumption of H_0 . A small p -value would indicate that it is unlikely that H_0 is correct. The p -values can be translated into confidence regions which means that essentially we want to find the θ that sets the exclusion region at 68 (95) % confidence level (CL).

The first step would be to generate test data for many different θ . Luckily we do not have to simulate new events a hundred times but can sample from a single simulation. Therefore MADMINER uses again the morphing setup discussed previously. For a given event it had already obtained the constants p_0 , a and b and thus can calculate a weight for any parameter θ . Afterwards, the events are sampled with the usual acceptance/rejection method according to the new weights. We call the new samples the test data as opposed to the observed³ data.

For each of the new test samples we calculate the expected log-likelihood of observing the test data under the hypothesis H_0 . This is not the full log-likelihood of the event but is rather based on histograms of a single given observable which in our case will be the score. Naturally, the true value of θ maximizes the likelihood. The *test statistic* q for any θ is given by the ratio of the log-likelihood of θ and the best log-likelihood:

$$q(\theta) = -2 \log \frac{p(x|\theta)}{p(x|\theta_{\text{best}})} \quad (38)$$

The larger $q(\theta)$, the less compatible is θ with the observed data. From the test statistic we can directly obtain the corresponding p -value:

$$p(\theta) = \int_{q(\theta_{\text{obs}})}^{\infty} f(q, k) dq \quad (39)$$

With k the degrees of freedom (amount of parameters) and $f(q, k)$ the probability density function for χ^2 given by

$$f(q, k) = \frac{1}{2^{k/2} \Gamma(k/2)} q^{k/2-1} \exp(-q/2) \quad (40)$$

which is implemented through the SCIPY [26] package.

³In a real analysis this would be real data from experiments. In our case it comes from simulations as well.

2.3 Symbolic regression

In this section we introduce PYSR [27], a julia based python package for symbolic regression, that we will use to obtain an analytic function for the optimal observable. This task is the central aspect of this work and it is therefore important to understand the algorithm in detail to be able to adapt it if needed.

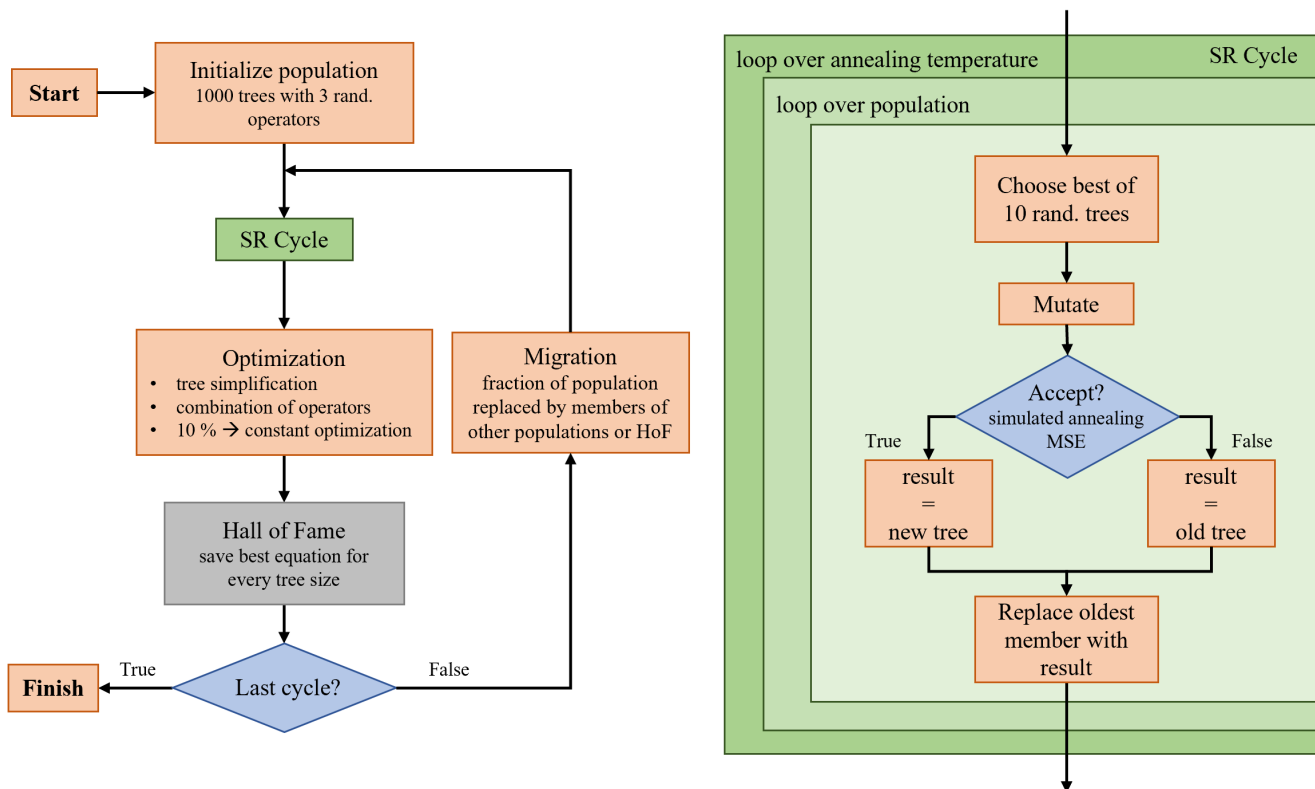


Figure 1: Diagrammatic representation of the PYSR algorithm.

For the given optimization task PYSR uses genetic programming to find a symbolic expression for the provided data. The basic idea is to start with a large set of random unfit configurations and to improve them by applying different kinds of operations, similar to natural genetic processes.

The algorithm uses tree diagrams for the construction of symbolic expressions. Such a tree consists of nodes that represent either an operator or an operand. Each operator node can connect to one node for unary operators like \sin or \exp or two nodes for binary operators like $+$ or $-$. The operands can either be variables from the data or constants from the algorithm. We provide PYSR with a list of operators of our choice to be used in the tree construction.

A diagrammatic description of the algorithm is depicted in Fig. 1. We start by initializing a set of populations. The hyperparameter `populations` sets the amount of populations and is per default set to the amount of processors (`procs`) used. The amount of individuals per populations is given by `npop`. Each individual is a formula represented by a tree. During initialization the trees are constructed through 3 random operators from the given operator list. The populations at first evolve separately where new individuals are created and old once discarded. They advance through many generations set by the hyperparameter `ncyclesperiterations`.

Starting with the current population a subsample of 10 trees is taken and the best one is chosen

according to its `score`⁴ given by:

$$\text{score} = \frac{\text{MSE}}{\text{baseline}} + \text{complexity} \cdot \text{parsimony} \quad (41)$$

where the MSE is the one given in Eq. (33), the baseline is the MSE between the data and the constant unity function and the complexity of a tree is given by the amount of nodes:

$$\text{complexity} = \#\text{nodes} \quad (42)$$

`parsimony` is a hyperparameter influencing how much equations with high complexity should be punished. If it is large, trees of low complexities are preferred.

The best tree from the subsample undergoes mutation to create a new tree. Such a mutation is a set of randomly chosen changes to the tree such as exchanging, adding, inserting or deleting a node. After mutation the decision needs to be made whether to accept the new tree to the population or not.

This is based on *simulated annealing*, a useful technique for a search of a global optimum in a high dimensional space. The idea is to accept worse solutions with a certain probability to prevent the algorithm from being stuck in a local optimum, while in each round the probability of accepting a worse solution is decreased. A new tree is accepted with the probability:

$$p = \exp\left(-\frac{\text{score}_{\text{new}} - \text{score}_{\text{old}}}{\text{alpha} \cdot T}\right) \quad (43)$$

with T the annealing temperature that is linearly decreased from 1 to 0 with every generation and `alpha` a hyperparameter. If the new tree is better than the old one ($\text{score}_{\text{new}} < \text{score}_{\text{old}}$), the new tree is accepted. If the new `score` is however larger, the acceptance probability is exponentially suppressed.

If a tree is accepted it replaces the oldest member of the population. It is important to make clear that it does not replace the tree that was chosen for mutation, since it was the best from the subsample and therefore a comparatively good fit it would be counterproductive to discard it from the population.

To create a new generation, this procedure is repeated for the amount of subsamples that are in the population, i.e. for a population of 1000 individuals it is performed 100 times. However in each loop the samples are chosen randomly for simplicity at the cost that not every individual might be considered. Afterwards the annealing temperature is reduced and the next generation is being created.

Once the last generation is obtained, the entire population undergoes an optimization process. This includes tree simplification and operator combination where for example a branch that sums two constants is replaced by a single constant. A constant optimization is performed only in 10 % of the cases since it is very time consuming. This is somewhat problematic since the constants generated during tree construction/mutation are random numbers from the normal distribution and naturally would not give a good fit. We will discuss a method to improve the outcome later in Sec. 3.4.

After optimization the hall of fame is created. It saves for each complexity the best result from all populations according to the MSE. A function with a higher complexity is only added if its MSE is lower than for the previous one. Since we not only obtain a single result but a list of

⁴This is not to be confused with the score that is the optimal observable.

formulas, we have the opportunity to choose a result that fits our needs balancing low expressivity and good MSE.

Before the start of the next iteration, a migration step is performed. Equations are replaced by members from other populations or the hall of fame according to the parameters `fractionReplaced` and `fractionReplacedHof`. The hyperparameter `niterations` sets the amount of iterations performed in total. We usually choose this number to be very large since having access to the hall of fame at all times we can stop the algorithm once its performance reaches a plateau.

Another relevant hyperparameter is `maxsize` that sets the maximal complexity of a function. We can also add constraints for every operator restricting the complexity of its branch to obtain readable results. Unless further specified, we will use the following settings:

- `populations=10`
- `npop=1000`
- `niterations=300`
- `ncyclesperiteration=200`
- `alpha=1`
- `parsimony=10-7`
- `fractionReplaced=0.5`
- `fractionReplacedHof=0.2`
- `maxsize=50`

3 ZH production

For our first attempt on applying symbolic regression to learn the optimal observable we consider the ZH production process shown in Fig. 2. The main advantage of this toy process is that it only has 2 final state particles. From energy-momentum conservation we only have 2 degrees of freedom in the center of mass frame. This keeps the analysis simple and gives us the opportunity to focus on the performance of our symbolic regression tool.

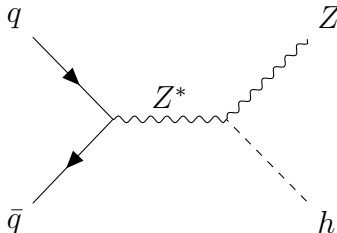


Figure 2: Feynman diagram for ZH production

The physical parameter of interest is the Wilson coefficient of one of the dimension 6 operators introduced in Tab. 1 that influences the ZZH -vertex. We will focus on one parameter only to keep the analysis simple. From the list of CP -conserving operators all but \mathcal{O}_{GG} are valid choices. Narrowing it down we require a covariant derivative of the Higgs field to see the influence of the operator in the momentum distribution. This leaves us only with \mathcal{O}_B and \mathcal{O}_W . From these two we choose the one with the simpler expression:

$$\mathcal{O}_B = \frac{ig'}{2}(D^\mu\phi)^\dagger D^\nu\phi B_{\mu\nu} \quad (44)$$

The corresponding Wilson coefficient is f_B . We will start by deriving the Feynman rules from this operator followed by the calculation of the matrix element for this process. Therefore we will consider a simplified process excluding the contribution of the photon propagator and taking only u -quarks in the initial state. The result can be used to calculate the score analytically and compare it to MADMINER results. Afterwards we will discuss the symbolic regression results for 2 different values of the Wilson coefficient for the simplified process, the process including the photon propagator and the one including all quarks. Detector and shower simulation as well as the determination of the confidence intervals we leave for the more complicated process of WBF in Sec. 4.

3.1 Feynman rules and matrix element

Feynman rules

We will use the conventions introduced in Sec. 1.1 and start by inserting the expression for the covariant derivative of the Higgs field from Eq. (5) into the operator in Eq. (44):

$$\begin{aligned} \mathcal{L}_{\text{EFT}} &\supset \frac{f_B}{\Lambda^2} \frac{ig'}{2} (D^\mu\phi)^\dagger D^\nu\phi B_{\mu\nu} \\ &= \frac{f_B}{\Lambda^2} \frac{ig'}{2} \left(\partial^\mu\phi^\dagger - i\frac{g}{2}\sigma^k W^{k\mu}\phi^\dagger - i\frac{g'}{2}B^\mu\phi^\dagger \right) \left(\partial^\nu\phi + i\frac{g}{2}\sigma^k W^{k\nu}\phi + i\frac{g'}{2}B^\nu\phi \right) B_{\mu\nu} \end{aligned} \quad (45)$$

Inserting the expression for the Higgs field as in Eq. (9), multiplying out the brackets and keeping only the relevant terms yields:

$$\supset \frac{f_B}{\Lambda^2} \frac{ig'}{2} \left\{ \frac{\partial^\mu h}{\sqrt{2}} \left(\frac{ig'}{2} B^\nu - \frac{ig}{2} W^{3\nu} \right) \frac{v}{\sqrt{2}} + \left(\frac{-ig'}{2} B^\mu + \frac{ig}{2} W^{3\mu} \right) \frac{v}{\sqrt{2}} \frac{\partial^\nu h}{\sqrt{2}} \right\} B_{\mu\nu} \quad (46)$$

$$= \frac{f_B}{\Lambda^2} \frac{g'v}{8} \left\{ \partial^\mu h (-g' B^\nu + gW^{3\nu}) + (g' B^\mu - gW^{3\mu}) \partial^\nu h \right\} B_{\mu\nu} \quad (47)$$

The expressions in the round brackets we can already identify with the Z -boson defined in Eq. (11). Simultaneously we insert the field strength tensor for $B_{\mu\nu}$ given in Eq. (3):

$$= \frac{f_B}{\Lambda^2} \frac{g'v}{8} \sqrt{g^2 + g'^2} \{ (\partial^\mu h) Z^\nu - Z^\mu (\partial^\nu h) \} (\partial_\mu B_\nu - \partial_\nu B_\mu) \quad (48)$$

Multiplying out we can collect terms by a cautious interchange of indices:

$$= \frac{f_B}{\Lambda^2} \frac{g'v}{4} \sqrt{g^2 + g'^2} \{ (\partial^\nu h) Z^\mu (\partial_\nu B_\mu) - Z^\mu (\partial^\nu h) (\partial_\mu B_\nu) \} \quad (49)$$

Further we insert the momentum operator $p_\mu = -i\partial_\mu$ and obtain:

$$= \frac{f_B}{\Lambda^2} \frac{g'v}{4} \sqrt{g^2 + g'^2} (Z^\mu p_H^\nu h p_\mu^B B_\nu - p_H^\nu h Z^\mu p_\nu^B B_\mu) \quad (50)$$

$$= \frac{f_B}{\Lambda^2} \frac{g'v}{4} \sqrt{g^2 + g'^2} (p_\nu^H p_\mu^B - p^H \cdot p^B g_{\mu\nu}) h Z^\mu B^\nu \quad (51)$$

We can again use Eq. (11) to replace the B -field and obtain the final result:

$$= \frac{f_B}{\Lambda^2} \frac{g'v}{4} (p_\nu^H p_\mu^B - p^H \cdot p^B g_{\mu\nu}) h Z^\mu (gA^\nu - g'Z^\nu) \quad (52)$$

From this expression we get two different vertex factors. One of them accounts for the AZH coupling, an interaction that does not exist in the Standard Model. It is given by:

$$-i \frac{f_B}{\Lambda^2} \frac{gg'v}{4} (p_\nu^H p_\mu^A - p^H \cdot p^A g_{\mu\nu}) \quad (53)$$

We will discuss its influence in Sec. 3.5. The second is the desired vertex for the ZZH coupling. Writing it down we have to be careful since we have two identical particles, we need to symmetrize the whole term:

$$\frac{f_B}{\Lambda^2} \frac{g'^2 v}{4} (p_\nu^H p_\mu^{Z_1} + p_\mu^H p_\nu^{Z_2} - g_{\mu\nu} p^H \cdot (p^{Z_1} + p^{Z_2})) h Z_1^\mu Z_2^\nu \quad (54)$$

Taking into account the Standard Model contribution derived in Sec. 1.1 the whole expression for the ZZH -vertex is given by:

$$iL_{\mu\nu} = i \frac{2m_Z^2}{v} g_{\mu\nu} + i \frac{f_B}{\Lambda^2} \frac{g'^2 v}{4} (p_\nu^H p_\mu^{Z_1} + p_\mu^H p_\nu^{Z_2} - g_{\mu\nu} p^H \cdot (p^{Z_1} + p^{Z_2})) \quad (55)$$

Matrix element

The matrix element for the process shown in Fig. 2 is given by:

$$i\mathcal{M} = \bar{\nu}(p_1) \left[-i \frac{g}{2 \cos \theta_W} \gamma_\mu (V - A\gamma_5) \right] u(p_2) \left[\frac{-i}{k^2 - m_Z^2} \left(g^{\mu\nu} - \frac{k^\mu k^\nu}{m_Z^2} \right) \right] iL_{\nu\lambda} \epsilon^\lambda \quad (56)$$

with p_1 and p_2 being the momenta of the incoming quarks, k the momentum of the propagator, $L_{\nu\lambda}$ the vertex from Eq. (55) and ϵ^λ the polarization of the outgoing Z -boson. V and A are the corresponding vector and axial-vector couplings of the Z -boson. Their value depends on the quark flavor.

We can write the squared matrix element averaging over initial spins and summing over final state polarizations in the following way:

$$|\mathcal{M}|^2 = \frac{1}{4} \frac{g^2}{4c_\theta^2} \bar{\nu}(p_1) \gamma^\mu (V - A\gamma_5) u(p_2) \bar{u}(p_2) \gamma^\rho (V - A\gamma_5) \nu(p_1) \\ \times \left(\frac{-g_{\mu\nu} + \frac{k_\mu k_\nu}{m_Z^2}}{k^2 - m_Z^2} \right) \left(\frac{-g_{\rho\sigma} + \frac{k_\rho k_\sigma}{m_Z^2}}{k^2 - m_Z^2} \right) L^{\nu\lambda} L^{\sigma\kappa} \epsilon_\lambda \epsilon_\kappa^* \quad (57)$$

We replace the spinors with the trace over γ -matrices and give the result:

$$= \frac{g^2}{4c_\theta^2} (V^2 + A^2) [p_1^\rho p_2^\mu + p_1^\mu p_2^\rho - g^{\rho\mu} (p_1 \cdot p_2)] \left(\frac{-g_{\mu\nu} + \frac{k_\mu k_\nu}{m_Z^2}}{k^2 - m_Z^2} \right) \left(\frac{-g_{\rho\sigma} + \frac{k_\rho k_\sigma}{m_Z^2}}{k^2 - m_Z^2} \right) L^{\nu\lambda} L^{\sigma\kappa} \epsilon_\lambda \epsilon_\kappa^* \quad (58)$$

The first simplification we can perform is to neglect the terms containing k_μ in the propagator, contracted with the fermion current they give 0 in the limit of massless fermions. This can be easily shown in the following way:

$$[p_1^\rho p_2^\mu + p_1^\mu p_2^\rho - g^{\rho\mu} (p_1 \cdot p_2)] k_\mu = p_1^\rho (p_2 \cdot k) + p_2^\rho (p_1 \cdot k) - k^\rho (p_1 \cdot p_2) \quad (59)$$

$$= p_1^\rho (p_2 \cdot p_1) + p_2^\rho (p_1 \cdot p_2) - (p_1^\rho + p_2^\rho) (p_1 \cdot p_2) = 0 \quad (60)$$

Where we used $k = p_1 + p_2$. For the matrix element squared this yields:

$$|\mathcal{M}|^2 = \frac{g^2}{4c_\theta^2} \frac{(V^2 + A^2)}{(k^2 - m_Z^2)^2} [p_1^\rho p_2^\mu + p_1^\mu p_2^\rho - g^{\rho\mu} (p_1 \cdot p_2)] g_{\mu\nu} g_{\rho\sigma} L^{\nu\lambda} L^{\sigma\kappa} \left(-g_{\lambda\kappa} + \frac{p_{Z\lambda} p_{Z\kappa}}{m_Z^2} \right) \quad (61)$$

$$= \frac{g^2}{4c_\theta^2} \frac{(V^2 + A^2)}{(k^2 - m_Z^2)^2} [p_{1\sigma} p_{2\nu} + p_{1\nu} p_{2\sigma} - g_{\sigma\nu} (p_1 \cdot p_2)] L^{\nu\lambda} L^{\sigma\kappa} \left(-g_{\lambda\kappa} + \frac{p_{Z\lambda} p_{Z\kappa}}{m_Z^2} \right) \quad (62)$$

where we performed the sum over the polarizations.

This is as far as we can go if we want to keep the expression restricted to one line. According to the expression in Eq. (35) we will look at the terms for SM contribution p_0 , the interference term $a\theta$ and the quadratic term $b\theta^2$ separately. For the SM part the vertex structure is simply given by:

$$L^{\nu\lambda} L^{\sigma\kappa} \Big|_{SM} = \frac{4m_z^4}{v^2} g^{\nu\lambda} g^{\sigma\kappa} \quad (63)$$

We can arrive at a rather simple expression for p_0 in the following way:

$$p_0 = \frac{g^2(V^2 + A^2)m_Z^4}{v^2c_\theta^2(k^2 - m_Z^2)^2} [p_{1\sigma}p_{2\nu} + p_{1\nu}p_{2\sigma} - g_{\sigma\nu}(p_1 \cdot p_2)] g^{\nu\lambda}g^{\sigma\kappa} \left(-g_{\lambda\kappa} + \frac{p_{Z\lambda}p_{Z\kappa}}{m_Z^2} \right) \quad (64)$$

$$= \frac{g^2(V^2 + A^2)m_Z^4}{v^2c_\theta^2(k^2 - m_Z^2)^2} [p_{1\sigma}p_{2\nu} + p_{1\nu}p_{2\sigma} - g_{\sigma\nu}(p_1 \cdot p_2)] \left(-g^{\nu\sigma} + \frac{p_Z^\nu p_Z^\sigma}{m_Z^2} \right) \quad (65)$$

$$= \frac{g^2(V^2 + A^2)m_Z^4}{v^2c_\theta^2(k^2 - m_Z^2)^2} \left[(p_1 \cdot p_2) + \frac{2}{m_Z^2}(p_1 \cdot p_Z)(p_2 \cdot p_Z) \right] \quad (66)$$

$$= \frac{g^2(V^2 + A^2)m_Z^2}{v^2c_\theta^2(k^2 - m_Z^2)^2} 2E_1E_2 [m_Z^2 + (E^Z - p_z^Z)(E^Z + p_z^Z)] \quad (67)$$

$$= \frac{g^2(V^2 + A^2)m_Z^2}{v^2c_\theta^2(k^2 - m_Z^2)^2} 2E_1E_2 [2m_Z^2 + p_T^2] \quad (68)$$

In the last step we used the following notation for the momenta imposing energy-momentum conservation:

$$p_1 = \begin{pmatrix} E_1 \\ \vec{0} \\ E_1 \end{pmatrix} \quad p_2 = \begin{pmatrix} E_2 \\ \vec{0} \\ -E_2 \end{pmatrix} \quad p^Z = \begin{pmatrix} E^Z \\ \vec{p}_T \\ p_z^Z \end{pmatrix} \quad p^H = \begin{pmatrix} E^H \\ -\vec{p}_T \\ p_z^H \end{pmatrix} \quad (69)$$

with the transverse momentum $\vec{p}_T = (p_x \ p_y)^T$.

