Symbolic regression for precision LHC physics

Manuel Morales-Alvarado
INFN, Sezione di Trieste
SISSA
Trieste, Italy
mmorales@sissa.it

Daniel Conde
IFIC
Universidad de Valencia
Valencia, Spain
daniel.conde@ific.uv.es

Josh Bendavid
Massachusetts Institute
of Technology
Cambridge, Massachusetts, USA
josh.bendavid@cern.ch

Veronica Sanz
IFIC
Universidad de Valencia
Valencia, Spain
veronica.sanz@uv.es

Maria Ubiali
DAMTP
University of Cambridge
Cambridge, UK
m.ubiali@damtp.cam.ac.uk

Abstract

We study the potential of symbolic regression (SR) to derive compact and precise analytic expressions that can improve the accuracy and simplicity of phenomenological analyses at the Large Hadron Collider (LHC). As a benchmark, we apply SR to equation recovery in quantum electrodynamics (QED), where established analytical results from quantum field theory provide a reliable framework for evaluation. This benchmark serves to validate the performance and reliability of SR before extending its application to structure functions in the Drell-Yan process mediated by virtual photons, which lack analytic representations from first principles. By combining the simplicity of analytic expressions with the predictive power of machine learning techniques, SR offers a useful tool for facilitating phenomenological analyses in high energy physics.

1 Introduction

SR is a machine learning task that discovers symbolic models by searching for simple analytic expressions that minimise both prediction error and model complexity. Unlike traditional methods, SR does not fit parameters to a potentially overparametrised model but instead finds concise formulas to describe data. This approach combines the power of machine learning with the clarity of analytical expressions, enabling the extraction of simple formulas from potentially complex datasets.

In LHC physics, some quantities have analytic expressions, while others require expensive fits or iterative algorithms for evaluation, lacking universally known formulas. Previous studies have applied SR in LHC contexts [1, 2], often comparing the derived models to known analytical results. However, a key motivation for this work arises in scenarios where no reference expression exists, prompting the need to assess the reliability of SR methods.

The structure of this work is as follows. Sect. 2 briefly introduces the basics of SR. Sect. 3 validates the methodology by recovering known analytical expressions from noisy data. Sect. 4 presents an SR-derived result for the Drell-Yan structure function with virtual photons, which cannot be obtained from first principles. We conclude in Sect. 5.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

2 Symbolic regression

SR is a supervised learning method that discovers closed-form analytical expressions for input-output relationships without completely predefined functional forms [1–5]. Unlike linear regression or neural networks, SR optimises simultaneously for both accuracy and simplicity.

We use the PySR package [6], a multipopulation evolutionary algorithm that evaluates symbolic expressions as expression trees. It can be highly effective for parameter spaces of moderate size [7, 8]. Fig. 1 shows an example tree representing the equation $3.1y \cdot (x^2 + 1)$. In this study, equations are

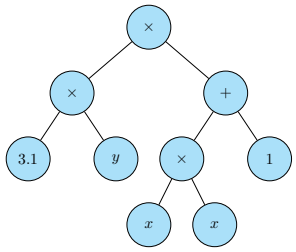

Figure 1: Example of an expression tree.

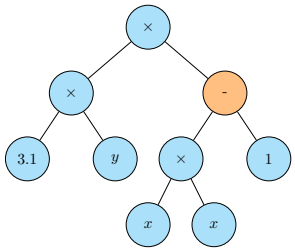

Figure 2: Mutation of an expression tree.

evaluated using the MSE loss function. During optimisation, expression trees mutate over iterations to improve based on a selection criterion, as shown in Fig. 2, where a '$+$' in Fig. 1 mutates to a '$-$'.

Trees can also combine through crossover, and complexity, defined by node count, is managed via multiobjective optimisation. More details are available in the original reference.

The selection criterion in PySR guides the evolutionary algorithm in choosing the fittest expression trees. There are three criteria: `accuracy`, which selects the model with the lowest loss; `score`, defined as the negative derivative of log-loss with respect to complexity, selects the model with highest decrease in loss with marginally higher complexity; and `best`, which selects the model with the highest score, provided its loss does not exceed 1.5 times that of the most accurate model.