For the interference term the vertex factor can be written as:

$$L^{\nu\lambda}L^{\sigma\kappa} \Big|_{interference} = \frac{f_B}{\Lambda^2} m_Z^2 g^2 [p_H^\lambda k^\nu + p_H^\nu p_Z^\lambda - p_H \cdot (k + p_Z) g^{\nu\lambda}] g^{\sigma\kappa} \quad (70)$$

The first term proportional to k^ν can be neglected following the same reasoning as in Eq. (60). Inserting this again into the general form of the matrix element given in Eq. (62) yields:

$$a\theta = \frac{f_B}{\Lambda^2} \frac{g^2 g'^2 m_Z^2 (V^2 + A^2)}{4c_\theta^2(k^2 - m_Z^2)^2} [p_{1\sigma}p_{2\nu} + p_{1\nu}p_{2\sigma} - g_{\sigma\nu}(p_1 \cdot p_2)] \times [p_H^\nu p_Z^\lambda - p_H \cdot (k + p_Z) g^{\nu\lambda}] \left(-g_\lambda^\sigma + \frac{p_{Z\lambda}p_Z^\sigma}{m_Z^2} \right) \quad (71)$$

Expanding the brackets in the second line we get:

$$= \frac{f_B}{\Lambda^2} \frac{g^2 g'^2 m_Z^2 (V^2 + A^2)}{4c_\theta^2(k^2 - m_Z^2)^2} [p_{1\sigma}p_{2\nu} + p_{1\nu}p_{2\sigma} - g_{\sigma\nu}(p_1 \cdot p_2)] \times \left[-p_H^\nu p_Z^\sigma + p_H \cdot (k + p_Z) g^{\nu\sigma} + p_H^\nu p_Z^\sigma - p_H \cdot (k + p_Z) \frac{p_Z^\nu p_Z^\sigma}{m_Z^2} \right] \quad (72)$$

$$= \frac{f_B}{\Lambda^2} \frac{g^2 g'^2 m_Z^2 (V^2 + A^2)}{4c_\theta^2(k^2 - m_Z^2)^2} [p_{1\sigma}p_{2\nu} + p_{1\nu}p_{2\sigma} - g_{\sigma\nu}(p_1 \cdot p_2)] p_H \cdot (k + p_Z) \left(g^{\nu\sigma} - \frac{p_Z^\nu p_Z^\sigma}{m_Z^2} \right) \quad (73)$$

We recognize that we obtain a term proportional to the SM as in Eq. (65) including an additional momentum dependence. We simplify and obtain:

$$= \frac{f_B g^2 g'^2 (V^2 + A^2)}{\Lambda^2 4c_\theta^2 (k^2 - m_Z^2)^2} 2E_1 E_2 (p_H \cdot (k + p_Z)) [2m_Z^2 + p_T^2] \quad (74)$$

For the last step we insert $k = p_Z + p_H$:

$$= \frac{f_B g^2 g'^2 (V^2 + A^2)}{\Lambda^2 4c_\theta^2 (k^2 - m_Z^2)^2} 2E_1 E_2 (m_H^2 + 2(p_H \cdot p_Z)) [2m_Z^2 + p_T^2] \quad (75)$$

The calculation for the quadratic term can be performed in full analogy. The only difference to the calculation of the interference term is that we have one more term for which we have to contract the indices. Thus we only write the final result:

$$b\theta^2 = \frac{f_B^2 g^2 g'^4 v^2 (V^2 + A^2)}{\Lambda^4 64c_\theta^2 (k^2 - m_Z^2)^2 m_Z^2} 2E_1 E_2 (m_H^2 + 2(p_H \cdot p_Z))^2 [2m_Z^2 + p_T^2] \quad (76)$$

To write out the full matrix element squared we combine the terms from Eq. (68), (75) and (76) to obtain:

$$|\mathcal{M}|^2 = \frac{g^2 (V^2 + A^2)}{c_\theta^2 (k^2 - m_Z^2)^2} 2E_1 E_2 \left(\frac{m_Z}{v} + \frac{f_B g'^2 v}{\Lambda^2 8m_Z} (m_H^2 + 2p_H p_Z) \right)^2 (2m_Z^2 + p_T^2) \quad (77)$$

3.2 Kinematic observables

Before we start our analysis of the score, which will be a non-trivial function of kinematic observables, we need to choose the observable quantities we want our function to depend on.

From the matrix element in Eq. (56) we can already see that p_T will be a useful observable. Furthermore we have a scalar product of the two outgoing particles which includes the momenta in the z -direction. Instead of p_z we will use the pseudorapidity η as it is usually done in high energy physics. It is given by:

$$\eta = -\ln \left(\tan \frac{\theta}{2} \right) \quad (78)$$

where θ is the angle between the particle momentum and the positive direction of the beam axis. For consideration we choose the sum of the pseudorapidities $\eta_+ = \eta_Z + \eta_H$ and their difference

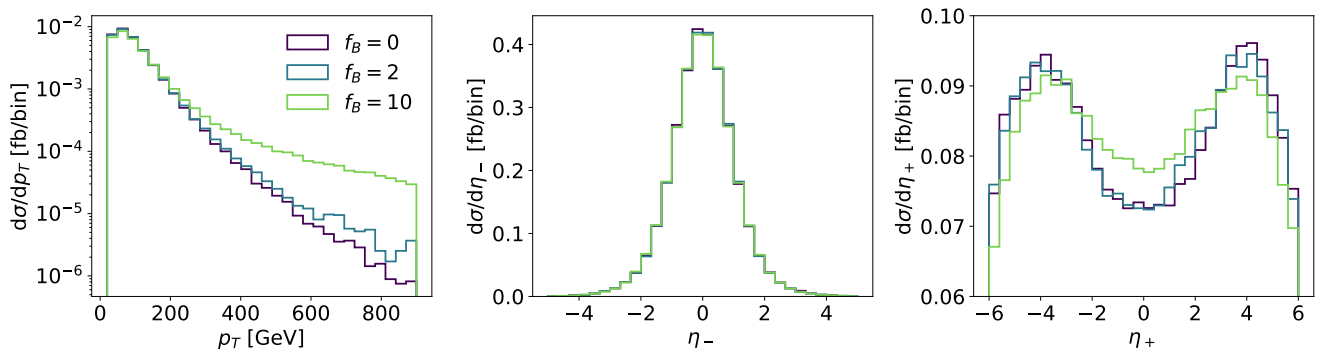


Figure 3: Kinematic distributions for different Wilson coefficients with $\eta_\pm = \eta_Z \pm \eta_H$

$\eta_- = \eta_Z - \eta_H$. We can look at the distribution of these variables to determine which of them are sensitive to a change of the Wilson coefficient.

As described in Sec. 2.1 we generate 500k events with MADGRAPH5 using the EWDIM6 [28] model file that implements our operator. We perform the simulation for 3 values of the Wilson coefficient: $f_B = 0, 2, 10$. Even though the last one is ruled out by experiments it will be useful to us for illustrative purposes. The distributions for the kinematic observables are depicted in Fig. 3.

At first sight it appears that the Wilson coefficient affects only p_T and η_+ while the η_- distribution does not show any dependence on f_B . This is however an artifact of looking at one dimensional histograms. To decide whether η_+ or η_- is sensitive to new physics, we have to look at 2d histograms of these two variables. To eliminate any possible correlations with p_T we construct the histograms for different slices in p_T . We will focus on areas with $p_T > 300$ GeV which are more sensitive to new physics. For the histograms we compare $f_B = 0$ and $f_B = 10$. Additionally, we take the ratio of the histograms as a proxy for the score and divide the histogram for $f_B = 10$ by the one for $f_B = 0$.

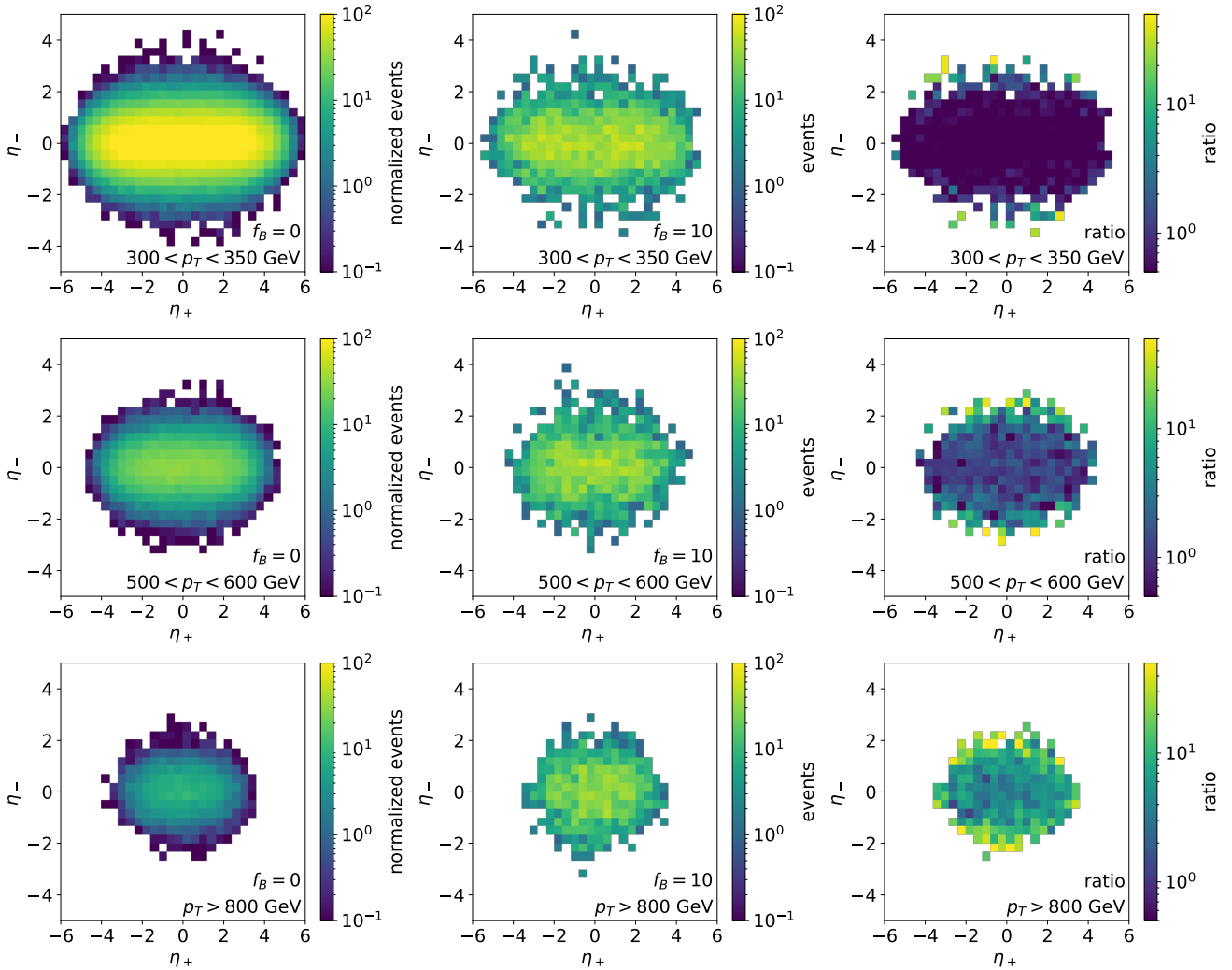


Figure 4: Kinematic correlation of η_- and η_+ . The right column shows the histogram ratio as a proxy for the score.

To have enough statistics in the high p_T -region for $f_B = 0$ we generate more events and normalize the histogram accordingly. The results are shown in Fig. 4. The histogram ratio shows that indeed, unlike the suggestion of the one dimensional histograms in Fig. 3, η_- is the sensitive observable for the Wilson coefficient. We clearly see that towards large $|\eta_-|$ the ratio increases while it does not change in the η_+ direction.

From these histograms we can also understand the one dimensional distribution for η_+ . We see that towards larger p_T the distribution in the $\eta_+ - \eta_-$ plane becomes more narrow and more round. This is simply because large η_+ correspond to cases where both final state particles fly in the same direction with a high p_z , such a scenario is highly unlikely if also the transverse momentum is required to be large and is therefore limited by energy-momentum conservation. Thus, for a large Wilson coefficient, where we have more events in the high p_T -region, these events will naturally be at lower $|\eta_+|$ values and for low Wilson coefficients we have more events in the low p_T -region of which a significant part will have large $|\eta_+|$ values.

For further analysis we therefore discard η_+ and only use p_T and η_- as the relevant kinematic observables for the score fit.

3.3 Score for f_B

Now that we established what variables we are going to use, we can analyze the score in terms of these observables. As a starting point we calculate the score analytically and compare it to MADMINER results.

Therefore we approximate the score with the expression in Eq. (36). For the full score we would have to take the term of Eq. (37) into account which is not analytically calculable. The

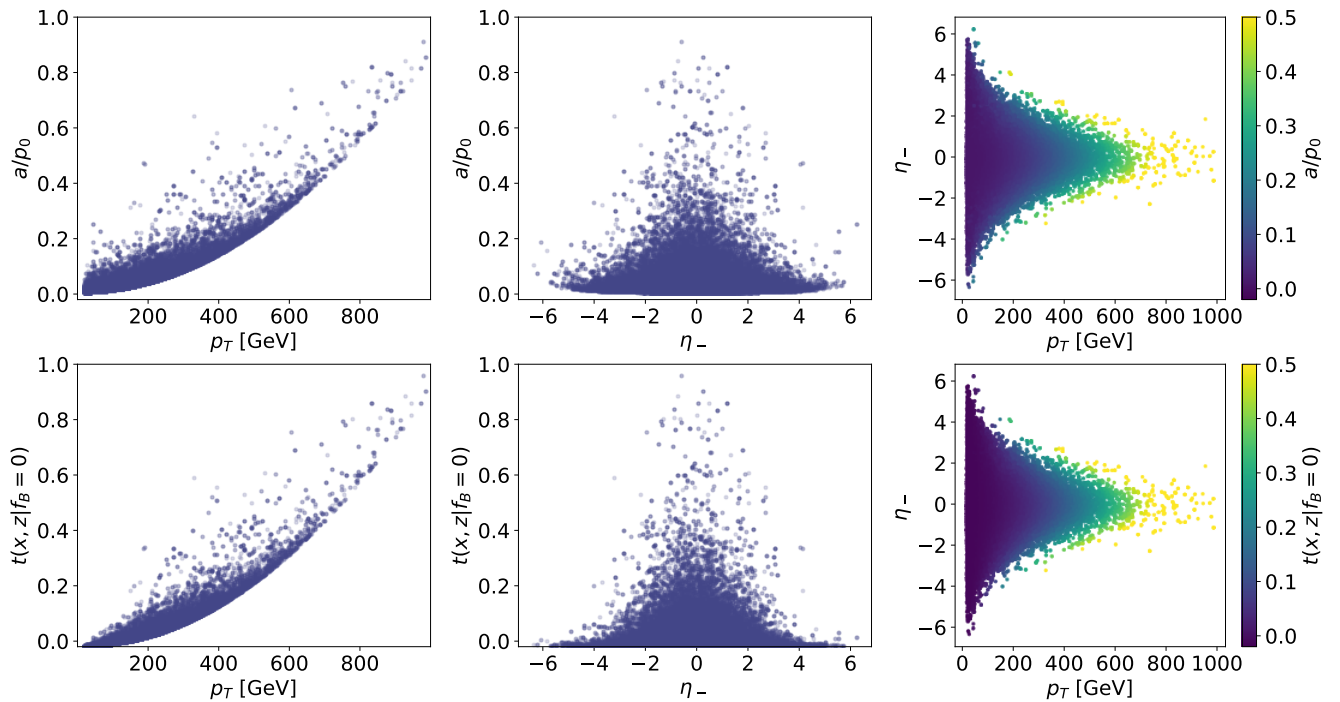


Figure 5: Joint score for $f_B = 0$. Top: calculated approximation. Bottom: result from MADMINER.

approximate score in the case of the SM is then given by:

$$t(x, z|f_B = 0) \approx \frac{a + 2bf_B}{p_0 + af_B + bf_B^2} \Big|_{f_B=0} = \frac{a}{p_0} = \frac{f_B g'^2 v^2}{4\Lambda^2 m_Z^2} (m_H^2 + 2p_H p_Z) \quad (79)$$

In the last step we inserted the results from the squared matrix element calculation of Eq. (68) and (75). This obtained a very compact result. The calculated score is shown in the top row of Fig. 5. The lower row shows the full score obtained through MADMINER. The first difference we see is a constant offset between the two results. This is due to the constant term of the total cross section that we explicitly neglected in our approximation. Furthermore we see some discrepancies in the $p_T \lesssim 215$ GeV region. This value corresponds to the sum of the masses of the final states. It is an artifact originating from MADGRAPH5 due to a condition that the center of mass energy should at least be twice as large as the sum of the external masses for the calculation of the matrix element to ensure numerical stability. Since this region is not significantly important to us we can neglect it for further considerations. We conclude that our analytic calculation sufficiently agrees with the result from MADMINER.

A few more words can be said about the distribution of the score. First of all the score grows with p_T . We remember that the score is a derivative of the log-likelihood with respect to θ . The bigger the change, the larger the derivative. Therefore this behavior is expected since we already established that high p_T -regions are more sensitive to a change of the Wilson coefficients. For the η_- distribution in the middle plots it looks like the score is larger for small η_- but this is an artifact from the phase space. Looking at the 2d plots on the right hand side we see that for a given slice in p_T the score grows towards large $|\eta_-|$. This corresponds to the observation we made with the 2d histogram ratios in Fig. 4 where the ratio increased towards larger $|\eta_-|$. Additionally we want to point out that in the 2d plot the score shows a smooth color gradient without any noise. This indicates that the score is fully described by these 2 variables and there are no other degrees of freedom in the system which confirms the discussion in Sec. 3.2.

In the next step we want to qualitatively understand how the score changes when increasing the Wilson coefficient. In Fig. 6 we show the score for $f_B = 0, 2, 10$ for each variable while keeping the other variable fixed. We see that in the sensitive regions, i.e. large p_T and large $|\eta_-|$, the score decreases with growing Wilson coefficient. This behavior might seem puzzling at first, as for higher score values we would expect a bigger increase. However, we need to remember that the

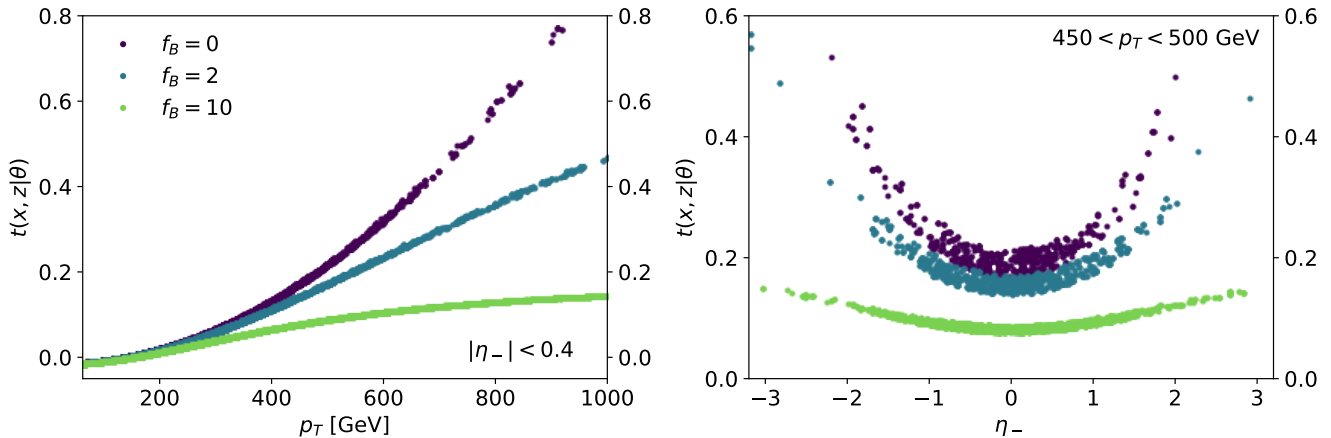


Figure 6: Score for different Wilson coefficients f_B .

	$\theta \ll 1$	$\theta \gg 1$
approximation	leading term $\frac{a}{p_0} + \frac{1}{p_0} \left(2b - \frac{a^2}{p_0} \right) \theta$	quadratic term $\frac{2}{\theta}$
scaling	constant or linear	decreasing with θ

Table 2: Approximations for the first term of the joint score in Eq. (36).

score considers the relative change and not the absolute one since it takes the derivative of the log-likelihood and not the likelihood itself. Therefore it is the relative change that decreases while the absolute one still increases.

This discussion can be done in a far more general framework. Taking again the general expression for the first term of the joint score from Eq. (36) we can perform two approximations, for $\theta \gg 1$ and $\theta \ll 1$. In the first case of large θ we can simply neglect the constant and linear terms in the denominator as well as the interference term in the numerator. We remain with:

$$t(x, z | \theta \gg 1) \approx \frac{2}{\theta} \quad (80)$$

This is exactly what we see in Fig. 6 for $f_B = 10$ where the score is limited from above by 0.2. For small θ we perform a first order Taylor expansion. The joint score is then given by:

$$t(z | \theta) \approx \frac{a}{p_0} + \frac{1}{p_0} \left(2b - \frac{a^2}{p_0} \right) \theta \quad (81)$$

The joint score will decrease as long as $2b < a^2/p_0$. To what extent this approximation is appropriate and when it breaks is highly dependent on the parameters a , b and p_0 which of course depend on the phase-space.

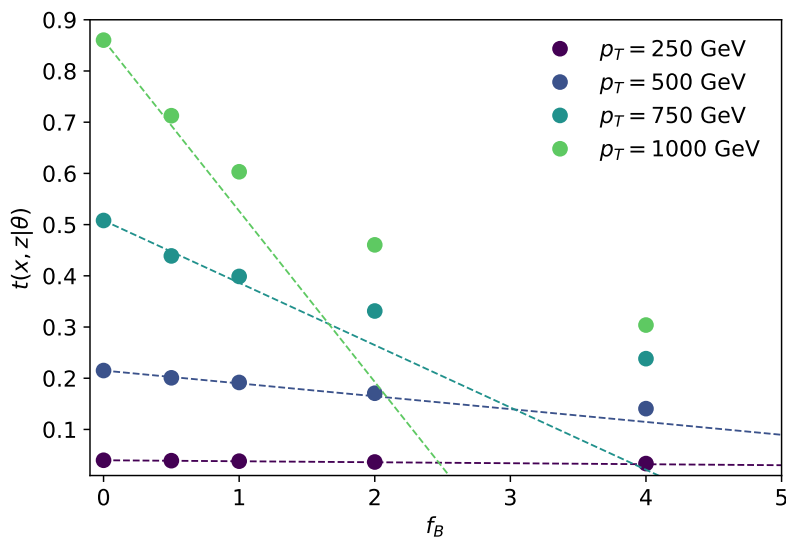


Figure 7: Joint score for small Wilson coefficients f_B for different p_T with $|\eta_-| \approx 0$. Dotted line is the Taylor expansion of Eq. (81).

To see how this applies on our data we additionally generate events for $f_B = 0.5, 1, 4$. For each Wilson coefficient we compare the true joint score to the prediction from the Taylor expansion. We calculate the score approximation for different values of p_T while fixing $|\eta_-| \approx 0$ for simplicity. The results are shown in Fig. 7. Note that to avoid the problem with the missing constant in our calculation we only fit the slope to the true score value at $f_B = 0$.

We see that for $f_B \leq 0.5$ the approximation fits for all p_T regions as predicted. For $p_T \leq 500$ GeV the approximation is even valid almost upto $f_B = 4$ and for $p_T \leq 750$ GeV upto $f_B = 1$. As it was the case for large θ the score decreases here as well. To conclude, while the approximation is valid for many regions in phase space, it is not wise to neglect higher orders of the Taylor expansion if we focus our analysis on the high p_T regions which are most sensitive to θ .

3.4 Symbolic regression on joint score

After establishing a general understanding of the joint score as well as its dependence on kinematic observables, we can apply symbolic regression to find a formula for the score. Since we are looking at a simplified process without any latent variables, the score is essentially the joint score and we expect the fit error to go to 0.

Some modifications on the data set are necessary to use PYSR. First of all we scale our input data in the following way:

$$x_p = p_T/m_H \quad x_\eta = |\eta_-| \quad (82)$$

This ensures that all quantities are dimensionless and in the same order of magnitude. For the PYSR algorithm this is helpful because of the random constant sampling in the tree construction or mutation process. From Sec. 2.3 we remember that a weakness of the algorithm was that there is no constant optimization after mutation and before acceptance, but only at the end of a full SR iteration and only for 10 % of the population. Having the input variables in the same order of magnitude increases the chances that the random constants are at least in the right order of magnitude.

Furthermore, the PYSR run time scales very badly with the size of the data set. Reducing the data set by taking for instance every 500-th event is problematic, since most of the data points lie in the low score region which are of less interest to us, whereas high-scored data points are very sparse. This can be seen best in Fig. 5. With such a sampling we will simply lose the data of interest. Instead, we sample points according to intervals in p_T , i.e. for each interval we take a fixed amount of data points and collect 800 phase-space points in total. This of course leads to a bias in the fit as the reduced data set is not representative of the old one. To balance this effect we introduce an optimization step in which we take the PYSR output and fit the constants on the original data set. We rely on the python package LMFIT [29] for non-linear optimization and curve fitting which is based on SCIPY.OPTIMIZE [26].

3.4.1 Polynomial functions for $f_B = 0$

From Fig. 6 and 5 we expect the true function to be a polynomial in x_p and x_η . To establish a baseline we fit polynomials of degrees 2 to 4 using LMFIT. The results are shown in Fig. 8 with the fit parameters given in Tab. 3. They are compared to the best PYSR result discussed below. We see that already the polynomial of degree 2 describes most of the data well. For a better accuracy in the high $|\eta_-|$ regions terms including x_η^3 are needed. These are the origin of the major improvement in the MSE for the polynomials of degree 3 and 4.

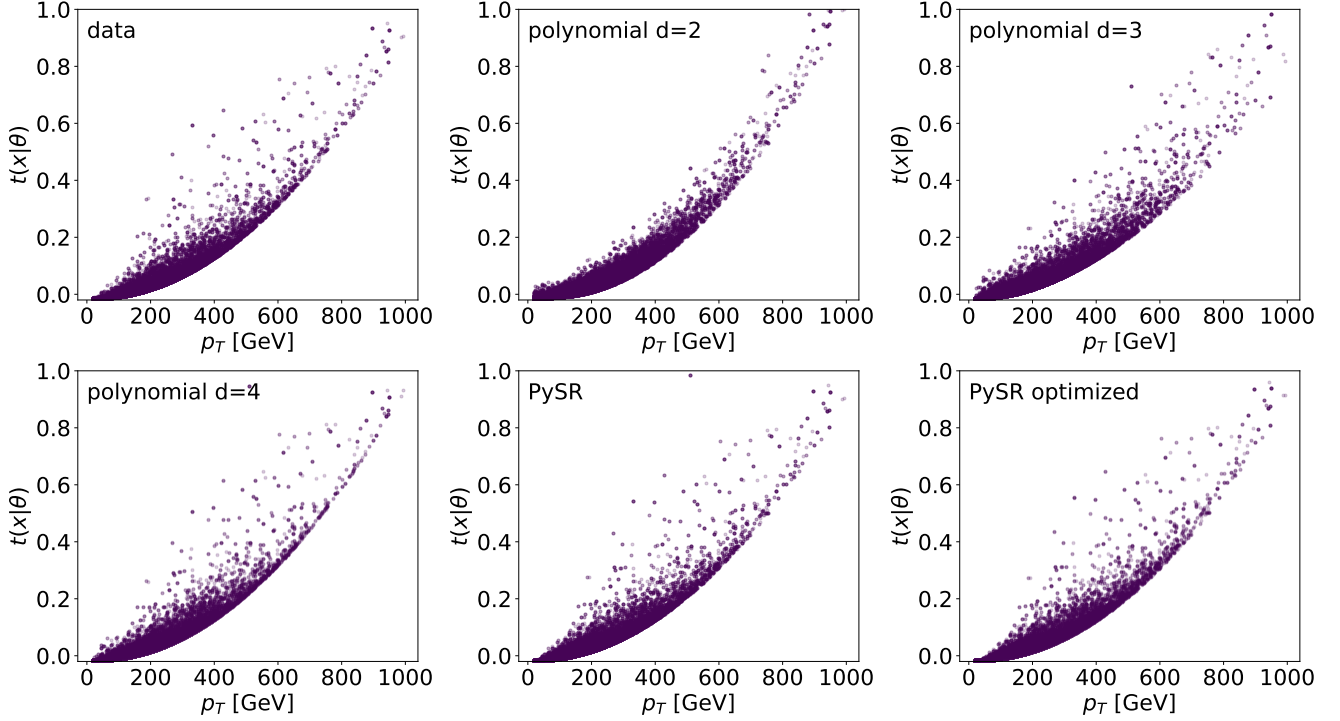


Figure 8: Score as a function of p_T for the polynomial fits and the PySR output, including optimization, for the simplified ZH setup with $f_B = 0$, corresponding to Tab. 3.

	polynomial $d = 2$	polynomial $d = 3$	polynomial $d = 4$	PySR	PySR optimized
MSE	$3.49 \cdot 10^{-3}$	$8.16 \cdot 10^{-4}$	$1.28 \cdot 10^{-4}$	$1.23 \cdot 10^{-4}$	$7.65 \cdot 10^{-5}$
dof	6	10	15	9	9
1	-0.03145	-0.1810	-0.1231	-0.1495	-0.134807(46)
x_p	-0.2022	0.4871	-0.06404	-0.01553	-0.036030(78)
x_η	-0.1783	0.1837	-0.04830	0.0045	0.002083(55)
x_p^2	0.1805	0.1303	0.1612	0.1453	0.148277(26)
$x_p x_\eta$	0.2303	-0.3434	0.1124	-0.01553	-0.00787(10)
x_η^2	0.02861	-0.1036	0.06492	-	-
x_p^3	-	-0.001788	$-4.504 \cdot 10^{-4}$	-	-
$x_p^2 x_\eta$	-	0.1022	-0.03152	0.01854	0.022835(68)
$x_p x_\eta^2$	-	0.1449	-0.1551	-	-
x_η^3	-	0.01001	-0.01976	$6.333 \cdot 10^{-4}$	0.0013648(50)
x_p^4	-	-	$6.936 \cdot 10^{-5}$	-	-
$x_p^3 x_\eta$	-	-	-0.002264	-	-
$x_p^2 x_\eta^2$	-	-	0.07835	0.005143	-0.002813(67)
$x_p x_\eta^3$	-	-	0.03080	-0.007064	-0.011333(26)
x_η^4	-	-	0.001368	-	-
$x_p^2 x_\eta^3$	-	-	-	0.01970	0.023525(22)

Table 3: Polynomial score functions for the simplified ZH setup with $f_B = 0$. The right column shows the results from a post-processing fit to the PySR function. All values for $t(x_p, x_\eta) \times 10$.

We can conclude that a polynomial of degree 4 is enough to describe the data and that in such a case it would not be necessary to make use of an involved fitting tool like PYSR. For illustrative purposes we do it anyway.

For the general settings described in Sec. 2.3 using the operators $+$, $-$, \times , square and cube we obtain the hall of fame with the most prominent results given in Tab. 4. The figure on the right hand side shows the MSE after optimization for all formulas. In the algorithm of PYSR more complex functions are only accepted to the hall of fame if they come with an improvement in the MSE. This behavior can be slightly distorted after optimization for a couple of reasons. There can be a function with a higher complexity but fewer free parameters so its predecessor has a larger improvement during optimization. Another reason could be that the output function simply describes the reduced data set better than it does the full one, regardless of the constants. For our analysis we ignore these outliers.

It is interesting to look at the evolution of the hall of fame in more detail. Looking at low complexity functions, we can learn which terms are the most dominant ones much better than from the coefficients of the polynomial in Tab. 3. For instance, the function of complexity 10 has a comparable MSE as the polynomial of degree 2. Instead of 6 parameters we have only 3 and we see that the relevant terms are x_p^2 , $x_p^2 x_\eta$ and the constant. For higher complexities we observe that, correlation terms between x_p and x_η , especially $x_p^2 x_\eta^3$, are essential. Last is an order 5 term which is not present in the polynomial fits. It is the reason why the PYSR result gives a better MSE. On the other hand terms with a higher order in x_p than 2 do not appear at all. This is in agreement with the comparatively small coefficients for the polynomials in Tab. 3.

To conclude, symbolic regression provides a better and more compact formula than the polynomials. For our best result we have a factor 1.6 MSE improvement compared to the fourth-order polynomial having only 9 free parameters as opposed to 15. This is a result of the algorithm's tree construction. For example we can have a term $(a + x)^4$ which has a complexity of only 5 and just one free parameter but is a polynomial of degree 4. When expended, the prefactors are all correlated while in the polynomial fit we would have 5 independent parameters. These correlations can enforce too strong restrictions and limit the potential of improvement during optimization, but they also simplify the expressions. A big advantage of the algorithm is that irrelevant terms

cmpl.	dof	function	MSE
7	1	$ax_p(x_p + x_\eta)$	$3.81 \cdot 10^{-2}$
10	3	$ax_p^2(b + x_\eta) - c$	$2.49 \cdot 10^{-3}$
14	3	$ax_p^2 + bx_p^2 x_\eta^2 - c$	$6.64 \cdot 10^{-4}$
22	4	$ax_p^2 + bx_p^2 x_\eta^2 - cx_p x_\eta - d$	$3.09 \cdot 10^{-4}$
32	6	$a(x_p^2 + x_\eta) + bx_p^2 x_\eta - (cx_p - d)^2$ $+ ex_p^2 x_\eta^3 - f$	$2.06 \cdot 10^{-4}$
34	7	$a(x_p^2 + x_\eta) + bx_p^2 x_\eta - (cx_p - d)^2$ $+ ex_\eta^3(x_p - f)^2 - g$	$7.77 \cdot 10^{-5}$
49	9	$ax_p^2 + bx_p^2 x_\eta - cx_\eta(x_p - d)$ $+ ex_\eta^3(x_p - f)^2 + gx_p^2 x_\eta^2 - hx_p - i$	$7.65 \cdot 10^{-5}$

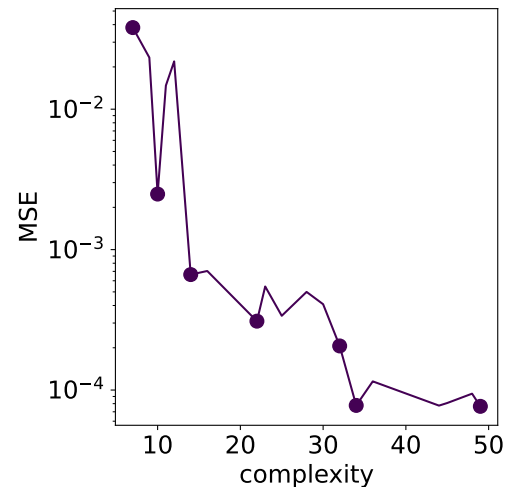


Table 4: Hall of fame for simplified ZH setup for $f_B = 0$. The last formula corresponds to the PYSR result shown in Tab. 3. MSE is given for $t(x_p, x_\eta) \times 10$.

are removed from the tree and only important once are kept. These on the other hand are not restricted by the order of polynomial unless explicitly specified.

Last but not least, even though we have no latent variables in the system and the joint score is given by a simple expression in Eq. (79), our fit performance stagnates at a MSE of $7.65 \cdot 10^{-5}$. This is due to the fact that we use η_- instead of p_z^Z and p_z^H . The connection between the two parametrizations includes log and tan which are operators we did not give to PYSR and therefore cannot obtain a perfect result. For all practical purposes the obtained approximations are sufficient.

3.4.2 Rational function for $f_B = 10$

From the previous discussion in Sec. 3.3 we deduce that a polynomial function will not be a good fit on the data due to the score being limited from above by $2/\theta$ for large θ . We therefore extend our search to the space of rational functions by including the division operator into PYSR. We increase the maximal complexity to 75 as we expect this regression task to be more complicated than the previous one.

From the obtained hall of fame we choose the following result:

$$t(x_p, x_\eta | f_B = 10) = ax_p - b + \frac{c(x_\eta + d)}{e + \frac{f}{x_p((x_p - g)^4 + h)(i(x_\eta - j)^2 + k)}} \quad (83)$$

Corresponding data is shown in Fig. 9 and Tab. 5. We see that, unlike for the previous case, the raw PYSR result shows a poor fit when it comes to the general shape of the score distribution.

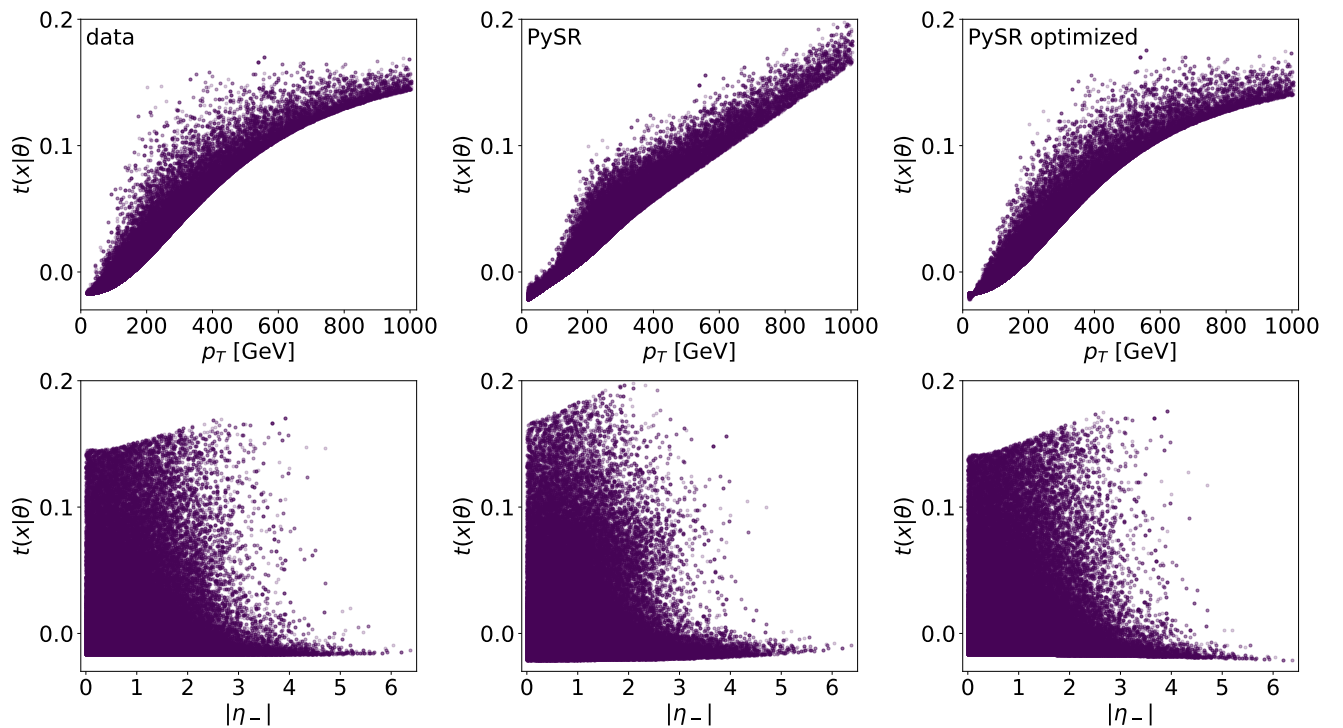


Figure 9: Score as a function of p_T and η_- for the rational PYSR output for the simplified ZH setup with $f_B = 10$, corresponding to Tab. 5.

	PYSR default	PYSR optimized		
		Eq.(83)	Eq.(84)	Eq.(85)
MSE	$8.85 \cdot 10^{-4}$	$7.52 \cdot 10^{-5}$	$7.38 \cdot 10^{-5}$	$5.42 \cdot 10^{-5}$
a	0.2201	0.02318(20)	0.01534(20)	0.00805(17)
b	0.2427	0.169067(79)	0.166262(71)	0.166229(67)
$c^{(\prime)}$	0.0249	6.2(10)	0.09973(32)	0.06691(36)
$d^{(\prime)}$	0.7070	13.667(54)	1.5949(23)	1.712(22)
e	0.1405	56.6(96)	-	-
$f^{(\prime)}$	0.7046	374(42)	2.680(21)	1.928(18)
g	0.2855	-13.834(86)	18.56(14)	23.54(20)
$h^{(\prime)}$	0.1270	$-3.945(96) \cdot 10^4$	$7.97(23) \cdot 10^{-6}$	$1.206(33) \cdot 10^{-5}$
$i^{(\prime)}$	0.5750	$2.05(30) \cdot 10^{-5}$	0.42702(54)	0.05091(78)
j	0.3189	0.336749(58)	0.32375(55)	-0.5942(55)
k	0.1192	$4.61(67) \cdot 10^{-5}$	-	-
y	fixed 2	fixed 2	fixed 2	3.3771(78)
z	fixed 4	fixed 4	fixed 4	3.5724(43)

Table 5: Rational function for the simplified ZH setup with $f_B = 10$. We compare the original parameters from PYSR to different optimization. All values for $t(x_p, x_\eta) \times 10$.

However, there is a significant improvement from the optimization fit of more than a factor of 10 in the MSE.

Comparing the values in the first and second column of Tab. 5 we can understand why the PYSR result is a poor fit. All constants are in the same order of magnitude while the true values from the optimization span over several orders of magnitude. Additionally, the function in Eq. (83) suffers from flat directions, i.e. fully correlated parameters. This causes problems both for the optimization within PYSR as well as for our post-processing as can be seen by the large fit errors. This can be taken care of by redefining the parameters of the above function in the following way:

$$t(x_p, x_\eta | f_B = 10) = ax_p - b + \frac{c'x_\eta + d'}{1 + \frac{f'}{x_p (h' (x_p + g)^4 - 1) (i' (x_\eta - j)^2 + 1)}}, \quad (84)$$

with

$$c' = \frac{c}{e} \quad d' = \frac{cd}{e} \quad f' = \frac{f}{ehk} \quad h' = \frac{1}{h} \quad i' = \frac{i}{k}$$

Thus, we have removed the two redundant parameters e and k and obtained a much more stable fit result with a slight improvement in the MSE. It is shown in the right panel of Fig. 9.

One more step that can be done in post-processing, is to check whether the exponents in the fraction are a reasonable choice. The exponents used by PYSR are fixed through the operators square and cube. To allow for non-integer values we would have to include the power operator. However, this will complicate the regression process and produce results that will be difficult to interpret. There is unfortunately no option to restrict the power operator to use constants only. Even if we restrict the operator complexity to 1, we could still obtain expressions like $x_p^{x_\eta}$ which are not useful to us. Instead we take the integer value of the exponents of our result and turn

them into fitable parameters. The score reads now:

$$t(x_p, x_\eta | f_B = 10) = ax_p - b + \frac{c'x_\eta + d'}{1 + \frac{f'}{x_p (h'|x_p + g|^z - 1) (i'|x_\eta - j|^y + 1)}} \quad (85)$$

We take absolute values to prevent the expression from being complex. We observe a significant shift in one of the parameters $z = 2 \rightarrow 3.37$ as well as an improvement in the MSE of about 35 %. For a real analysis the expression of Eq. (84) would be sufficient. The MSE is comparable to the one obtained in Sec. 3.4.1.

Ultimately, from this example we learn that post-processing is essential for rational functions with fully correlated parameters if we want to accurately fit regions with high score values. However, this has to be taken with a grain of salt: there is nothing that guarantees that a poor fit will become good after optimization. A bad parametrization might just be what it is: a bad parametrization.