3 Equation rediscovery from QED

In this section, we apply SR to the process $e^+e^- \to \gamma^* \to \mu^+\mu^-$ at leading order, testing its ability to recover the angular cross-section distribution in the massless limit. From QED, the distribution is:

$$\frac{d\sigma}{d\cos\theta} = \frac{\pi\alpha^2}{2s}(1 + \cos^2\theta), \tag{1}$$

where θ is the angle between the outgoing muon and incoming electrons, α the QED coupling, and s the squared centre-of-mass energy, treated as a constant. Radiative corrections and higher-order effects are not included, so α is also treated as constant. SR will aim to rediscover this equation from simulated samples.

To train the regressor, we generate $100k$ events at $\sqrt{s} = 1$ TeV using MadGraph5_aMC@NLO [9, 10] without kinematic cuts. From these events, we extract $\cos\theta$ distributions for various binnings and show the corresponding samples to the regressor. As Eq. (1) indicates, $\cos\theta$ is the key kinematic variable.

In Tab. 1, we present the analytical equations derived from the simulated distributions for different binnings. Comparing with Eq. (1), we observe that the `accuracy` criterion often fails to recover the equation across different binning levels, fitting the noise. In contrast, `best`, successfully recovers the correct equation with finer binning. Notably, `score` consistently identifies the correct equation, even with very fine binning, demonstrating that simplicity can be an effective factor.

Realistic simulations must account for uncertainties and fluctuations in the data. Figs. 3 and 4 show the absolute and normalised distributions for 30 bins, comparing the simulation's central value (with 1-standard deviation Poisson uncertainty at the level of the event count in each bin), the SR result,

Table 1: Equations according to the three selection criteria for different bin sizes with $c_\theta \equiv \cos\theta$. The numbers that appear in these expressions have been approximated to the 5th decimal place.

Bins	Accuracy	Best	Score
10	$c_\theta(c_\theta + 0.00798)(0.00111 \cdot c_\theta^3 + 0.03459) + 0.03503$	$c_\theta^2(0.00111 \cdot c_\theta + 0.03459) + 0.03503$	$0.03459 \cdot c_\theta^2 + 0.03503$
20	$c_\theta(c_\theta + 0.01825)(-0.00155 \cdot c_\theta(c_\theta - 0.05138) + 0.03579) + 0.03485$	$c_\theta(0.03447 \cdot c_\theta + 0.00064) + 0.03498$	$0.03447 \cdot c_\theta^2 + 0.03498$
200	$c_\theta^2(-0.64647 \cdot c_\theta(0.00119 \cdot c_\theta - 0.00151) + 0.03495) + 0.03495$	$0.03447 \cdot c_\theta^2 + 0.03495$	$0.03447 \cdot c_\theta^2 + 0.03495$

and the analytic equation. SR demonstrates excellent agreement with the true functional form, even in cases where the simulation deviates by more than one standard deviation from it. This shows SR's ability to recover not only accurate distributions but also the correct functional dependence derived from first principles in QED.

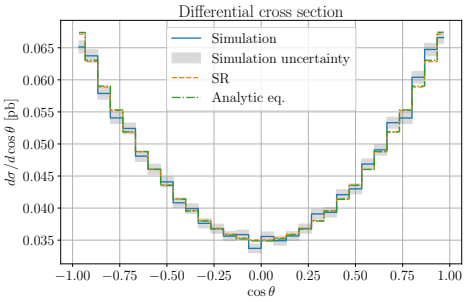

Figure 3: Absolute distribution for 30 bins.

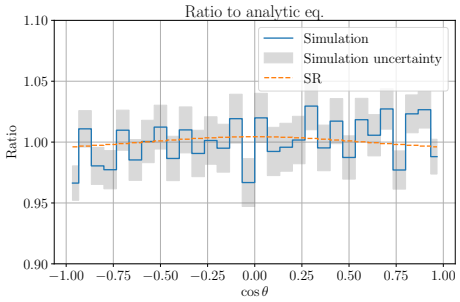

Figure 4: Normalised distribution for 30 bins.

Having verified a known natural law across various settings, we can now explore how SR can shed light on expressions that do not have a known closed analytical formula. This is, for example, the case of parton distribution functions and structure functions.

4 Proton structure functions

Parton distribution functions (PDFs) are essential for calculating observables at hadron colliders, as discussed in [11, 12] and references therein. They represent momentum distribution of partons within hadrons, which must be combined with partonic cross sections to produce physical predictions to compare with experimental data. PDFs cannot be calculated from first principles, as they encapsulate the non-perturbative regime of quantum chromodynamics where the strong coupling becomes too large for perturbative methods to converge. Instead, PDFs have to be fitted from experimental data. This is achieved at the state of the art by using fixed functional forms or neural networks with hundreds of trainable weights [13–15].