3.5 Photon propagator

For the next step we include the photon propagator which we derived in Sec. 3.1 simulating a realistic process for finite f_B . We still consider only u -quarks for simplicity. Note that the presence of the photon propagator does not add any latent variables. For the squared matrix element of the process the contributions of the photon- and Z -propagator are first summed and then squared.

Looking at the joint score shown in Fig. 10 we see how it differs from the process with Z only in Fig. 8 and 9.

In the case of the SM the joint score was given by a/p_0 , the interference term normalized over the SM term. The new physics term in the matrix element is now a sum of the Z and the photon contribution. Since the photonic vertex factor comes with a minus sign relative to the ZZH vertex as shown in Eq. (53), the interference term is smaller. While the score previously reached values up to 1, the maximal values in this case are at about 0.4.

With a smaller interference term we expect the quadratic term to be more dominant for $f_B = 10$. We already discussed in Sec. 3.3 that once we can neglect everything but the quadratic

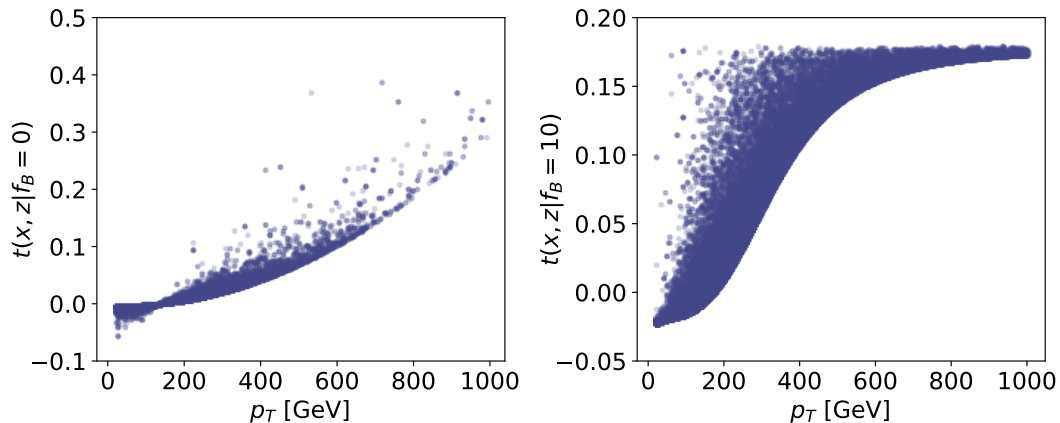


Figure 10: Joint score for ZH process including photon propagator for $f_B = 0, 10$.

term for large Wilson coefficients, the score is bound by a constant. Indeed we see that this bound is more prominent than it was for the Z only process.

In the following we fit the score in the same manner as we did in the previous chapter.

3.5.1 Results for $f_B = 0$

For $f_B = 0$ the data is even simpler to fit than it was before. We therefore just give the best result from the hall of fame:

$$t(x_p, x_\eta | f_B = 0) \times 10 = -a + b(x_p - c)^2 + dx_\eta(x_p - e)^2 - fx_\eta^2(x_p - g)^2 + hx_\eta^3(x_p - i)^2 \quad (86)$$

with

$$\begin{aligned} a &= 0.025668(21) & b &= 0.048981(17) & c &= 0.19081(51) & d &= 0.008893(27) \\ e &= 0.6110(32) & f &= 1.9(11) \cdot 10^{-7} & g &= 167(47) & h &= 0.0078760(65) & i &= 0.48588(52) \end{aligned}$$

With a MSE of $3.05 \cdot 10^{-5}$ the fit is a factor of 2 better than for the process without photon in Sec. 3.4.1. The parameters g and h have large errors indicating a high correlation. Therefore one can assume the contribution of $fx_\eta^2x_p^2$ is negligible and replace $fx_\eta^2(x_p - g)^2$ with fx_η^2 . This yields $f = 0.0055627(52)$, a much more stable result, while the MSE is unchanged.

In general the formula has a very similar structure to the ones given in Tab. 4. The highest order in x_p is again 2 and the important $x_\eta^3x_p^2$ term is present.

3.5.2 Results for $f_B = 10$

While for $f_B = 0$ we obtained an even better and simpler fit than we did for the process without the photon, the situation turns out to be far more complex for $f_B = 10$ due to the sharper bound. With the previous hyperparameter settings not only do we not obtain any reasonable parametrizations, but the algorithm does not include any functions with a higher complexity than 30. Increasing the amount of populations or iterations did not cause an improvement. The reason behind this performance is the acceptance probability of the simulated annealing.

During mutation a certain change to the tree can be in principle the right one, but because of a wrong constant its MSE might be much worse compared to the tree before mutation and since no constant optimization is performed in this step, the acceptance probability might simply be too low to accept the new result. Increasing the hyperparameter `alpha` in Eq. (43) by several orders of magnitude would result in a higher acceptance probability but the convergence of the algorithm would suffer severely. Instead we find a way to modify the acceptance probability which we justify in the following.

Neglecting `parsimony` the original formula for the acceptance probability was given by

$$p_{\text{old}} = \exp\left(-\frac{\text{MSE}_{\text{new}} - \text{MSE}_{\text{old}}}{\text{alpha} \cdot T \cdot \text{baseline}}\right) \quad (87)$$

with the baseline being the MSE to the constant unit function. The main problem is that only the absolute difference between the new (after mutation) and old (before mutation) MSE is considered. This makes the algorithm dependent on absolute numbers rather than the relative improvement. Therefore we introduce a new acceptance probability considering the relative change of the MSE given by:

$$p_{\text{new}} = \exp\left(-\frac{\text{MSE}_{\text{new}} - \text{MSE}_{\text{old}}}{\text{alpha} \cdot T \cdot \text{MSE}_{\text{old}}}\right) \quad (88)$$

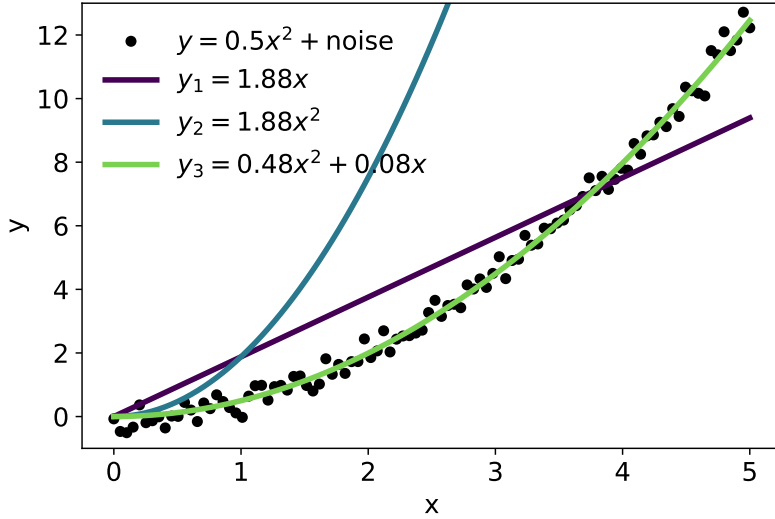


Figure 11: Toy data justifying new acceptance probability.

We demonstrate the effect of this new acceptance formula on a very simple toy example shown in Fig. 11. First we generate toy data for the function $y = 0.5x^2$ adding noise. To simulate a bad fit in the first round of the algorithm, we construct a linear function $y_1 = mx$. Assuming this linear function undergoes a constant optimization we get $m = 1.88$ with $\text{MSE} = 1.90$. In the next round this linear function is chosen for mutation. A very simple and likely mutation that would directly bring us to the correct shape, is to insert a node with the squaring operator to obtain $y_2 = 1.88x^2$. Not changing the constant however results in $\text{MSE} = 240$ and an acceptance probability of $10^{-3}\%$ according to Eq. (87) with `alpha` = 100. Therefore the algorithm is likely to reject the new and correct form of the function without a constant optimization even for a high `alpha`. The new acceptance probability would give 28 %, a much more promising result.

For data that is easy to describe, the inefficiency caused by p_{old} is balanced by a large amount of mutation attempts. For our toy model it could add a new branch to the tree with a quadratic term like $y_3 = 1.88x + bx^2$. According to p_{old} , the formula with for example $b = 0.4$ would be accepted with a probability of 47 %. If this function gets accepted and optimized we get $b' = 0.48$ and $m' = 0.08$ with $\text{MSE} = 0.077$ which is about as good as the MSE for $y = 0.5x^2$ due to the noise.

For our $f_B = 10$ data, for the process including the photon propagator, the amount of mutations needed to balance this inefficiency is beyond the limit of computer power available to us, so we have to rely on p_{new} . It makes sure that the algorithm can get out of local minimas more easily no matter what the absolute value of the MSE is. Choosing `alpha` = 100 gives for a relative difference of one order of magnitude an acceptance probability of 90 %, for 2 orders 37 % when $T = 1$.

With the new acceptance probability we obtain the hall of fame given in Tab. 6. Unlike with p_{old} we obtain formulas with higher complexity as expected. We choose the expression of complexity 73 and show it in the last panel of Fig. 12.

cmpl.	dof	function	MSE
16	5	$ax_p + bx_\eta - c(d - ex_p)^2$	$1.57 \cdot 10^{-2}$
22	6	$ax_p + bx_\eta - c(d - ex_p)^2 + f/x_p$	$9.46 \cdot 10^{-3}$
30	8	$(ax_p - b)/(cx_p^3 + d + e(x_p - x_\eta + f_p + g_p/x)/x) - h$	$3.82 \cdot 10^{-3}$
42	9	$(ax_p - b)/(cx_p^3 + d + e(x_p + f - (gx_\eta - h/x_p^2)/x)/(x_p + x_\eta/x_p)) - i$	$1.22 \cdot 10^{-3}$
45	8	$(x_p - a)/(bx_p^3 + c + d(x_p + e - f(x_\eta - g/x_p^2(x_p + y)))/x_p)/(x_p + x_\eta^2/x_p) - h$	$7.96 \cdot 10^{-4}$
47	10	$(x_p - a)/(bx_p^3 + c + d(x_p + e - f(x_\eta - g/(hx_p^2(x_p + x_\eta) + i)))/x_p)/(x_p + x_\eta^2/x_p) - j$	$6.71 \cdot 10^{-4}$
50	10	$(x_p - a)/(bx_p^3 + c + d(x_p + e - f(x_\eta - g/(hx_p^2(x_p + x_\eta^2 - x_\eta) + i)))/x_p)/(x_p + x_\eta^2/x_p) - j$	$6.03 \cdot 10^{-4}$
63	14	$(ax_p - b)/(cx_p^3 + d + e(x_p + f - g(hx_p^2 + x_\eta + i - j)/(kx_p^2(x_p + (x_\eta - l)^2 - x_\eta) + m)))/x_p)/(x_p + x_\eta^2/x_p) - n$	$5.64 \cdot 10^{-4}$
73	15	$(ax_p - b)/(cx_p^3 + d(x_p - e(fx_p^2 + x_\eta + g - h/(i(j - x_p)^2(x_p + x_\eta^2) + k)))/x_p)/(x_p + lx_\eta(mx_px_\eta + x_\eta)) + n) - o$	$1.44 \cdot 10^{-4}$

Table 6: Hall of fame for process with $f_B = 10$ including photon. MSE given for $t(x_p, x_\eta) \times 10$

Writing it in a more readable way:

$$t(x_p, x_\eta | f_B = 10) = \frac{ax_p - b}{cx_p^3 + d \frac{x_p - \frac{e}{x_p} \left(fx_p^2 + x_\eta + g - \frac{h}{i(j - x_p)^2(x_p + x_\eta^2) + k} \right)}{x_p + lx_\eta(mx_px_\eta + x_\eta)} + n - o \quad (89)$$

We can again simplify and redefine parameters:

$$t(x_p, x_\eta | f_B = 10) = \frac{x_p - b'}{c'x_p^3 + \frac{e'x_\eta + f'x_p^2 + g' - \frac{h'}{i'(j - x_p)^2(x_p + x_\eta^2) + 1}}{x_p^2 + lx_px_\eta^2(mx_p + 1)}} + n' - o \quad (90)$$

with

$$b' = \frac{b}{a} \quad c' = \frac{c}{a} \quad e' = e \frac{d}{a} \quad f' = \frac{d}{a}(1 - ef) \quad g' = e \frac{d}{a}g \quad h' = \frac{d}{a} \frac{h}{k} \quad i' = \frac{i}{k} \quad n' = \frac{n}{a}$$

This way we removed the parameters a , d and k having 12 degrees of freedom left.

Similar to what we did in Sec. 3.4.1 we can compare this result to some standard rational function fits. For the numerator we choose a polynomial of order 4 and for the denominator we try degrees 1 to 4. The results are shown in Fig. 12. The best performance is of course obtained for the largest amount of parameters. A fourth-order polynomial with 2 variables has 15 degrees of freedom. Putting the same kind of polynomial in the numerator as well we have one correlation that can be removed so that we have 29 parameters. For the other rational functions we respectively have 24, 20 and 17 free parameters. Comparing the MSE, PYSR is somewhere in between the $d = 2$ and $d = 3$ results. It fails to perfectly fit the upper edge like $d = 3$ or $d = 4$ though the amount of points in this area is not significant. The best of the rational functions has more than twice as many parameters but improves the MSE only by 30 %, the function with the $d = 3$ denominator has 8 more parameters and an improvement of only 10 %.

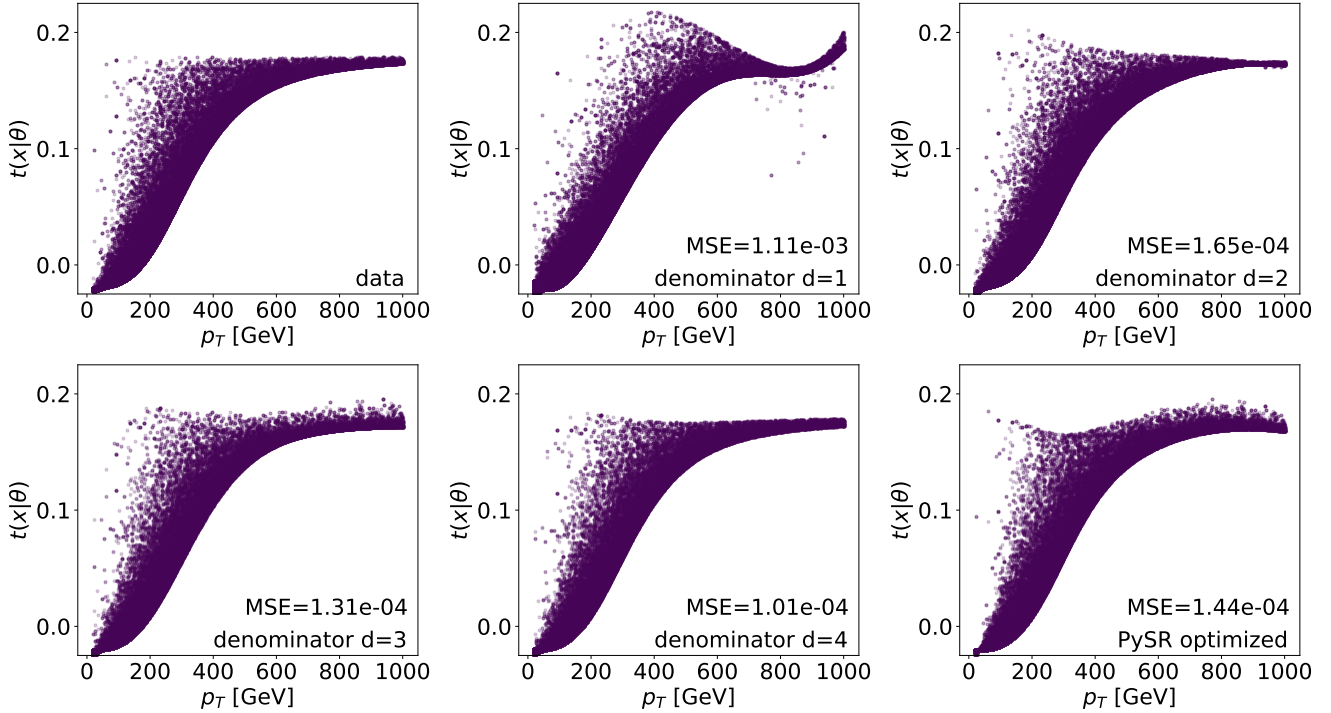


Figure 12: Score as a function of p_T for rational fits and the PySR output, including optimization, for the simplified ZH setup including photon propagator for $f_B = 10$. The numerator of the rational function fits is a fourth-order polynomial. MSE given for $t(x_p, x_\eta) \times 10$

To conclude, we first showed that an acceptance probability considering the relative improvement of the MSE is needed for an algorithm to be efficient enough to regress on complex data. With the new results we showed again that PySR gives compact functions with a minimal amount of free parameters which are favored over the long expressions of basic rational functions with a fixed polynomial order in numerator and denominator.

3.6 Two quark flavors

Finally, we include all incoming quarks and generate events for the full partonic process

$$pp \rightarrow ZH$$

In the distribution of the score we encounter a double branch structure shown in Fig. 13. We now have a discrete latent variable in our process which is the quark flavor. The Z -boson couples differently to up-type quarks than it does to down-type quarks, as the vector coupling has a dependence on the quark charge and isospin. Because of the photon contribution in the matrix element, which couples equally to all quarks, the Z -coupling does not cancel in the fraction of the score as it did in the case without photon in Eq. (79).

Comparing Fig. 10 where we only had u -quarks to 13, we see that for $f_B = 0$ the lower branch corresponds to up-type quarks and the upper to down-type quarks⁵. The situation is reversed for the case of $f_B = 10$.

⁵Due to the nature of the PDFs we mostly have u - and d -quarks in the data set, however c - and s -quarks can appear as well. Since particle charge and isospin are the only relevant quantities for the Z -coupling, we cannot distinguish between the quark generations.

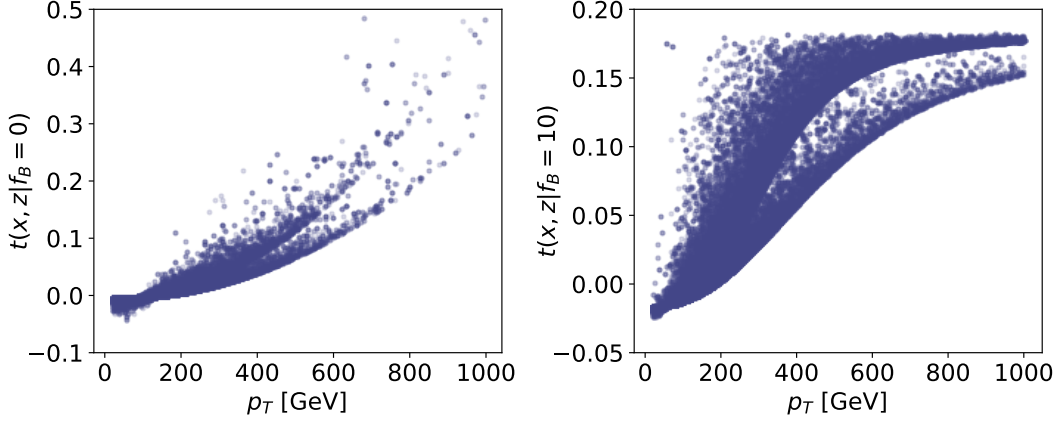


Figure 13: Joint score for ZH process including all quarks for $f_B = 0, 10$.

For the case of $f_B = 0$ only the interference term between SM and new physics is relevant for the score. The vector and axial-vector couplings for the different quark types are given by:

$$\begin{aligned}
 V_u &= \frac{1}{2} - \frac{8}{3} \sin^2 \theta_W & A_u &= \frac{1}{2} \\
 V_d &= -\frac{1}{2} + \frac{2}{3} \sin^2 \theta_W & A_d &= -\frac{1}{2}
 \end{aligned}$$

Therefore calculating $(V^2 + A^2)$ we find that this term is larger for d -quarks.

In the interference term, as discussed in the beginning of Sec. 3.5, the photon contribution is subtracted due to the relative minus sign of the vertex factors. Since the photon couples equally to all quarks, only the difference in the Z -contribution matters for the score. Hence, down-type quarks have a higher score.

As for the case of $f_B = 10$ where the quadratic term of the new physics contribution is dominant, the $2/f_B$ bound is reached once the interference term is negligible. As it is smaller for the up-type quarks, the score approaches the limit faster in this case.

In this chapter we want to investigate the performance of PySR for this double-branch structure. The expectation is to find a marginalized result that lies between the two branches.

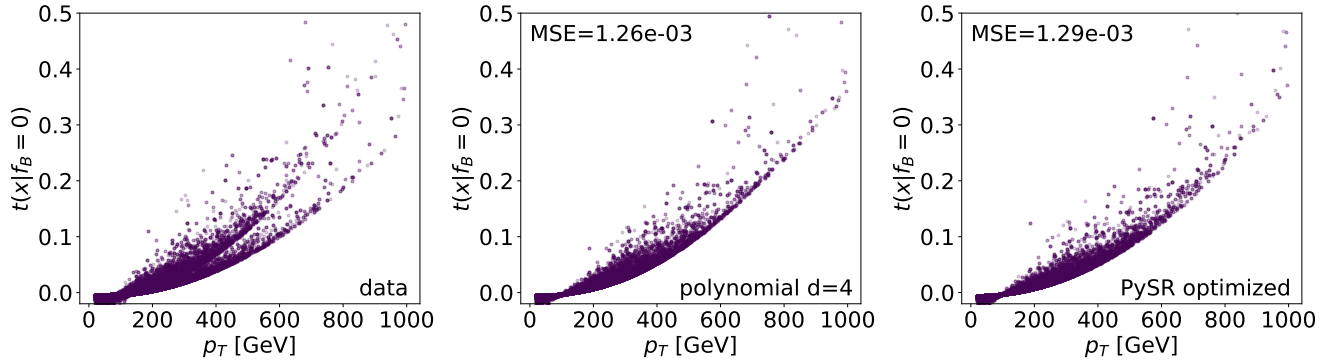


Figure 14: Score as a function of p_T for fourth-order polynomial fit and PySR including optimization, for the full partonic ZH setup for $f_B = 0$. MSE given for $t(x_p, x_\eta) \times 10$

3.6.1 Results for $f_B = 0$

The hall of fame is displayed in Tab. 7 of which the best result of complexity 29 is shown in the right panel of Fig. 14 next to the polynomial baseline. The two fits show a comparable performance. In this case we expect a higher MSE than in the previously discussed cases as due to the latent variable the MSE cannot go to 0 even in theory. In Fig. 14 we see that the fit lies approximately in between the two branches. However, due to the $|\eta_-|$ dependence the branches overlap so that it is difficult to see which data points belong to which branch and where the fit lies in between, i.e. is it in the middle or closer to one of the branches? For a deeper analysis we have to look at distributions of the two variables where we fix one of them. We combine this with a detailed study of the hall of fame.

From Tab. 7 we choose the functions of complexity 7, 9, 11, 13 and 29 for a detailed study. A display of these functions is shown in Fig. 15. The upper panels show the p_T -distribution for a fixed $|\eta_-|$ and the lower panels show the $|\eta_-|$ -distributions for 2 values of p_T .

We start with the simplest expression of complexity 7 consisting only of a squared term in p_T and a correlation term linear in p_T and $|\eta_-|$. It has a single free parameter. Looking at the upper two panels we see that the fit matches the data for small p_T but significantly undershoots both branches for larger values. The $|\eta_-|$ -distributions show a good agreement with the data only at low $|\eta_-|$ and low p_T . For a formula with low complexity this behavior is expected since these are the regions containing most of the data.

Clearly, one parameter cannot be enough to describe both the dependencies in p_T and in $|\eta_-|$. The additional parameter for complexity 9 comes with a significant improvement in the MSE. The prefactor of the p_T term is now ab and only a bit larger than the parameter a of the previous function, so there is no significant difference for data with $\eta_- = 0$ or low p_T values. The correlation term with its own constant is now squared in p_T instead of linear. This causes a significant improvement in the high p_T regions for large $|\eta_-|$ as can be seen in the upper right panel.

If the correlation term is additionally also squared in η_- , as it is for complexity 11, we observe an improvement in the lower left panel for low p_T regions. High p_T regions are barely affected by this, since the higher p_T , the smaller are the maximal values for η_- due to phase-space reasons discussed in Sec. 3.2. For values of $|\eta_-| < 1.5$ such a squared term doesn't win by a lot against a linear one. Since the improvement affects only few events the MSE changes only very insignificantly

cmpl.	dof	function	MSE
7	1	$ax_p(x_p + x_\eta)$	$a = 0.0375$ $6.51 \cdot 10^{-3}$
9	2	$ax_p^2(x_\eta + b)$	$a = 0.0203$ $ab = 0.0406$ $4.35 \cdot 10^{-3}$
11	2	$ax_p^2(x_\eta^2 + b)$	$a = 0.0111$ $ab = 0.0462$ $4.32 \cdot 10^{-3}$
13	3	$ax_p^2 + bx_px_\eta^2 - c$	$a = 0.0648$ $b = 0.0088$ $c = 0.0625$ $1.96 \cdot 10^{-3}$
17	4	$ax_p^2 + bx_px_\eta^2 - cx_\eta - d$	$1.84 \cdot 10^{-3}$
19	4	$ax_p^2 + bx_px_\eta^2 - cx_p - dx_\eta$	$1.74 \cdot 10^{-3}$
21	5	$ax_p^2 + bx_px_\eta^2 - cx_p - dx_\eta + e$	$1.72 \cdot 10^{-3}$
27	6	$ax_p^2 + bx_px_\eta^2 - cx_px_\eta - dx_p + ex_\eta + f$	$1.63 \cdot 10^{-3}$
28	7	$ax_p^2(bx_\eta - c)^2 + dx_px_\eta^2 + ex_p^2 - fx_p - gx_\eta$	$1.43 \cdot 10^{-3}$
29	8	$ax_p^2 + b(x_\eta^2 + c)(x_\eta(dx_p - e)(x_p - f) + x_p + g) - h$	$1.29 \cdot 10^{-3}$

Table 7: Hall of fame for full partonic ZH setup with $f_B = 0$. All values are given for $t(x_p, x_\eta) \times 10$

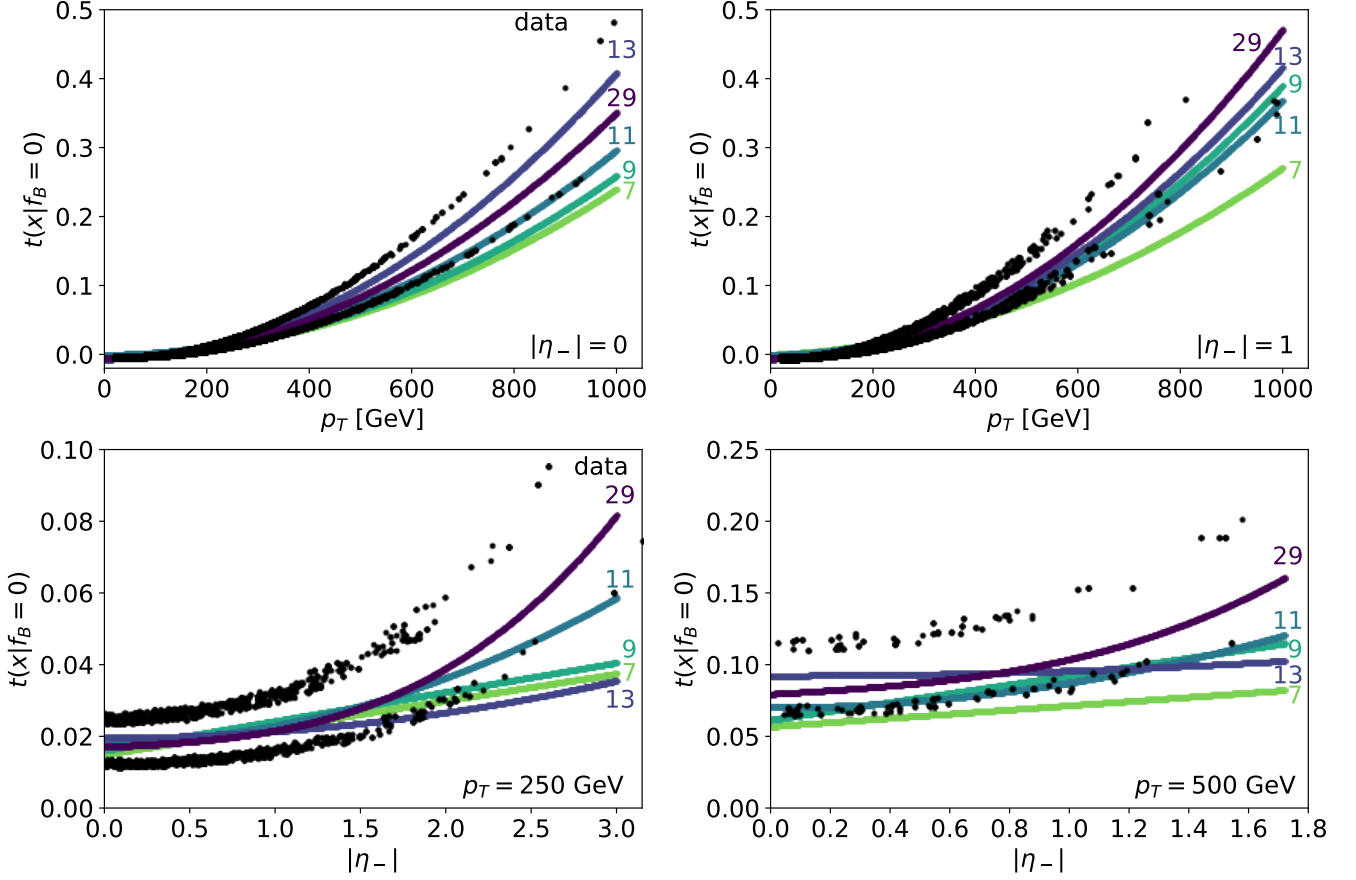


Figure 15: Sliced kinematic distributions for the score of the full partonic ZH process with $f_B = 0$, showing the HoF results of Tab. 7.

between complexity 9 and 11.

For the following equations the tree structure changes and the bracket we had in the previous expressions "opens up". This is important as the squared p_T -term is not simultaneously used for both the pure p_T dependence and the correlation term. For complexity 13 this allows the correlation term to be linear in p_T and squared in $|\eta_-|$. Additionally we have a constant term appearing for the first time. This constant is necessary for a good parametrization of the score due to the constant contribution from the total cross section. Comparing to complexity 11 the prefactor of the squared p_T term is about 1.5 times larger. As a result we observe a good agreement with the data in the p_T distributions. While all previous fits were very close or even below the lower branch, complexity 13 lies in between the branches with a tendency towards the upper branch. However, in the $|\eta_-|$ -distribution the new equation seems to be less than optimal and loses against complexity 11. But since there is a lot more data in regions with high p_T and low $|\eta_-|$ than there is in regions with high $|\eta_-|$ and low p_T complexity 13 performs better in the overall MSE than the previous one.

With growing complexity we see how the equation is being "fine-tuned" by gradually adding linear terms in $|\eta_-|$ and p_T to improve the fit. The tree structure changes again at complexity 28 and makes room for more efficient and compact parametrizations. The function of complexity 29 consists of several brackets being multiplied with each other rather than a sum of terms. Such a structure allows for more correlation terms while keeping the amount of parameters low. Expanding the brackets we mostly find the same terms as we saw for low complexities. A new and

	Z only	Z and γ	all quarks
MSE	$7.68 \cdot 10^{-5}$ ($7.65 \cdot 10^{-5}$)	$3.05 \cdot 10^{-5}$	$1.30 \cdot 10^{-3}$ ($1.29 \cdot 10^{-3}$)
a	0.136316(29)	0.025668(21)	0.038766(15)
b	0.148896(27)	0.048981(17)	0.062109(84)
c	0.12362(21)	0.19081(51)	0.1537(25)
d	0.019829(88)	0.008893(27)	$-1.33(78) \cdot 10^{-4}$
e	0.1552(25)	0.6110(32)	-8.0(24)
f	$5.02(92) \cdot 10^{-4}$	$1.9(11) \cdot 10^{-7}$	0.01385(25)
g	0.959(69)	167(47)	0.364(20)
h	0.022549(28)	0.0078760(65)	0.00470(11)
i	0.24413(51)	0.48588(52)	0.375(14)

Table 8: Coefficients for PYSR result of Eq. (86) from one quark setup including photon propagator optimized on all 3 data sets. The MSE of the original optimized PYSR result of the data set is given in brackets for comparison. All values for $t(x_p, x_\eta) \times 10$.

important correlation term $x_\eta^3 x_p^2$ appears. We already discussed its significance in Sec. 3.4.1, it drastically improves the fit for phase-space regions with large $|\eta_-|$. We see how this parametrization fits even the phase-space areas with a small amount of points. In all 4 panels the fit lies exactly between the branches. Moreover, this formula is consistent with the ones obtained in all previous studies of $f_B = 0$ data, showing that there is no conceptual difference in the dependencies on the kinematic variables of the three considered processes. The variations caused by the different coupling constants can be absorbed by the free parameters. To illustrate that we choose the result of one of the considered cases and optimize it for all three data sets. We choose the analytic expression obtained from the one-quark process including the photon propagator in Eq. (86) because of its simple form. Applying it on the one-quark process with Z only and the two quark process we obtain the coefficients given in Tab. 8. We also take the MSE of this fit and compare it to the MSE corresponding to the original result of the data set. The differences between the MSE are negligible proving that indeed one expression is enough to describe all 3 data sets.

Such a comparison of different PYSR results can be helpful to point out differences and similarities of two parametrizations. Because PYSR is based on a random construction of trees, we will never obtain the exact result twice and for complex expression the similarities cannot always be recognized at first glance. Moreover, as we are working with data sets where we cannot reduce the error to 0 there is no global minimum in the function space, i.e. there is no single expression that describes the data perfectly, but there are many that do it sufficiently well.

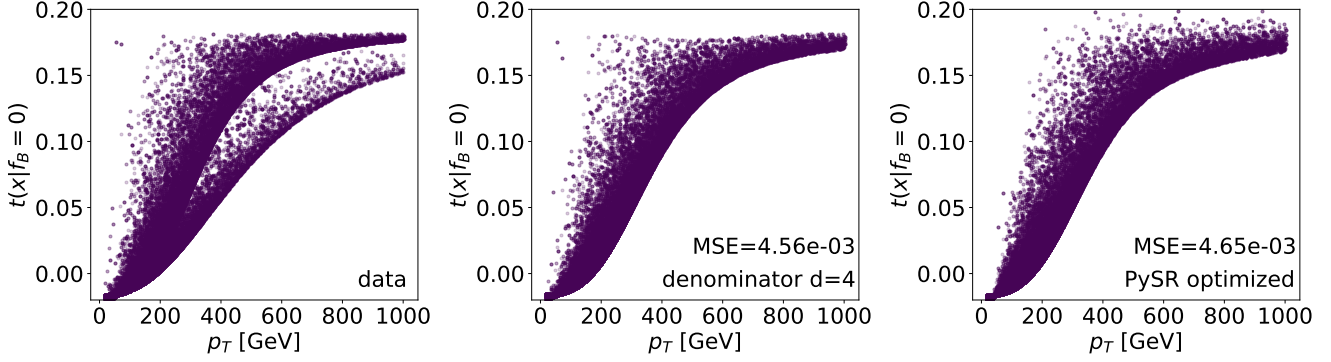


Figure 16: Score as a function of p_T for fourth-order polynomial fit and PySR including optimization, for the full partonic ZH setup for $f_B = 10$. MSE given for $t(x_p, x_\eta) \times 10$

3.6.2 Results for $f_B = 10$

In analogy to the previous chapter we will discuss the hall of fame results for $f_B = 10$. Unlike for the previous case, this data does not have a simple leading term p_T^2 that accounts for most of the data. It is dominated by the limited growth behavior which is difficult to describe without the use of exponential functions. We therefore do not expect useful expressions from the low complexity results. The hall of fame results are shown in Tab. 9 and Fig. 17.

Starting with the p_T -distribution of the upper panels we quickly see that all equations up to complexity 23 are essentially linear in p_T with many of them overlapping. Of course a linear function is by no means sufficient to describe this data. Even though a x_p^2 term would improve the fit in regions at about $p_T < 400$ GeV it would drastically overshoot the limit for large p_T and is therefore not considered by PySR.

Also in the $|\eta_-|$ -distributions the low complexity results show poor fits and barely any improvement among each other. The correlation term $x_p^4 x_\eta$ appearing in the denominator for complexity 16, 23 and 25 is supposed to implement the limited growth behavior. In the MSE we see its impact in the improvement from complexity 10 to 16 but its expressivity is too limited to show a visible effect.

From complexity 16 to 25 the MSE shows only minor improvements from the insertion of linear and cubed x_η terms. A significant improvement we see only once the polynomial in the denominator consists of many terms including higher orders to ensure the cut off. Now that we have a more efficient limitation in the function, a cubed contribution in p_T can be added to the

cmpl.	dof	function		MSE
10	3	$ax_p + bx_\eta^3 - c$	$a = 0.3487 \quad b = 0.0043 \quad c = 0.3492$	$1.61 \cdot 10^{-2}$
16	4	$ax_p - b/(cx_p^4 x_\eta + d)$	$a = 0.3032 \quad b = 0.0960 \quad c = 0.0213 \quad d = 0.3033$	$1.26 \cdot 10^{-2}$
20	4	$ax_p - b/(cx_p^5 x_\eta + d)$	$a = 0.2860 \quad b = 0.0942 \quad c = 0.0117 \quad d = 0.3005$	$1.21 \cdot 10^{-2}$
23	5	$ax_p + bx_\eta^3 - c/(dx_p^4 x_\eta + e)$		$1.19 \cdot 10^{-2}$
25	7	$ax_p + bx_\eta^3 + cx_\eta - d/(ex_p^4(x_\eta + f) + g)$		$1.14 \cdot 10^{-2}$
45	12	$ax_p + bx_\eta - c(x_p - d)^3 + e - f/(gx_p^3 x_\eta^3 - x_\eta(hx_p + i) + j(x_p + k)^6 + l)$		$4.65 \cdot 10^{-3}$
51	13	$ax_p + bx_\eta - c(x_p - d)^3 + e - f/(gx_p^3 x_\eta^3 - x_\eta(h + i) + j(x_p + k)^6 + l + m/x_p)$		$4.65 \cdot 10^{-3}$

Table 9: Hall of fame for the complete ZH setup with $f_B = 10$. MSE is given for $t(x_p, x_\eta) \times 10$.

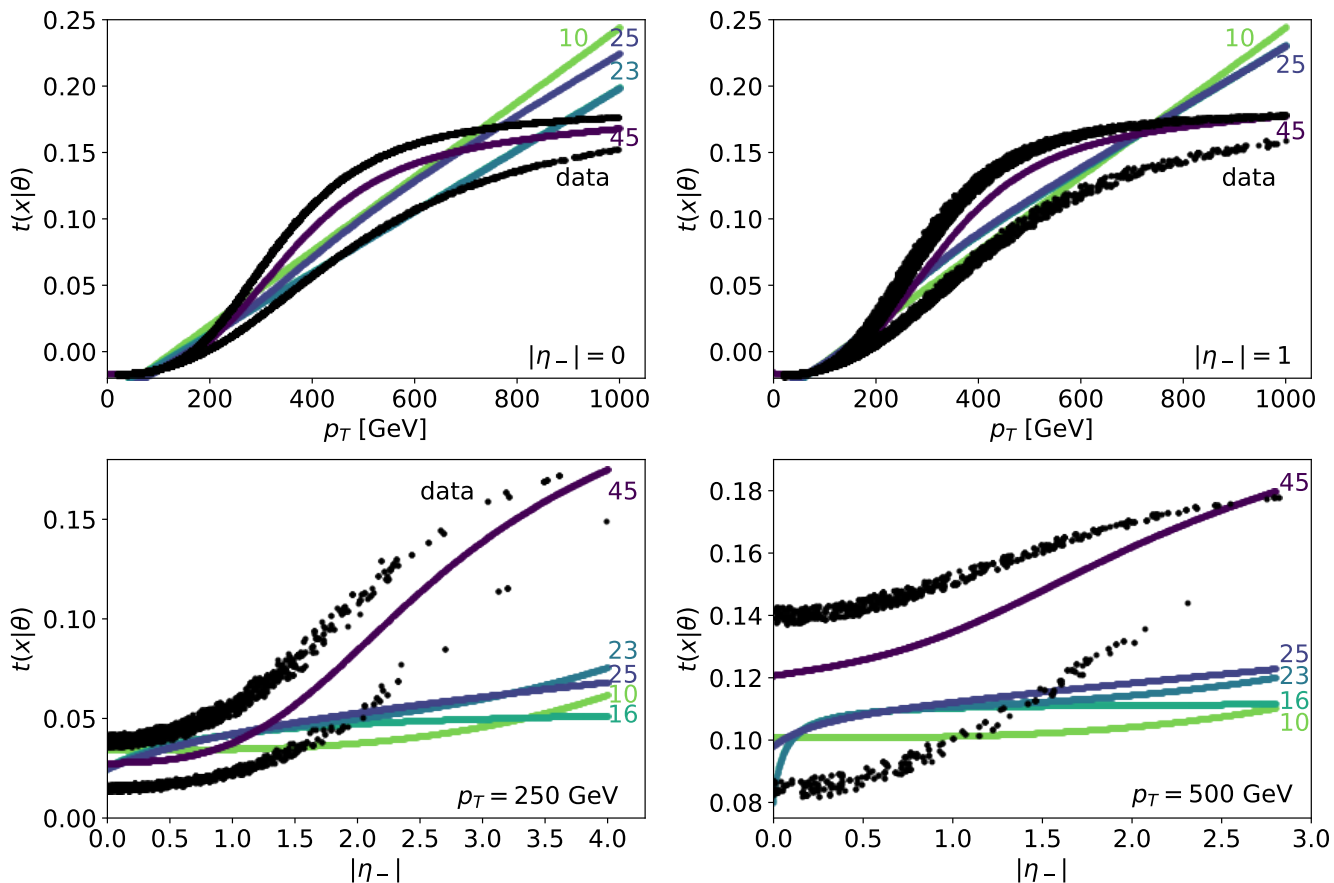


Figure 17: Sliced kinematic distributions for the score of the full partonic ZH process with $f_B = 10$, showing the HoF results of Tab. 9.

polynomial part of the equation. This matches the low p_T data better than it was possible with the linear term. We see in Fig. 17 how the function for complexity 45 describes most of the data well. A slight overshoot in areas of high p_T and high $|\eta_-|$ can be observed. However, as we can see from Tab. 9, adding more terms into the denominator for complexity 51 does not improve the MSE and since only few points are affected by this overshoot we accept the result obtained here. Note, that in the case of one quark flavor we had an expression of complexity 73, a much more complex result.

From Fig. 17 we can clearly see that the upper branch is more populated than the lower one. This comes from the fact that we have more u -quarks in a pp collision than d -quarks. Therefore it is reasonable that the fit will be closer to the upper branch, though it is hard to tell by eye where exactly it should be. Essentially, this issue corresponds to finding out where the theoretical limit of the MSE lies. A convenient way to do so is to introduce a proxy described in the following.

We create a 2d histogram in the two kinematic variables. The value of each bin is given by the average score of the data points in the bin. Such a histogram is shown in the left panel of Fig. 18. For each bin we can calculate the MSE between the score of each data point and the average score. The average over all bins would give us a theoretical limit for the MSE.