Many differential observables depend on structure functions (SFs), which are weighted combinations of PDFs [16, 17]. Like PDFs themselves, SFs lack closed analytical expressions. In this section, we introduce the first SR approach to derive SFs. Our goal is to obtain accurate and compact analytical expressions that can effectively model these functions, offering a more straightforward and clear understanding of their behaviour compared to current techniques.

We consider the leading order Drell-Yan (DY) process $p\ p \to \gamma^* \to \mu^+\ \mu^-$ at $\sqrt{s} = 1$ TeV and generate $100k$ events. No standard cuts are applied on the event generation. We use the CT10 NLO

PDF set [18, 19]. The DY double differential cross section is given by

$$\frac{d^2\sigma}{dMdy} = \frac{1}{3s}\frac{8\pi\alpha^2}{3M}\left(\sum_q Q_q^2(f_q(x_1,\tau)f_{\overline{q}}(x_2,\tau) + f_{\overline{q}}(x_1,\tau)f_q(x_2,\tau)\right) \equiv \frac{1}{3s}\frac{8\pi\alpha^2}{3M}F(M,y), \tag{2}$$

where M and y are, respectively, the invariant mass and the rapidity of the muon pair, $x_1 = \sqrt{\tau}e^y$ and $x_2 = \sqrt{\tau}e^{-y}$ are the parton momentum fractions that cannot be bigger than 1, and $\tau = M^2/s$. The sum in parentheses runs over the $q = u, d, s, c$ quark flavours, Q_q is their respective electric charge, and f_q are their associated PDFs. We are exclusively interested in the SF $F(M, y)$ at high resolution in the kinematic coverage as the rest of the distribution is known.

The SF values from the simulation are shown in Fig. 5, calculated by reweighting the double differential distribution with the prefactor $\frac{1}{3s}\frac{8\pi\alpha^2}{3M}$ as per Eq. (2). Applying SR to this distribution yields a hall-of-fame set of models at various complexities, with selected examples shown in Table 2. The `best` model, with a complexity of 33 (highlighted in gray), is shown in Fig. 6. These results are compared to the central replica values of the PDF set, shown in Fig. 7. From the figures, we observe

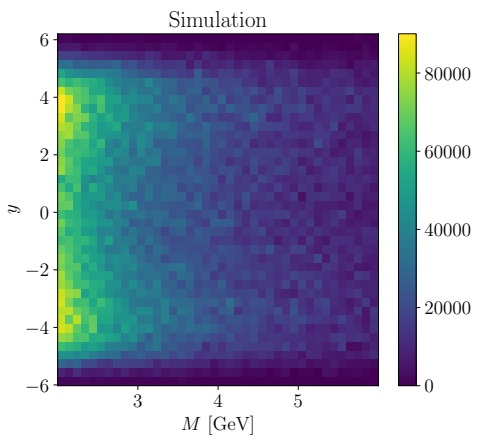

Figure 5: SF values obtained from the reweighted simulation.

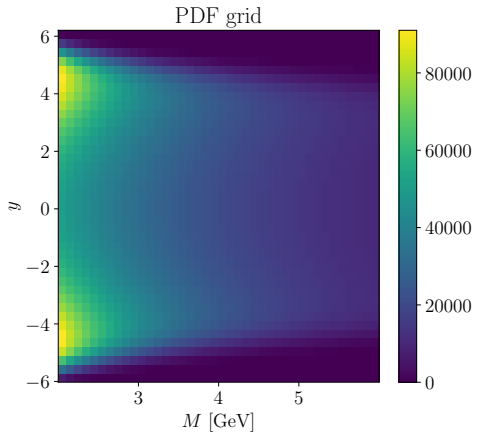

Figure 6: SF values obtained from the central value PDF grids.