Of course this result depends highly on the amount of bins we choose: too many bins will result in bins containing only one data point leading to an error of 0. Having too few bins we do not only average over events in two branches for the same point in phase-space, but also over a larger phase-space which will result in a too large MSE. In order to decide which amount of

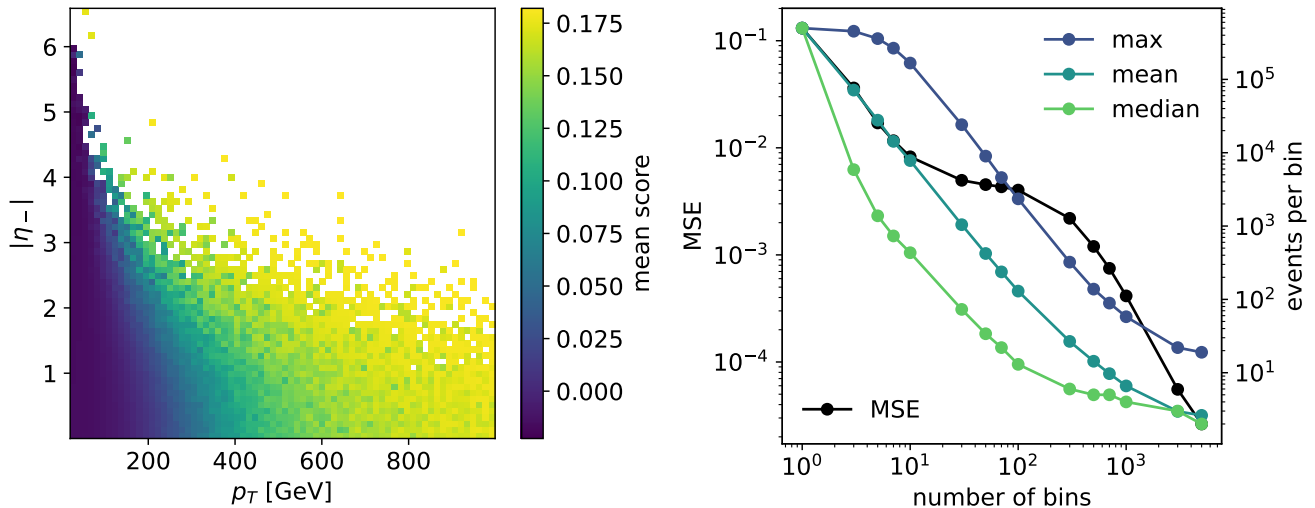


Figure 18: Left: Mean score for 70×70 bins. Right: average MSE between mean score of bin and score in bin compared to amount of events per bin.

bins to choose we plot the average MSE in dependence of the bin amount, while keeping track of the maximal, mean and median amount of events per bin. This is displayed in the right panel of Fig. 18. In the area of 30-100 bins the MSE does not change significantly so we deduce that any amount of bins within this interval is reasonable.

Since the gap between the two branches is different for different regions in p_T , the MSE of the proxy should be compared individually for different p_T -intervals. In Tab. 10 we show the MSE for the rational function fit with denominator and numerator of degree 4, the PySR result of complexity 45 and the proxy for 50 and 100 bins. Our fit results are comparable to the 50 bin proxy. PySR is slightly worse for very large p_T .

To conclude, in this case we found that the double branch structure is easier to fit than it was in the case of one quark. Due to the lower branch, the high p_T region is smeared out and the theoretical score limit is reached for less events loosening the conditions of the cutoff requirement for the fit.

	denom. d=4	PySR	proxy (50 b.)	proxy (100 b.)
$p_T < 250$ GeV	1.11e-03	1.20e-03	1.18e-03	1.01e-03
$250 < p_T < 500$ GeV	4.61e-02	4.62e-02	4.54e-02	4.17e-02
$500 < p_T < 750$ GeV	3.71e-02	3.72e-02	3.52e-02	2.98e-02
$p_T > 750$ GeV	1.22e-02	1.31e-02	1.11e-02	8.12e-03

Table 10: Comparing MSE of proxy to rational function fit and PySR result for different p_T bins

4 Weak Boson Fusion

In this section we are going beyond our toy model of the ZH -production and look at weak-boson-fusion instead. Having 3 particles in the final state, this process is fundamentally more difficult to describe due to the increased amount of degrees of freedom. We are going to apply the same methods to obtain the optimal observable for a dimension-6 Wilson coefficient and connect this analysis to the question of CP-violation in the VVH -interaction. As before, we start by introducing the process and studying the data at parton level. Afterwards we add detector effects for a realistic analysis and conclude with a comparison of confidence intervals for different observables.

4.1 Feynman rules, matrix element, CP-observable

Assuming only u - and d -quarks in the initial state the Feynman diagram is given by:

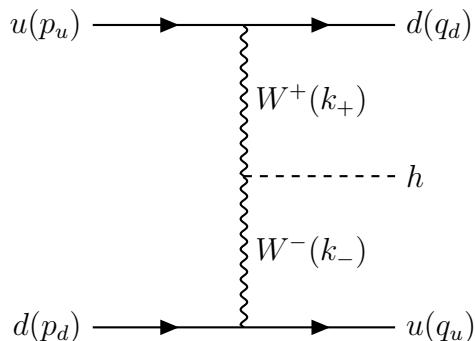


Figure 19: Feynman diagram for weak boson fusion

The process of weak-boson-fusion is dominated by the contribution of the W -propagators. Therefore, we choose the WWH -coupling for a detailed study. In this chapter we additionally want to investigate the CP-violation, which is why the operator of our choice from Tab. 1 is the following:

$$\mathcal{O}_{W\widetilde{W}} = -(\phi^\dagger\phi)\widetilde{W}_{\mu\nu}^k W^{\mu\nu k} = -(\phi^\dagger\phi)\epsilon_{\mu\nu\rho\sigma} W^{\rho\sigma k} W^{\mu\nu k} \quad (91)$$

Feynman rules

In analogy to Sec. 3.1 we start by deriving the Feynman rule for the vertex. We insert the expressions for the Higgs field from Eq. (9) and the field strength tensor from Eq. (3) and keep only terms contributing to the WWH -coupling:

$$\mathcal{L}_{EFT} \supset -\frac{f_{W\widetilde{W}}}{\Lambda^2}(\phi^\dagger\phi)\epsilon_{\mu\nu\rho\sigma} W^{\rho\sigma k} W^{\mu\nu k} \quad (92)$$

$$\supset -\frac{f_{W\widetilde{W}}v}{\Lambda^2}\epsilon_{\mu\nu\rho\sigma}(\partial^\rho W^{\sigma k} - \partial^\sigma W^{\rho k})(\partial^\mu W^{\nu k} - \partial^\nu W^{\mu k})h \quad (93)$$

Expanding the brackets yields:

$$= -\frac{f_{W\widetilde{W}}v}{\Lambda^2}\epsilon_{\mu\nu\rho\sigma}(\partial^\rho W^{\sigma k}\partial^\mu W^{\nu k} - \partial^\sigma W^{\rho k}\partial^\mu W^{\nu k} - \partial^\rho W^{\sigma k}\partial^\nu W^{\mu k} + \partial^\sigma W^{\rho k}\partial^\nu W^{\mu k})h \quad (94)$$

We rewrite the expression in terms of physical fields using $W^\pm = (W^1 \pm iW^2)/\sqrt{2}$. We neglect the W^3 -contribution in this calculation for simplicity, but note that they will give rise to ZZH , AAH and ZAH couplings.

$$\supset -\frac{f_{W\widetilde{W}}^v}{\Lambda^2}\epsilon_{\mu\nu\rho\sigma}(k_+^\rho W^{+\sigma}k_-^\mu W^{-\nu} + k_-^\rho W^{-\sigma}k_+^\mu W^{+\nu} - k_+^\sigma W^{+\rho}k_-^\mu W^{-\nu} - k_-^\sigma W^{-\rho}k_+^\mu W^{+\nu} - k_+^\rho W^{+\sigma}k_-^\nu W^{-\mu} - k_-^\rho W^{-\sigma}k_+^\nu W^{+\mu} + k_+^\sigma W^{+\rho}k_-^\nu W^{-\mu} + k_-^\sigma W^{-\rho}k_+^\nu W^{+\mu})h \quad (95)$$

$$= -4\frac{f_{W\widetilde{W}}^v}{\Lambda^2}\epsilon_{\mu\nu\rho\sigma}k_+^\sigma k_-^\rho W^{+\mu}W^{-\nu}h \quad (96)$$

In the last step we renamed indices and combined terms, taking into account that the Levi-Civita tensor is fully anti-symmetric. We use k_\pm for the momenta of W^\pm . Together with the SM contribution from Sec. 1.1 the vertex factor can be written as:

$$iL_{\mu\nu} = 2i\frac{m_W^2}{v}g_{\mu\nu} - 4i\frac{f_{W\widetilde{W}}^v}{\Lambda^2}\epsilon_{\mu\nu\rho\sigma}k_+^\sigma k_-^\rho \quad (97)$$

Matrix element

The matrix element is given by:

$$i\mathcal{M} = \bar{u}(q_d)\Gamma^\mu u(p_u) \left(-i\frac{g_{\mu\alpha} + \frac{k_{+\mu}k_{+\alpha}}{m_W^2}}{k_+^2 - m_W^2} \right) iL^{\alpha\beta} \left(-i\frac{g_{\nu\beta} + \frac{k_{-\nu}k_{-\beta}}{m_W^2}}{k_-^2 - m_W^2} \right) \bar{u}(q_u)\Gamma^\nu u(p_d) \quad (98)$$

where p denotes the momenta of the incoming and q the outgoing quarks. The $qq'W$ vertex is given by:

$$\Gamma^\mu = -i\frac{g}{2\sqrt{2}}\gamma^\mu(1 - \gamma^5)V_{qq'} \quad (99)$$

with $V_{qq'}$ being the corresponding CKM matrix entry. We can easily show that terms proportional to k_\pm will drop out in the limit of massless fermions:

$$\begin{aligned} \bar{u}(q_d)k_{+\mu}\gamma^\mu(1 - \gamma^5)u(p_u) &= \bar{u}(q_d)\not{k}_+(1 - \gamma^5)u(p_u) \\ &= \bar{u}(q_d)(\not{p}_u - \not{q}_d)(1 - \gamma^5)u(p_u) \\ &= -\bar{u}(q_d)\not{q}_d(1 - \gamma^5)u(p_u) + \bar{u}(q_d)(1 + \gamma^5)\not{p}_u u(p_u) \\ &= -\bar{u}(q_d)m_d(1 - \gamma^5)u(p_u) + \bar{u}(q_d)(1 + \gamma^5)m_u u(p_u) \end{aligned} \quad (100)$$

We inserted $k_+ = p_u - q_d$ and made use of the Dirac equation in the last step. Since both terms in the final expression are proportional to the quark masses they can be neglected. The simplified matrix element can thus be written as:

$$i\mathcal{M} = -i\frac{(\bar{u}(q_d)\Gamma^\mu u(p_u))L_{\mu\nu}(\bar{u}(q_u)\Gamma^\nu u(p_d))}{(k_-^2 - m_W^2)(k_+^2 - m_W^2)} \quad (101)$$

For the matrix element squared we evaluate the traces of the γ -matrices separately. For the first term we obtain:

$$\sum \left(\bar{u}(q_d)\Gamma^{*\mu}u(p_u) \right) \left(\bar{u}(p_u)\Gamma^{\mu'}u(q_d) \right) = \frac{g^2}{8}V_{ud}^2\text{Tr} \left[\not{q}_d\gamma^\mu(1 - \gamma^5)\not{p}_u\gamma^{\mu'}(1 - \gamma^5) \right] \quad (102)$$

$$= \frac{g^2}{4} V_{ud}^2 q_{d\lambda} p_{u\lambda'} \text{Tr} \left[\gamma^\lambda \gamma^\mu \gamma^{\lambda'} \gamma^{\mu'} (1 - \gamma^5) \right] \quad (103)$$

$$= g^2 V_{ud}^2 q_{d\lambda} p_{u\lambda'} \left[g^{\lambda\mu} g^{\lambda'\mu'} + g^{\lambda'\mu} g^{\lambda\mu'} - g^{\lambda\lambda'} g^{\mu\mu'} + i\epsilon^{\lambda\mu\lambda'\mu'} \right] \quad (104)$$

and similarly for the second term:

$$\sum \left(\bar{u}(q_u) \Gamma^{*\nu} u(p_d) \right) \left(\bar{u}(p_d) \Gamma^{\nu'} u(q_u) \right) \quad (105)$$

$$= g^2 V_{ud}^2 q_{u\kappa} p_{d\kappa'} \left[g^{\kappa\nu} g^{\kappa'\nu'} + g^{\kappa'\nu} g^{\kappa\nu'} - g^{\kappa\kappa'} g^{\nu\nu'} + i\epsilon^{\kappa\nu\kappa'\nu'} \right] \quad (106)$$

The SM contribution to the WWH -vertex is given by:

$$L_{\mu\nu}^* L_{\mu'\nu'} \Big|_{\text{SM}} = 4 \frac{m_W^4}{v^2} g_{\mu\nu} g_{\mu'\nu'} \quad (107)$$

The prefactor of 4 cancels with the averaging over the initial spins. Taking all the ingredients together and contracting the indices we end up with the following matrix element squared:

$$|\overline{\mathcal{M}}|^2 \Big|_{\text{SM}} = g^4 V_{ud}^4 \frac{m_W^4}{v^2} \frac{(q_d \cdot q_u)(p_u \cdot p_d)}{(k_-^2 - m_W^2)^2 (k_+^2 - m_W^2)^2} \quad (108)$$

For the interference term the Lorentz structure is given by:

$$L_{\mu\nu}^* L_{\mu'\nu'} \Big|_{\text{interference}} = -16 \frac{m_W^2 f_W \widetilde{W}}{\Lambda^2} g_{\mu\nu} \epsilon_{\mu'\nu'\rho\sigma} k_+^\sigma k_-^\rho \quad (109)$$

For the squared matrix element we obtain:

$$|\overline{\mathcal{M}}|^2 \Big|_{\text{interference}} = -8 g^4 V_{ud}^4 \frac{m_W^2 f_W \widetilde{W}}{\Lambda^2} \frac{((p_d \cdot p_u) + (q_u \cdot q_d))}{(k_-^2 - m_W^2)^2 (k_+^2 - m_W^2)^2} \epsilon_{\mu\nu\rho\sigma} p_d^\mu p_u^\nu q_d^\rho q_u^\sigma \quad (110)$$

To compute the approximate joint score for this single process we take the ratio of the interference and the SM term and get:

$$t(x, z | f_W \widetilde{W} = 0) \approx -8 \frac{v^2}{m_W^2 \Lambda^2} \frac{(p_d \cdot p_u) + (q_u \cdot q_d)}{(q_d \cdot q_u)(p_u \cdot p_d)} \epsilon_{\mu\nu\rho\sigma} p_d^\mu p_u^\nu q_d^\rho q_u^\sigma \quad (111)$$

CP-observable

By contracting four 4-momenta with the Levi-Civita tensor we can construct a pseudo-scalar that is odd under parity transformation and time reversal. It is therefore a CP-violating observable [30]. In our case of weak boson fusion there is only one way to obtain such a quantity:

$$\epsilon_{\mu\nu\rho\sigma} p_d^\mu p_u^\nu q_d^\rho q_u^\sigma \quad (112)$$

This term is included in the expression for the score in Eq. (111) which is expected from a CP-violating operator.

In order to evaluate this expression we are going to adapt the notation from [30] by assigning the initial and final momenta to the positive and negative hemisphere defined by the positive and negative z-direction (beam axis). We therefore enforce the restriction that for both initial

and final states there is one quark in each hemisphere. The quark indices are replaced by $+$ and $-$ to indicate their hemisphere rather than their quark flavor. For the initial momenta we have $p_{\pm} = (E_{\pm}, 0, 0, \pm E_{\pm})$.

To evaluate the Levi-Civita tensor we remember that it is fully anti-symmetric so that terms with two equal indices give 0 and the exchange of two indices leads to a sign flip. We write:

$$\begin{aligned} \epsilon_{\mu\nu\rho\sigma} p_+^{\mu} p_+^{\nu} q_+^{\rho} q_-^{\sigma} &= \epsilon_{1423} (-E_+ E_- q_{+x} q_{-y}) + \epsilon_{1432} (-E_+ E_- q_{+y} q_{-x}) \\ &\quad + \epsilon_{4123} (E_+ E_- q_{+x} q_{-y}) + \epsilon_{4132} (E_+ E_- q_{+y} q_{-x}) \end{aligned} \quad (113)$$

$$= (-E_+ E_- q_{+x} q_{-y}) - (-E_+ E_- q_{+y} q_{-x}) - (E_+ E_- q_{+x} q_{-y}) + (E_+ E_- q_{+y} q_{-x}) \quad (114)$$

$$= 2E_+ E_- (q_{+y} q_{-x} - q_{+x} q_{-y}) \quad (115)$$

We insert the relations $q_{\pm x} = q_{\pm T} \cos \phi_{\pm}$ and $q_{\pm y} = q_{\pm T} \sin \phi_{\pm}$ and recognize the trigonometric identity:

$$= 2E_+ E_- q_{+T} q_{-T} (\sin \phi_+ \cos \phi_- - \cos \phi_+ \sin \phi_-) \quad (116)$$

$$= 2E_+ E_- q_{+T} q_{-T} \sin(\phi_+ - \phi_-) = 2E_+ E_- q_{+T} q_{-T} \sin \Delta\phi \quad (117)$$

The CP-violating observable therefore primarily depends on the azimuthal angle between the two outgoing quarks as well as their transverse momenta and the energies of the initial quarks. Inserting this result into the score of Eq. (111) we obtain:

$$t(x, z | f_{W\widetilde{W}} = 0) \approx -8 \frac{v^2}{m_W^2 \Lambda^2} \frac{2E_+ E_- + (q_+ \cdot q_-)}{(q_+ \cdot q_-)} q_{+T} q_{-T} \sin \Delta\phi \quad (118)$$

At this point we want to emphasize, that the score is the optimal observable of the CP-violating Wilson coefficient $f_{W\widetilde{W}}$ and not CP-violation in general. Therefore it is not simply given by the expression in Eq. (117) but includes additional momenta dependencies. Therefore we expect $\sin \Delta\phi$ to be the leading term of the optimal observable and additional non-trivial polynomial of the jet momenta for a more accurate fit.

4.2 Score for $f_{W\widetilde{W}}$

Unlike for the process in Sec. 3 where we separately looked at the contributions of the different propagators, in this section we will directly look at the full partonic process

$$pp \rightarrow qq'H \quad (119)$$

which includes the previously mentioned Feynman diagrams with ZZH -, AAH - and ZAH -vertices.

The kinematic distributions are shown in the upper panels of Fig. 20. For the case of the Standard Model with $f_{W\widetilde{W}} = 0$ the distribution in $\Delta\phi$ is symmetric while for $f_{W\widetilde{W}} = 1$ we can see the predicted sine-shape. For the momentum distribution we again see an increase of events in the tail for a finite Wilson coefficient due to the momentum dependence in the operator. $p_{T,1}$ corresponds to the quark with higher energy, the distribution for $p_{T,2}$ is similar. We also show the distribution in $\Delta\eta$ which indicates only a small difference between the two histograms. It is unclear at this point, if the slight increase in events at lower $\Delta\eta$ for higher $f_{W\widetilde{W}}$ is meaningful or just an artifact from looking at 1-dimensional histograms and includes correlations with p_T . To

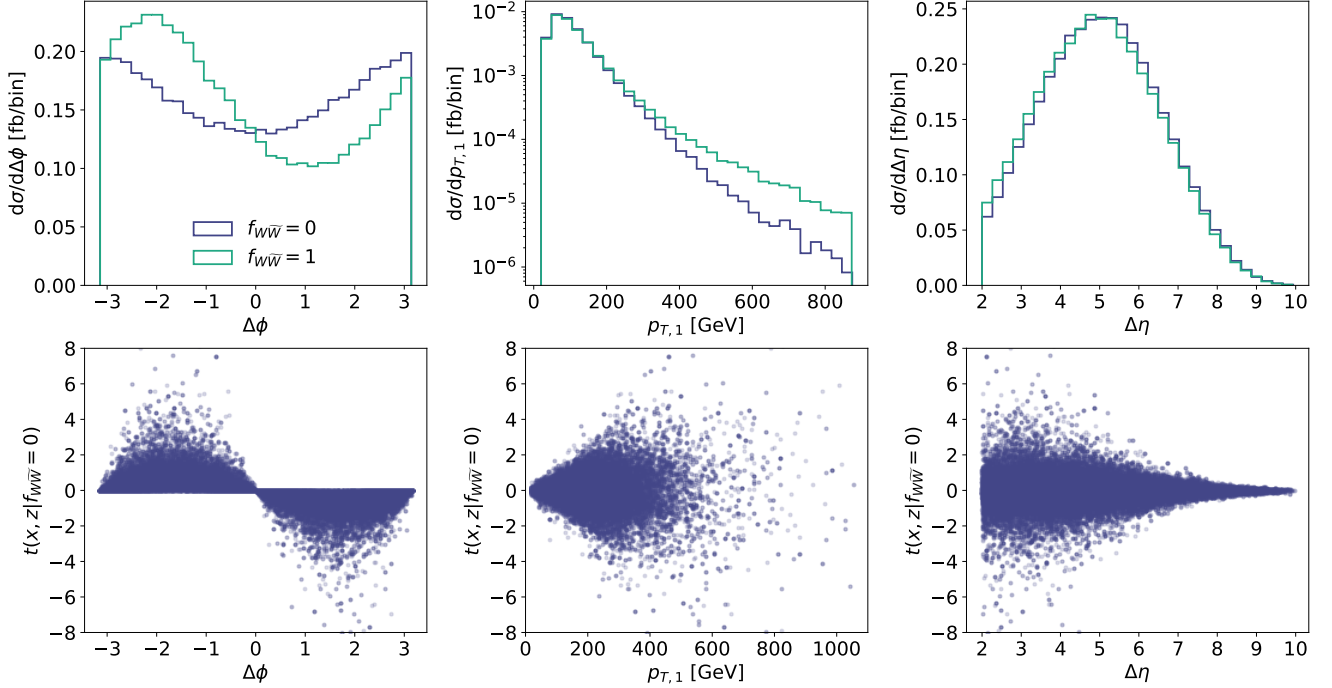


Figure 20: Top: Kinematic distributions for the WBF process at parton level for $f_{W\widetilde{W}} = 0, 1$. Bottom: Joint score in dependence of different kinematic observables.

examine this further, we would have to look at multi-dimensional histograms similar to Sec. 3.2 and keep other variables fixed. However, due to the increased amount of degrees of freedom in this process, such an analysis would be quite involved and is unnecessary. We will show later in Sec. 4.3 that including $\Delta\eta$ in the symbolic regression gives better observables.

The joint score for the SM is shown in the lower panels of Fig. 20. Most prominently, we see the predicted dependence on $\sin \Delta\phi$. Events with a high score appear at large p_T values and small $\Delta\eta$ as suggested by the 1-dimensional histograms.

In the next step we want to see how and why the score distribution changes for larger Wilson coefficients. Firstly, we compare the $\Delta\phi$ distribution of the score for $f_{W\widetilde{W}} = 0, 0.5, 1$. This is shown in the upper panels of Fig. 21. The evolution here is quite interesting: starting with a sine, the score turns into a spoon. Between $f_{W\widetilde{W}} = 0$ and $f_{W\widetilde{W}} = 0.5$ we see that in the latter points with positive score appear in the region of positive $\Delta\phi$, an unpopulated area in the case of the SM. Here, one can still clearly recognize the sine shape, a trait that is almost completely lost for the Wilson coefficient of 1.

To understand this peculiar behavior we have to remember the general score behavior from Tab. 2. We also need to consider that while the interference term is proportional to $\sin \Delta\phi$, the quadratic term is symmetric and proportional to $\sin^2 \Delta\phi$. In areas where the quadratic term dominates, the score is bound by $2/\theta$. We can clearly see this effect in the area of negative $\Delta\phi$ where the upper bound for $f_{W\widetilde{W}} = 0.5$ is at 4 and for $f_{W\widetilde{W}} = 1$ at 2.

For positive $\Delta\phi$ the balance between the negative interference term and the positive quadratic term determines the value of the score. We separate the events in 3 categories: $t(\theta) < 0$, $0 < t(\theta) < 2/\theta$ and $t(\theta) > 2/\theta$. The first case of negative score values corresponds to the situation of the SM. The quadratic term in this case is small, its absolute value is smaller than the one of the interference term so their sum stays negative. However, when the quadratic term is larger than the interference term, the score is positive and above the $2/\theta$ bound. If the quadratic term

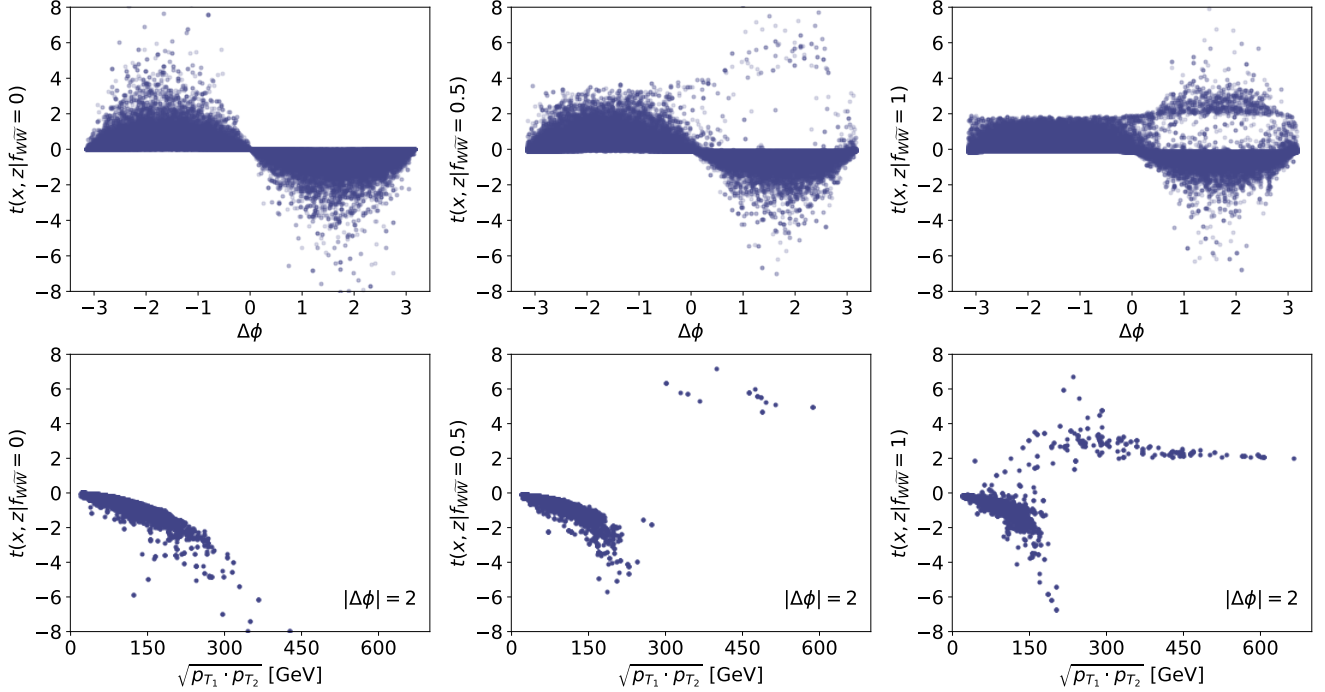


Figure 21: Top: Joint score for $\Delta\phi$ for different Wilson coefficients. Bottom: Joint score for geometric mean of the jet p_T for fixed $\Delta\phi$.

grows further it is bound by the $2/\theta$ limit from below. This can be clearly seen for $f_{W\widetilde{W}} = 1$ when looking at the lower plots of Fig. 21. They depict the geometric mean of the transverse momenta for a fixed value of $\Delta\phi$. We know that the quadratic term grows with p_T faster than the interference term so we see how the score converges to the limit. For 0.5 the quadratic term is not large enough to win against the interference term, therefore the limit is not reached in this case.

A small amount of points can be seen in the area of positive score below the limit. They correspond to small phase-space areas where the quadratic term is comparable to the interference, but results in a positive score contribution.

We also notice that the maximal p_T value for the negative score events is decreasing with growing Wilson coefficient which is once more because of the rapid growth of the quadratic term with p_T .

4.3 Symbolic regression at parton level

In this section we will apply symbolic regression on the full partonic process without shower or detector effects. We are going to examine the score in the case of the Standard Model as well as for the more complicated case of $f_{W\widetilde{W}} = 1$. Here we focus on analyzing the functional form, the operators and kinematic observables needed for a good MSE.

4.3.1 Results for $f_{W\widetilde{W}} = 0$

We start by providing PYSR with the kinematic observables we expect from the considerations in the previous chapter as well as from Eq. (118). We use the observables:

$$\{ x_{p,1}, x_{p,2}, \Delta\phi \} \quad \text{with} \quad x_{p,i} = \frac{p_{T,i}}{m_H}, \quad (120)$$

compl	dof	function	MSE
3	1	$a\Delta\phi$	1.30e-01
4	1	$a\sin(\Delta\phi)$	1.03e-01
5	1	$a\Delta\phi x_{p,1}$	9.93e-02
6	1	$a\sin(\Delta\phi)x_{p,1}$	4.94e-02
8	1	$a\sin(\Delta\phi)x_{p,1}x_{p,2}$	1.49e-02
10	2	$a\sin(\Delta\phi)x_{p,1}(x_{p,2} + b)$	1.40e-02
15	4	$a(b\Delta\phi^2 + x_{p,1})(x_{p,2} + c)(\sin(\Delta\phi) - d)$	1.36e-02

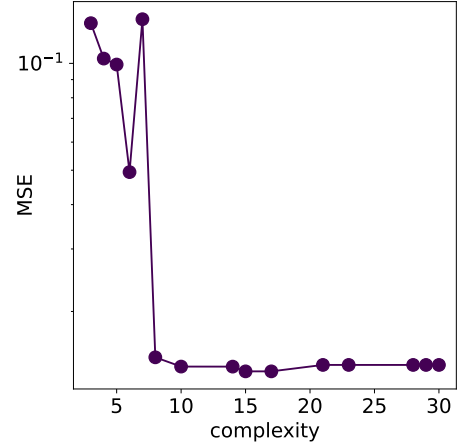


Table 11: Hall of fame for WBF Higgs production at parton level with $f_{W\widetilde{W}} = 0$.

For the operators we allow \sin , $+$, $-$, \times , square and cube. We set $\alpha=1.5$ and the maximal complexity to 30.

The hall of fame is shown in Tab. 11. We see that the largest drop in the MSE occurs once the algorithm learns the term $\sin\Delta\phi x_{p,1}x_{p,2}$ which was predicted by Eq. (118). No significant improvement happens for larger complexities. Since the MSE is still rather large we deduce that more information, i.e. more kinematic observables, is needed for a better fit. We add $\Delta\eta$ and obtain the hall of fame given in Tab. 12. The results of low complexity are similar to the ones obtained previously, but complexity 16 and 28 include $\Delta\eta$ and improve the MSE by almost a factor of 2. For further analysis we will use the result of complexity 16:

$$t(x_{p,1}, x_{p,2}, \Delta\phi, \Delta\eta | f_{W\widetilde{W}} = 0) = -x_{p,1}(x_{p,2} + c)(a - b\Delta\eta)\sin(\Delta\phi + d)$$

$$\text{with } a = 1.086(11) \quad b = 0.10241(19) \quad c = 0.24165(20) \quad d = 0.00662(32) \quad (121)$$

The constant d is small ensuring that $t \propto \sin\Delta\phi$.

compl	dof	function	MSE
3	1	$a\Delta\phi$	$1.30 \cdot 10^{-1}$
4	1	$\sin(a\Delta\phi)$	$2.75 \cdot 10^{-1}$
5	1	$a\Delta\phi x_{p,1}$	$9.93 \cdot 10^{-2}$
6	1	$-x_{p,1}\sin(\Delta\phi + a)$	$1.90 \cdot 10^{-1}$
7	1	$(-x_{p,1} - a)\sin(\sin(\Delta\phi))$	$5.63 \cdot 10^{-2}$
8	1	$(a - x_{p,1})x_{p,2}\sin(\Delta\phi)$	$1.61 \cdot 10^{-2}$
14	2	$x_{p,1}(a\Delta\phi - \sin(\sin(\Delta\phi)))(x_{p,2} + b)$	$1.44 \cdot 10^{-2}$
15	3	$-(x_{p,2}(a\Delta\eta^2 + x_{p,1}) + b)\sin(\Delta\phi + c)$	$1.30 \cdot 10^{-2}$
16	4	$-x_{p,1}(a - b\Delta\eta)(x_{p,2} + c)\sin(\Delta\phi + d)$	$8.50 \cdot 10^{-3}$
28	7	$(x_{p,2} + a)(bx_{p,1}(c - \Delta\phi) - x_{p,1}(d\Delta\eta + ex_{p,2} + f)\sin(\Delta\phi + g))$	$8.18 \cdot 10^{-3}$

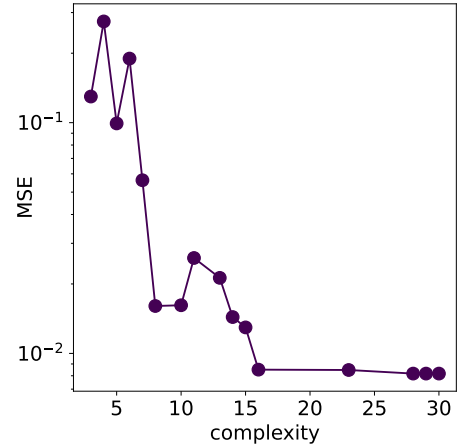


Table 12: Hall of fame for WBF Higgs production at parton level with $f_{W\widetilde{W}} = 0$ including $\Delta\eta$.

cmpl.	dof	function	MSE
3	1	$ax_{p,+}$	0.140
10	1	$a(-s_\phi + x_{p,+})/\Delta\eta^2$	0.111
12	2	$a + (-s_\phi + x_{p,+})/(b - s_\phi^3)$	0.080
15	3	$a + (bs_\phi + x_{p,+})/(c - s_\phi)$	0.074
17	4	$(as_\phi + x_{p,+})/(-s_\phi + b + c/\Delta\eta) - d$	0.066
23	3	$(-s_\phi x_{p,+}/x_{p,\times} + x_{p,+})/(a\Delta\eta^2/x_{p,\times} - s_\phi + b) - c$	0.063
29	7	$as_\phi - bs_\phi/(cs_\phi + dx_{p,\times}^4 + e/x_{p,\times}) + fx_{p,+} - g$	0.049
39	9	$as_\phi + bx_{p,+} - c(s_\phi + d)/(e/(fx_{p,+} + x_{p,+}) + x_{p,\times}^4/\Delta\eta^4) + g + (h - ix_{p,\times})/\Delta\eta$	0.039
49	7	$(as_\phi^3 + b(s_\phi^3 + x_{p,+}))/c\Delta\eta^2/(s_\phi^2(-\Delta\eta x_{p,\times} + x_{p,\times}^3 - dx_{p,\times}(\Delta\eta/(s_\phi + e/s_\phi) + s_\phi - x_{p,+}))) - fs_\phi + x_{p,\times}) + g$	0.040

Table 13: Hall of fame for WBF Higgs production with $f_{W\widetilde{W}} = 1$ using $x_{p,\times}$, $x_{p,+}$, s_ϕ and $\Delta\eta$

4.3.2 Results for $f_{W\widetilde{W}} = 1$

Judging by the complex shape of the score distribution for $f_{W\widetilde{W}} = 1$, it is going to be far more challenging for PYSR to find a good analytic expression for the optimal observable. Taking the same operators and variables that showed good and comprehensive results in the case of the SM does not lead to an equally successful result anymore. An adaptation of the input data is necessary to ease the regression process.

First of all, we implement the division operator to enable the fit of boundaries. Secondly, the dependence on $\sin \Delta\phi$ is not as apparent anymore as it was in the case of the SM (see Sec. 4.2). By providing the sine operator, the algorithm starts building equations not only with $\Delta\phi$ as argument, but also $\Delta\eta$ or p_T which we know is not justified from a physics perspective and is a waste of computational time and energy. Instead, we remove the sine operator and directly provide $\sin \Delta\phi$ as variable.

Additionally we deduce from symmetry considerations that any observable should depend equally on the two outgoing quark momenta. Alternatively to providing the quark momenta separately, we use the product and the sum of the momenta as symmetrized input. The new parameter basis is therefore given by:

$$\left\{ x_{p,\times} = \frac{\sqrt{p_{T,1}p_{T,2}}}{m_H}, x_{p,+} = \frac{p_{T,1} + p_{T,2}}{m_H}, s_\phi = \sin \Delta\phi, \Delta\eta \right\} \quad (122)$$

The corresponding hall of fame is given in Tab. 13 with its last two members shown in the lower panels of Fig. 22. Interestingly, even though the result of complexity 39 has a comparable MSE

cmpl.	dof	function	MSE
3	1	$ax_{p,\times}$	0.124
12	2	$ax_{p,\times}/(x_{p,\times}/\Delta\eta + \Delta\eta + b)$	0.116
15	2	$(s_\phi + a)(-s_\phi + x_{p,\times} - b)/(-s_\phi + x_{p,\times} + \Delta\eta/x_{p,\times})$	0.054
26	4	$a/(b - (s_\phi - c - d/(s_\phi^2 - s_\phi\Delta\eta - s_\phi/x_{p,\times} + ex_{p,\times}^2))/x_{p,\times})$	0.048
31	7	$a/(b - (s_\phi + (cs_\phi^2 - d)/(es_\phi^2x_{p,\times}^2 - s_\phi\Delta\eta + f) - g)/x_{p,\times})$	0.039

Table 14: Hall of fame for WBF Higgs production with $f_{W\widetilde{W}} = 1$ using $x_{p,\times}$, $\Delta\phi$ and $\Delta\eta$

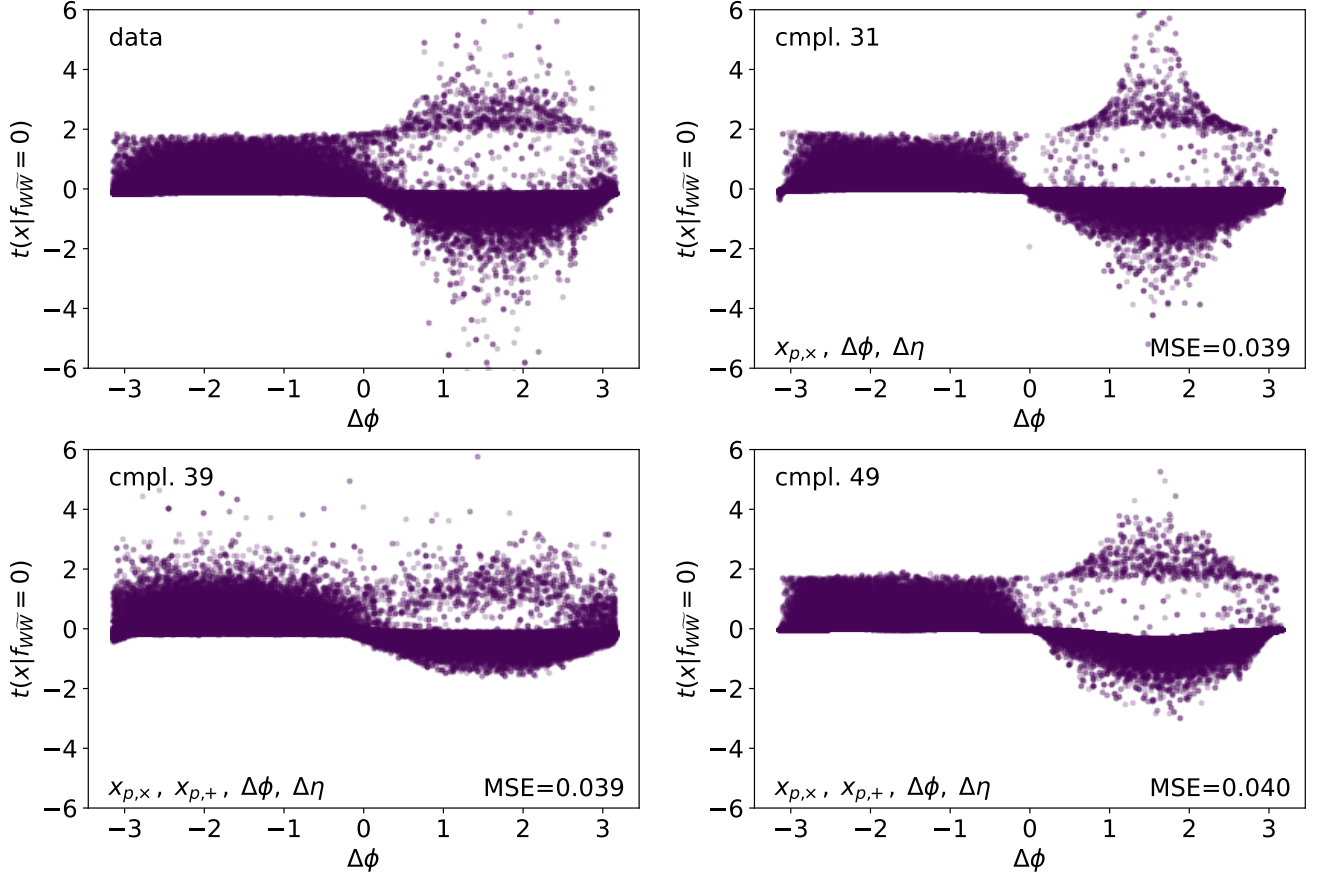


Figure 22: PYSR results for different complexities and input variables. Two lower panels correspond to Tab. 13, upper right fit corresponds to Tab. 14.

to the result of complexity 49, the shapes of the distributions are entirely different. In the simpler expression most of the prominent features and boundaries are lost, while they remain very clearly for complexity 49. The complicated expression appears to fit data with a high score better than the bulk of events at low score values. The simpler expression on the other hand, fits the low scored events better than the ones defining the limits, so that overall the two fits have comparable results.

In the next step we want to check if the sum of the momenta is a necessary quantity for the fit. We run PYSR with the exact same settings, but without $x_{p,+}$, again and obtain the hall of fame given in Tab. 14 with its last member shown in the upper right panel of Fig. 22. This function has the same MSE as the previously discussed ones while having a lower expressivity. The limits are preserved in this fit as well. The only distinct discrepancy is the narrow peak in the region of high score values affecting only an insignificant amount of data points. Due to the lower complexity and the smaller parameter basis we consider this a better result than the previous ones including $x_{p,+}$. Writing it out in its expanded and readable form it is given by:

$$t(p_{\times}, s_{\phi}, \Delta\eta | f_{W\tilde{W}} = 1) = \frac{a' x_{p,\times} (e' s_{\phi}^2 x_{p,\times} - s_{\phi} \Delta\eta - f')}{(b' x_{p,\times} + s_{\phi} - g') (e' s_{\phi}^2 x_{p,\times} - s_{\phi} \Delta\eta - f') - c' s_{\phi}^2 - d'}$$

with $a' = 0.75$ $b' = 0.38$ $c' = 4.2$ $d' = 4.6$ $e' = 1.1$ $f' = 0.26$ $g' = 0.21$ (123)

Technically we should have obtained this result already in the first run as the needed parameter

basis is a subset of the given one. This shows us that the performance of the PYSR algorithm scales badly with the amount of provided variables due to the random tree construction process. In theory this can be balanced by increasing the amount of populations and individuals.

4.4 Detector effects

In this section we want to investigate the effects of shower and detector simulations on the optimal observable. We expect that since the underlying physics is the same as for the partonic process, the functional form will not change significantly. Referring to Sec. 2.1 we use PYTHIA8 for shower simulation and DELPHES for a fast detector simulation including the anti- k_T jet clustering algorithm [31]. We exclude initial state radiation for simplicity. MADMINER still extracts the parton level information to obtain the joint score while for the fitting procedure only the final state observables are available to us.

The main influence of the detector on the data is the addition of noise that has a bad impact on the PYSR convergence. While in the case of $f_{W\widetilde{W}} = 0$ we obtain a similar expression to Eq. (121), the situation is more critical for $f_{W\widetilde{W}} = 1$. We find that instead of applying PYSR directly on the noisy data, it is convenient to take the previously obtained expressions from parton level data and perform an optimization fit on the detector level data.

In Tab.15 we compare the parameters from the fit optimizer on detector and parton level process. The change among the parameters is statistically significant but does not affect the general picture. The parameters stay within the same orders of magnitude and the relations among them are conserved. For instance, for $f_{W\widetilde{W}} = 0$ the parameter d is still small and the relation of a/b is comparable. With c being a constant added to a jet momenta it is most sensitive to the noise and therefore shows the largest discrepancy among the parameters.

Comparing the MSE we see that for $f_{W\widetilde{W}} = 0$ the detector noise increases the MSE by almost a factor of 2. However, for $f_{W\widetilde{W}} = 1$ the change is less than 10 %. Additionally the parameter changes are less significant than for the previous case. This indicates that the parametrization found for $f_{W\widetilde{W}} = 1$ already had large discrepancies with the parton level data and that these discrepancies are comparable to those induced by detector noise.

We conclude by saying that the functional form of a well understood PYSR result for parton level data can easily be optimized on the detector level data without loss of important information.

$f_{W\widetilde{W}} = 0$ Eq.(121)	parton level	detector	pull	$f_{W\widetilde{W}} = 1$ Eq.(123)	parton level	detector	pull
				a'	0.7490(14)	0.8792(31)	93.0
a	1.086(11)	0.9264(20)	14.5	b'	0.37800(94)	0.4160(19)	40.4
b	0.10241(19)	0.08387(35)	97.6	c'	4.218(18)	3.526(31)	38.4
c	0.24165(84)	0.3542(20)	134.0	d'	4.598(18)	4.759(32)	8.9
d	0.00662(32)	0.00911(67)	7.75	e'	1.1271(26)	1.0950(48)	1.2
MSE	$8.50 \cdot 10^{-3}$	$1.51 \cdot 10^{-2}$		f'	-0.2638(49)	-0.2325(68)	6.4
				g'	0.2063(19)	0.2057(34)	0.3
				MSE	$3.89 \cdot 10^{-2}$	$4.15 \cdot 10^{-2}$	

Table 15: Detector effect on the scores for WBF Higgs production, for fixed functional forms derived at parton level.