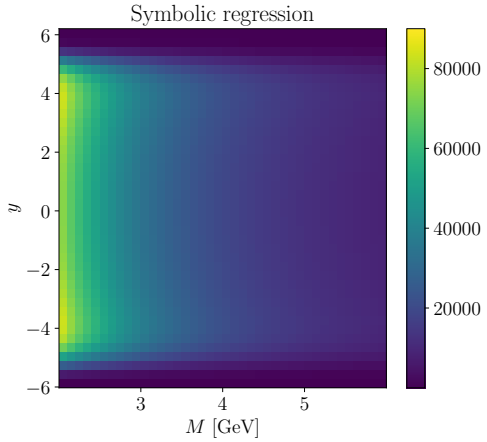

Figure 7: SF $F(M, y)$ values obtained with SR.

that the SR result provides a smooth function that approximates well the PDF grid. It successfully extrapolates to the unphysical regions of the kinematic coverage by suppressing the function (although it does not fully reach zero) and effectively reduces fluctuations at high invariant mass M. Moreover, the expressions in the table show that the SF can be parametrised at low complexity (3) using inverse power laws in M. At higher complexities, where there is resolution in y, the SR recognises that it

must be symmetric in y, as dictated by particle kinematics. Notably, the most complex expression (complexity 35), while more accurate by construction, achieves a lower score due to increased expression length.

Table 2: Selection of `best` SR expressions $F(M, y)$ with their complexities and scores. Constants are approximated for display purposes.

Complexity	Equation	Score
3	$\frac{9.16 \cdot 10^4}{M}$	0.359
33	$\frac{2.86 \cdot 10^5 \left(0.0461 \cdot 1.15^{y^2}\right)^{-0.0250 \cdot 1.15^{y^2}}}{M^2}$	0.387
35	$\frac{2.86 \cdot 10^5 \left(0.0461 \cdot 1.15^{y^2}\right)^{-0.0250 \cdot 1.15^{y^2}}}{M^2 + 0.117}$	0.00967

To compare the simulation and SR results to the PDF grid baseline, in Table 3 we use the following fit quality metrics:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \qquad \text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \qquad R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where n is the number of data points, y_i the actual value, $\hat{y}_i$ the predicted value, and $\bar{y}$ the mean of y_i. We see that SR achieves lower RMSE and MAE values, indicating a better overall fit quality with the PDF grid. Additionally, the higher R^2 value for the SR fit (0.9030 vs. 0.8898) reflects a stronger correlation. These results demonstrate the capability of SR to model the data with good precision.

Table 3: Comparison of metrics of the reweighted simulation and the SR fit with respect to the PDF grid.

Metric	Reweighted simulation	Symbolic regression
Root mean square error (RMSE)	6.09×10^3	5.72×10^3
Mean absolute error (MAE)	4.26×10^3	3.74×10^3
Coef. of determination (R^2)	0.8898	0.9030

5 Conclusion

We have explored the application of SR to derive simple yet accurate analytical formulas in the context of collider phenomenology. Using well-established QED processes as a benchmark, we validated the reliability of SR by recovering known analytical expressions from noisy data under varying conditions. Furthermore, we extended this methodology to SFs in Drell-Yan, showing the potential of SR to provide accurate and simple results even in scenarios where no closed-form solutions are available. This study highlights the utility of SR in simplifying and enhancing the analysis of complex datasets in high energy physics, and could be extended to study more sophisticated uncertainties, higher-dimensional distributions, or higher-order processes like those involving electroweak boson production with angular coefficients [20–25].

SR combines machine learning with analytical expressions, enabling accurate closed-form models from complex datasets. This work contributes to precision physics at the LHC and, more in general, machine learning-assisted discovery in high energy physics.

References

[1] Anja Butter, Tilman Plehn, Nathalie Soybelman, and Johann Brehmer. Back to the Formula – LHC Edition. 9 2021.

[2] Zhongtian Dong, Kyoungchul Kong, Konstantin T. Matchev, and Katia Matcheva. Is the machine smarter than the theorist: Deriving formulas for particle kinematics with symbolic regression. *Phys. Rev. D*, 107(5):055018, 2023.

[3] Suyong Choi. Construction of a Kinematic Variable Sensitive to the Mass of the Standard Model Higgs Boson in $H->WW^*->l^+\nu l^-\bar{\nu}$ using Symbolic Regression. *JHEP*, 08:110, 2011.

[4] Aurélien Dersy, Matthew D. Schwartz, and Xiaoyuan Zhang. Simplifying Polylogarithms with Machine Learning. 6 2022.

[5] Abdulhakim Alnuqaydan, Sergei Gleyzer, and Harrison Prosper. SYMBA: symbolic computation of squared amplitudes in high energy physics with machine learning. *Mach. Learn. Sci. Tech.*, 4(1):015007, 2023.

[6] M. Cranmer. Interpretable machine learning for science with pysr and symbolicregression.jl. *ArXiv*, abs/2305.01582, 2023.

[7] Shehu AbdusSalam, Steve Abel, and Miguel Crispim Romao. Symbolic regression for beyond the standard model physics. 2024.

[8] F. O. de Franca, M. Virgolin, M. Kommenda, M. S. Majumder, M. Cranmer, G. Espada, L. Ingelse, A. Fonseca, M. Landajuela, B. Petersen, R. Glatt, N. Mundhenk, C. S. Lee, J. D. Hochhalter, D. L. Randall, P. Kamienny, H. Zhang, G. Dick, A. Simon, B. Burlacu, Jaan Kasak, Meera Machado, Casper Wilstrup, and W. G. La Cava. Interpretable symbolic regression for data science: Analysis of the 2022 competition. 2023.

[9] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014.

[10] R. Frederix, S. Frixione, V. Hirschi, D. Pagani, H. S. Shao, and M. Zaro. The automation of next-to-leading order electroweak calculations. *JHEP*, 07:185, 2018. [Erratum: JHEP 11, 085 (2021)].

[11] Jon Butterworth et al. PDF4LHC recommendations for LHC Run II. *J. Phys. G*, 43:023001, 2016.

[12] Thomas Cridge. PDF4LHC21: Update on the benchmarking of the CT, MSHT and NNPDF global PDF fits. *SciPost Phys. Proc.*, 8:101, 2022.

[13] S. Bailey, T. Cridge, L. A. Harland-Lang, A. D. Martin, and R. S. Thorne. Parton distributions from LHC, HERA, Tevatron and fixed target data: MSHT20 PDFs. *Eur. Phys. J. C*, 81(4):341, 2021.

[14] Tie-Jiun Hou et al. New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC. *Phys. Rev. D*, 103(1):014013, 2021.

[15] Richard D. Ball et al. The path to proton structure at 1% accuracy. *Eur. Phys. J. C*, 82(5):428, 2022.

[16] R. Keith Ellis, W. James Stirling, and B. R. Webber. *QCD and collider physics*, volume 8. Cambridge University Press, 2 2011.

[17] Michael E. Peskin and Daniel V. Schroeder. *An Introduction to quantum field theory*. Addison-Wesley, Reading, USA, 1995.

[18] Jun Gao, Marco Guzzi, Joey Huston, Hung-Liang Lai, Zhao Li, Pavel Nadolsky, Jon Pumplin, Daniel Stump, and C. P. Yuan. CT10 next-to-next-to-leading order global analysis of QCD. *Phys. Rev. D*, 89(3):033009, 2014.

[19] Andy Buckley, James Ferrando, Stephen Lloyd, Karl Nordström, Ben Page, Martin Rüfenacht, Marek Schönherr, and Graeme Watt. LHAPDF6: parton density access in the LHC precision era. *Eur. Phys. J. C*, 75:132, 2015.

[20] John C. Collins and Davison E. Soper. Angular Distribution of Dileptons in High-Energy Hadron Collisions. *Phys. Rev. D*, 16:2219, 1977.

[21] E. Mirkes. Angular decay distribution of leptons from W bosons at NLO in hadronic collisions. *Nucl. Phys. B*, 387:3–85, 1992.

[22] E. Mirkes and J. Ohnemus. Angular distributions of Drell-Yan lepton pairs at the Tevatron: Order $\alpha - s^2$ corrections and Monte Carlo studies. *Phys. Rev. D*, 51:4891–4904, 1995.

[23] E. Mirkes and J. Ohnemus. W and Z polarization effects in hadronic collisions. *Phys. Rev. D*, 50:5692–5703, 1994.

[24] E. Mirkes and J. Ohnemus. Polarization effects in Drell-Yan type processes h1 + h2 —> (W, Z, gamma*, J / psi) + x. In *1994 Meeting of the American Physical Society, Division of Particles and Fields (DPF 94)*, pages 1721–1723, 8 1994.

[25] Georges Aad et al. A precise measurement of the Z-boson double-differential transverse momentum and rapidity distributions in the full phase space of the decay leptons with the ATLAS experiment at $\sqrt{s}$ = 8 TeV. 9 2023.