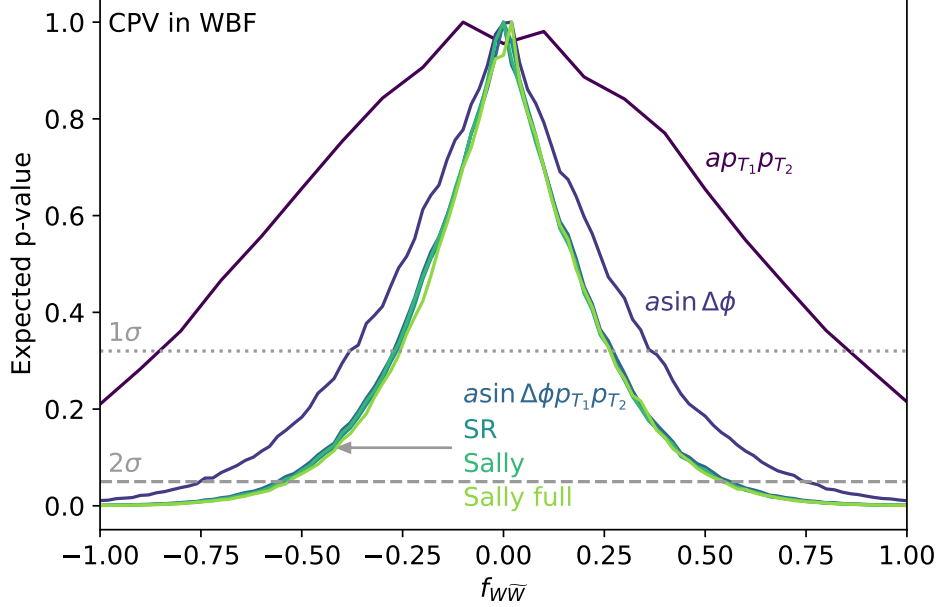


Figure 23: Confidence intervals for different score approximations or candidate observables assuming $f_{W\widetilde{W}} = 0$. SALLY uses the observables $p_{T,1}$, $p_{T,2}$, $\Delta\phi$ and $\Delta\eta$. SALLY full uses 18 kinematic observables describing the full event.

4.5 Exclusion limits

In the last section we are going to investigate the confidence intervals for different (optimal) observables. We establish a connection between the previously considered MSE and the 1σ exclusion limit. To that end we use MADMINER to compute the log-likelihood ratios and extract p -values as described in Sec. 2.2 for the process including detector effects assuming $f_{W\widetilde{W}} = 0$ and an integrated LHC luminosity of 139 fb^{-1} . We start by considering the observables

$$a_1 p_{T,1} p_{T,2} \quad a_2 \sin \Delta\phi \quad a_3 p_{T,1} p_{T,2} \sin \Delta\phi, \quad (124)$$

with $a_1 = -8.32(89) \cdot 10^{-7}$, $a_2 = -0.37370(94)$, and $a_3 = -5.5386(49) \cdot 10^{-5}$ and compare the results to the SR expression of Eq. (121). Finally, we compare these results to two neural nets using SALLY. For the first we use the same 4 observables as we did for symbolic regression. To prove that more kinematic observables do not provide additional physical information we train

(optimal) observable	MSE				reach	
	all	$ t(f_{W\widetilde{W}}) = 0.1 \dots 0.5$	$ t(f_{W\widetilde{W}}) > 0.5$	weighted	1σ	2σ
$ap_{T_1}p_{T_2}$	0.1576	0.0645	1.144	0.298	[-0.86,0.86]	—
$a \sin \Delta\phi$	0.0885	0.0163	0.680	0.223	[-0.38,0.36]	[-0.76,0.74]
$a \sin \Delta\phi p_{T_1}p_{T_2}$	0.0217	0.0076	0.163	0.056	[-0.28,0.28]	[-0.56,0.56]
SR Eq.(121)	0.0145	0.0059	0.103	0.031	[-0.26,0.26]	[-0.54,0.54]
SALLY	0.0129	0.0051	0.092	0.030	[-0.26,0.26]	[-0.56,0.54]
SALLY full	0.0048	0.0031	0.026	0.014	[-0.26,0.26]	[-0.54,0.54]

Table 16: MSE and exclusion limits for different approximations of the score or candidate optimal observable. The different scenarios correspond to Fig. 23.

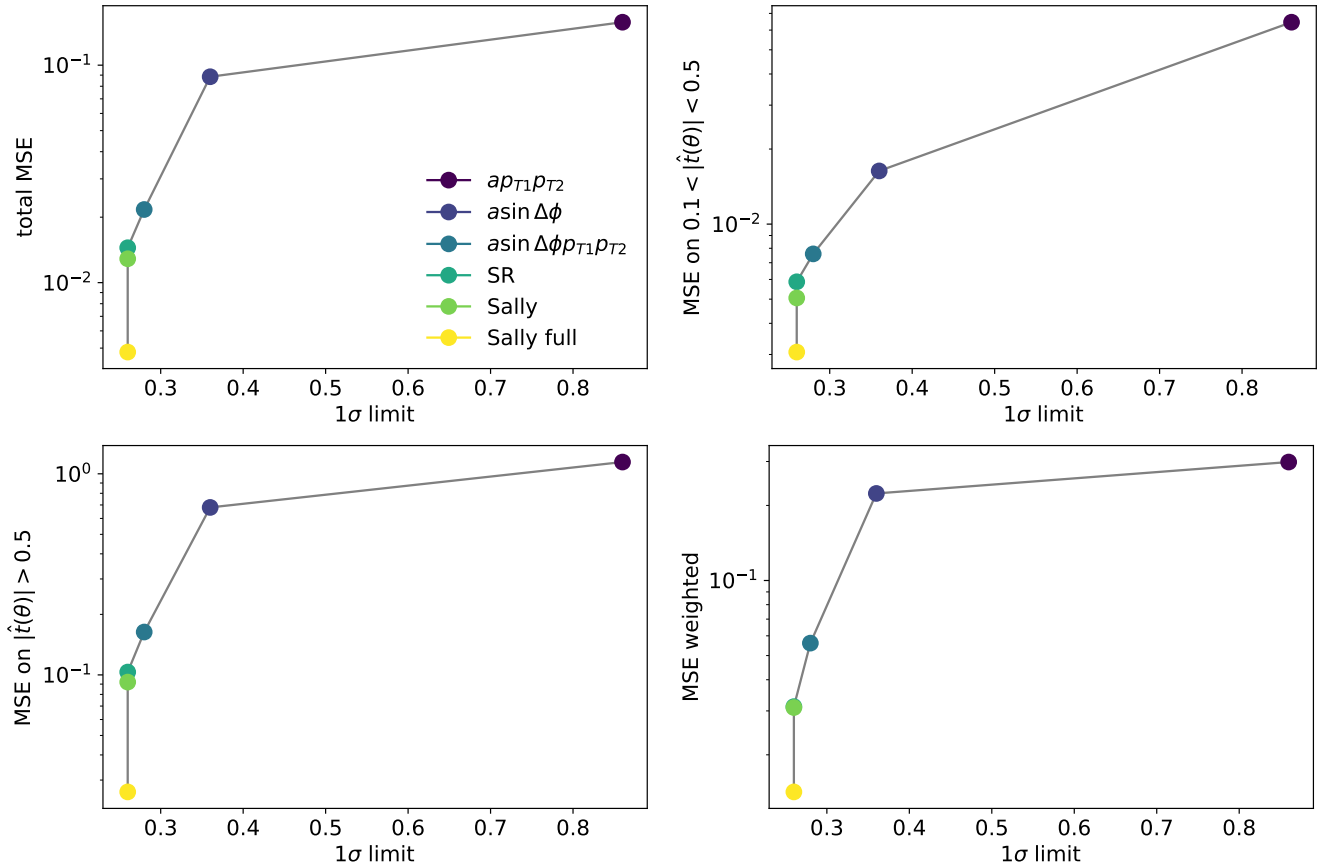


Figure 24: Scaling of the expected exclusion limits with the MSE for the different evaluations from Tab. 16.

a second neural net using 18 kinematic observables describing the full event. We call it "SALLY full". The limits are shown in Fig. 23.

As we expect the observable $ap_{T,1}p_{T,2}$ does not provide a good exclusion limit due to the absence of the most important kinematic observable $\Delta\phi$. For $a \sin \Delta\phi$ the confidence interval shrinks significantly. Combining the two expressions leads to further improvement. However, the symbolic regression result from Eq. (121), including additionally a dependence on $\Delta\eta$, shows only a minor improvement compared to $ap_{T,1}p_{T,2} \sin \Delta\phi$. SALLY and "SALLY full" give the same result marking the best limit that can be reached in theory.

Looking at the overall MSE in Tab. 16 it becomes apparent that a perfect description of all score areas does not necessarily lead to an improvement in the confidence intervals. We see that both SALLY networks significantly improve the MSE with respect to the PYSR result, but the σ -limits do not change at all. This is because a large amount of events is at low score values that do not influence the measurement of $f_{W\widetilde{W}}$ but have a significant impact on the MSE. Events with a high score are more sensitive to the Wilson coefficient but since they are very rare, they do not have a significant influence on the MSE.

We therefore evaluate the MSE for intermediate $t(f_{W\widetilde{W}}) = 0.1 \dots 0.5$ and for large score values $t(f_{W\widetilde{W}}) > 0.5$. We also consider a weighted MSE of the form:

$$\text{MSE}_{\text{weighted}} = \frac{1}{n} \sum_{i=1}^n g_i(x) (g_i(x) - t_i(x, z|\theta))^2 \quad (125)$$

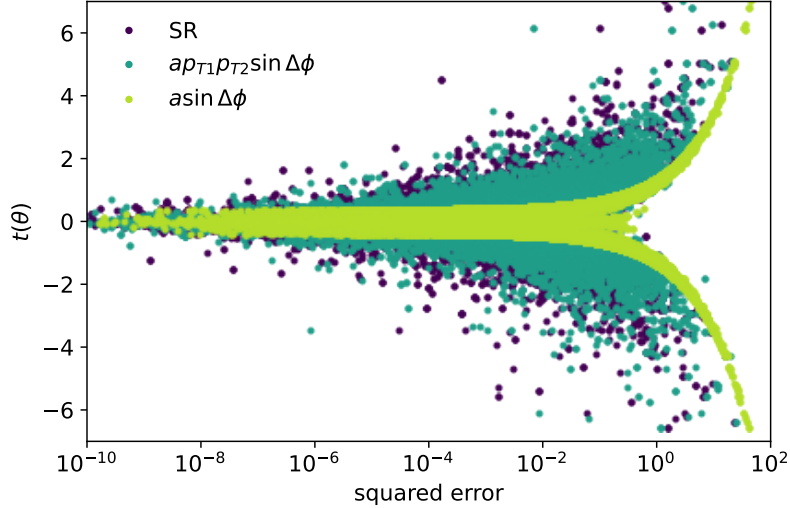


Figure 25: Squared error compared for different observables.

The results are listed in Tab. 16 and shown in Fig. 23. The overall MSE, the MSE for large score events as well as the weighted MSE all show similar behavior. In general, the limits improve with the MSE but a quantitative connection can not be derived. However, the MSE on intermediate score values appears to decrease linearly with the limits until it reaches the plateau.

For a more intuitive understanding we can look at the squared error for the PYSR result, $ap_{T,1}p_{T,2} \sin \Delta\phi$ and $a \sin \Delta\phi$ shown in Fig. 25. Between the last two parametrizations we can see a significant improvement in the squared error for intermediate and high score events. This is precisely the reason for the improved confidence intervals. In comparison, the difference between the PYSR result and $ap_{T,1}p_{T,2} \sin \Delta\phi$ is much smaller and therefore does not have a significant impact on the confidence intervals. The SALLY results show a similar behavior.

To sum up, we saw that our symbolic regression result is indeed better than conventional observables from theory approximations. However in this example, over complicated results do not improve the confidence intervals further. The performance of our analytic observable is equal to the performance of the neural net.

5 Summary and Outlook

The importance of machine learning in the modern LHC era is growing with the need for better analysis techniques for precision measurements. However, neural networks come with the disadvantage that they are often perceived as black boxes and are difficult to interpret. Especially considering that traditionally, analytical formulas derived from perturbative quantum field theory are the language of particle physicists. In this work we showed how symbolic regression can be used to combine both worlds by using a genetic algorithm to produce analytical expressions.

With the help of the symbolic regression tool PYSR we investigated optimal observables for VVH -couplings. For our toy process we looked at the associated ZH -production process to find the optimal observable for the dimension-6 Wilson coefficient f_B . First of all we considered a process with just one quark type in the initial state to exclude any latent variables and found an approximate function for the joint score which is calculated from matrix element information by MADMINER. For the case of $f_B = 0$ we found that the optimal observable can be described via a polynomial in the two variables p_T and $|\Delta\eta|$. Data obtained from higher Wilson coefficients like $f_B = 10$ suffers from saturation effects and therefore requires a rational function description. We found that including an additional Feynman diagram to this process as well as allowing for all quark types did not change the functional form for $f_B = 0$. For $f_B = 10$ we encountered problems with the PYSR convergence that we fixed by changing the simulated annealing acceptance formula.

Later we turned to a more involved process of weak-boson-fusion studying a CP-violating dimension-6 operator with the Wilson coefficient $f_{W\widetilde{W}}$. In the case of the SM we found simple symbolic expressions including the predicted sine-dependence of the azimuthal angle between the outgoing jets. For higher Wilson coefficients the obtained observables became significantly more complicated. We continued our analysis by including detector effects and found that optimizing the functional form obtained from parton level process on detector level data gives better results than applying PYSR directly on detector data. Finally we obtained confidence intervals for (optimal) observables from theoretical approximation, symbolic regression and neural nets. The exclusion limits of the PYSR expressions showed significant improvement compared to simple theoretical approximations like $\sin\Delta\phi$. The confidence interval of the PYSR result was equal to those from the neural nets indicating that symbolic regression can be used for this task instead of neural nets without any loss in performance.

While not applicable in all places where machine learning is used in particle physics today, symbolic regression can serve as a bridge between fundamental theory and complex experimental analyses.


Acknowledgments

First and foremost I would like to thank Prof. Tilman Plehn for the great opportunity of working on this project and his outstanding support and interest in the work throughout the entire year. I am very grateful to Anja Butter for her guidance, insightful feedback and helpful advice on countless problems throughout this project. Equally, I would like to express my gratitude to Johann Brehmer, not only for coming up with the project idea, but also for his continued interest in the project and support even after leaving research. Without his insights on the usage of MADMINER this work would not have been possible. In addition I would like to thank my colleagues for fun discussions and conversations during lunch and coffee breaks as well as for their patience for me eating slowly. Finally I want to thank my boyfriend Alon, for his never ending mental support and patience for listening to physicists problems, as well as my family for believing in me and never doubting my success.

Erklärung:

Ich versichere, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den 01.10.2021


.....

References

- [1] Ilaria Brivio and Michael Trott. “The standard model as an effective field theory”. In: *Physics Reports* 793 (2019), 1–98. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2018.11.002. URL: <http://dx.doi.org/10.1016/j.physrep.2018.11.002>.
- [2] Giampiero Passarino and Michael Trott. “The Standard Model Effective Field Theory and Next to Leading Order”. In: (2016). arXiv: 1610.08356 [hep-ph].
- [3] D. Atwood and A. Soni. “Analysis for magnetic moment and electric dipole moment form-factors of the top quark via $e^+e^- \rightarrow t\bar{t}$ ”. In: *Phys. Rev. D* 45 (1992), pp. 2405–2413. DOI: 10.1103/PhysRevD.45.2405.
- [4] M. Davier et al. “The Optimal method for the measurement of tau polarization”. In: *Phys. Lett. B* 306 (1993), pp. 411–417. DOI: 10.1016/0370-2693(93)90101-M.
- [5] M. Diehl and O. Nachtmann. “Optimal observables for the measurement of three gauge boson couplings in $e^+e^- \rightarrow W^+W^-$ ”. In: *Z. Phys. C* 62 (1994), pp. 397–412. DOI: 10.1007/BF01555899.
- [6] Otto Nachtmann and Felix Nagel. “Optimal observables and phase-space ambiguities”. In: *Eur. Phys. J. C* 40 (2005), pp. 497–503. DOI: 10.1140/epjc/s2005-02153-9. arXiv: hep-ph/0407224.
- [7] Johann Brehmer et al. “A guide to constraining effective field theories with machine learning”. In: *Physical Review D* 98.5 (2018). ISSN: 2470-0029. DOI: 10.1103/physrevd.98.052004. URL: <http://dx.doi.org/10.1103/PhysRevD.98.052004>.
- [8] Johann Brehmer et al. “Mining gold from implicit models to improve likelihood-free inference”. In: *Proceedings of the National Academy of Sciences* 117.10 (2020), 5242–5249. ISSN: 1091-6490. DOI: 10.1073/pnas.1915980117. URL: <http://dx.doi.org/10.1073/pnas.1915980117>.
- [9] Johann Brehmer et al. “Constraining Effective Field Theories with Machine Learning”. In: *Physical Review Letters* 121.11 (2018). ISSN: 1079-7114. DOI: 10.1103/physrevlett.121.111801. URL: <http://dx.doi.org/10.1103/PhysRevLett.121.111801>.
- [10] David B. Kaplan. “Five lectures on effective field theory”. In: (2005). arXiv: 0510023 [nucl-th].
- [11] Johann Brehmer. “New Ideas for Effective Higgs Measurements”. dissertation. Universität Heidelberg, 2017. URL: https://www.thphys.uni-heidelberg.de/~plehn/includes/theses/brehmer_d.pdf.
- [12] Johann Brehmer et al. *MadMiner: Machine learning-based inference for particle physics*. 2020. arXiv: 1907.10621 [hep-ph].
- [13] R. A. Fisher. “The detection of linkage with ”dominant” abnormalities”. In: *Annals of Eugenics* 6.2 (1935), pp. 187–201. DOI: <https://doi.org/10.1111/j.1469-1809.1935.tb02227.x>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1935.tb02227.x>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1935.tb02227.x>.

- [14] J. Alwall et al. “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations”. In: *Journal of High Energy Physics* 2014.7 (2014). ISSN: 1029-8479. DOI: 10.1007/jhep07(2014)079. URL: [http://dx.doi.org/10.1007/JHEP07\(2014\)079](http://dx.doi.org/10.1007/JHEP07(2014)079).
- [15] Johan Alwall et al. “MadGraph 5: going beyond”. In: *Journal of High Energy Physics* 2011.6 (2011). ISSN: 1029-8479. DOI: 10.1007/jhep06(2011)128. URL: [http://dx.doi.org/10.1007/JHEP06\(2011\)128](http://dx.doi.org/10.1007/JHEP06(2011)128).
- [16] Priscila de Aquino et al. “ALOHA: Automatic libraries of helicity amplitudes for Feynman diagram computations”. In: *Computer Physics Communications* 183.10 (2012), 2254–2263. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2012.05.004. URL: <http://dx.doi.org/10.1016/j.cpc.2012.05.004>.
- [17] Torbjörn Sjöstrand et al. “An introduction to PYTHIA 8.2”. In: *Computer Physics Communications* 191 (2015), 159–177. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2015.01.024. URL: <http://dx.doi.org/10.1016/j.cpc.2015.01.024>.
- [18] Tilman Plehn. “Lectures on LHC Physics”. In: *Lecture Notes in Physics* (2012). ISSN: 1616-6361. DOI: 10.1007/978-3-642-24040-9. URL: <http://dx.doi.org/10.1007/978-3-642-24040-9>.
- [19] Bo Andersson et al. “Parton Fragmentation and String Dynamics”. In: *Phys. Rept.* 97 (1983), pp. 31–145. DOI: 10.1016/0370-1573(83)90080-7.
- [20] J. de Favereau et al. “DELPHES 3: a modular framework for fast simulation of a generic collider experiment”. In: *Journal of High Energy Physics* 2014.2 (2014). ISSN: 1029-8479. DOI: 10.1007/jhep02(2014)057. URL: [http://dx.doi.org/10.1007/JHEP02\(2014\)057](http://dx.doi.org/10.1007/JHEP02(2014)057).
- [21] Johann Brehmer et al. *MadMiner technical documentation*. <https://madminer.readthedocs.io/en/latest/>.
- [22] *A morphing technique for signal modelling in a multidimensional space of coupling parameters*. Tech. rep. Geneva: CERN, 2015. URL: <http://cds.cern.ch/record/2066980>.
- [23] Charles R. Harris et al. “Array programming with NumPy”. In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: <https://doi.org/10.1038/s41586-020-2649-2>.
- [24] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. “On the Convergence of Adam and Beyond”. In: (2019). arXiv: 1904.09237 [cs.LG].
- [25] Glen Cowan et al. “Asymptotic formulae for likelihood-based tests of new physics”. In: *The European Physical Journal C* 71.2 (2011). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-011-1554-0. URL: <http://dx.doi.org/10.1140/epjc/s10052-011-1554-0>.
- [26] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.
- [27] Miles Cranmer. *PySR: Fast & Parallelized Symbolic Regression in Python/Julia*. Sept. 2020. DOI: 10.5281/zenodo.4041459. URL: <http://doi.org/10.5281/zenodo.4041459>.
- [28] Céline Degrande et al. “Effective field theory: A modern approach to anomalous couplings”. In: *Annals of Physics* 335 (2013), 21–32. ISSN: 0003-4916. DOI: 10.1016/j.aop.2013.04.016. URL: <http://dx.doi.org/10.1016/j.aop.2013.04.016>.

- [29] Matt Newville et al. *lmfit/lmfit-py 1.0.2*. Version 1.0.2. Feb. 2021. DOI: 10.5281/zenodo.4516651. URL: <https://doi.org/10.5281/zenodo.4516651>.
- [30] Johann Brehmer et al. “Better Higgs- CP tests through information geometry”. In: *Physical Review D* 97.9 (2018). ISSN: 2470-0029. DOI: 10.1103/physrevd.97.095017. URL: <http://dx.doi.org/10.1103/PhysRevD.97.095017>.
- [31] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. “The anti- k_t jet clustering algorithm”. In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph].