Phy-SRBench: A Physics Benchmark for Scientific Equation Discovery with Symbolic Regression

Nour Makke
Qatar Computing Research Institute, HBKU
Doha, Qatar
nmakke@hbku.edu.qa

Sanjay Chawla
Qatar Computing Research Institute, HBKU
Doha, Qatar
schawla@hbku.edu.qa

Abstract

Scientific discovery has driven conceptual and technological advances in all areas, including non-scientific domains. Symbolic regression is a machine learning subfield that focuses on automating scientific discovery towards making new discoveries. It suffers from key limitations in its key component, data benchmarks. This paper presents a new dataset benchmark originating from physics, specifically high-energy physics. These equations are domain-specific and can be expressed in different forms, thus allowing for different sets of input-output pairs.

1 Introduction

The ultimate goal of research at the interface of artificial intelligence (AI) and natural sciences is to build models that understand the world, accelerate scientific discovery, and gather relevant knowledge from a "science" perspective. Symbolic regression (SR) is a machine learning (ML) subfield that has gained growing interest in recent years for learning transparent AI systems, automating scientific understanding, and boosting scientific discovery. It focuses on learning analytical models from numerical data. Following the deep learning (DL) revolution in early 2010, SR methods have significantly evolved from search-based methods (e.g., heuristic search and genetic programming (GP) (1), reinforcement learning (2)) to modern learning-based methods that leverage large-scale synthetic data and transformers (3; 4), and hybrid techniques (5). A full review of SR methods can be found in (6; 7).

SR is intuitively suited for applications in the natural sciences that seek to learn the latent structure of the world through theoretical models, and particularly in the physical sciences, where laws are expressed as equations. Despite significant advancements in SR in recent years, its application to complex real-world and physics problems remains limited. Moreover, most existing applications utilize synthetic (i.e., simulated) data, while only a few studies employ experimental physics data, which is naturally noisy, to learn either fundamental laws or established physical models (8; 9; 10; 11). Applications of SR to experimental data with known physical models are critical for testing the credibility of SR at the practice level and evaluating its effectiveness, holding promise in SR. To enhance trust in SR for making new discoveries in science, it is essential to demonstrate the efficiency of SR in discovering fundamental laws (i.e., (8)) and phenomenological models (i.e., (10; 11)) using experimental physics data. This may be hard to achieve using existing SR dataset benchmarks.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

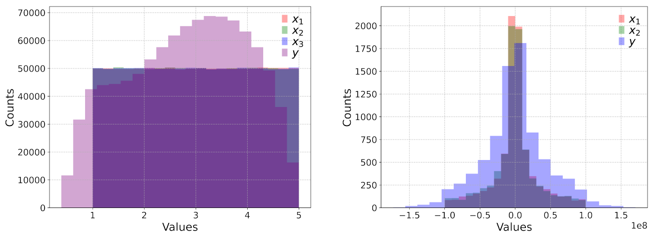

Figure 1: Distributions of the features (x) and the target model (y) for the (exemplary) physics equation $y = (u + v)/(1 + uv/c^2)$, where u and v are relativistic velocities and c is the speed of light. Left: In the AIFeynman benchmark (13), $\mathrm{x} = \{u, v, c\} \in \mathbb{R}^3$, all are uniformly generated in the range [1,5]. Right: In the the revised benchmark SRSD (12), $\mathrm{x} = \{u,\ v\} \in \mathbb{R}^2$ and $c = 2.99 \times 10^8$. The distributions are significantly different, leading to a noticeable difference in SR performance.

It is no surprise that synthetic SR datasets, which involve generating numerical points using an equation, are technically easier to create compared to the datasets required in other ML-related fields. However, high-quality datasets in SR remain challenging, given that SR is primarily developed to boost discoveries in science, and physical sciences are known for their scalability. Currently, the most widely used physics benchmark is the AIFeynman, which suffers key limitations (7), as illustrated in Fig. 1. A corrected version was proposed in (12), but it was not adopted in SR algorithmic works. In nearly all domains of physics, there is an enormous amount of high-statistics and multi-dimensional measurements of physical observables in the physical disciplines that could be used for SR applications.

2 Cross-sections in high-energy physics

The cross sections hold particular importance in high-energy physics (HEP). They are experimentally measured and subsequently compared to theoretical calculations. Such comparisons are essential to validate theoretical calculations and explore missing factors in theory when a mismatch is observed. Cross sections are calculated for different processes that involve different types of interactions, and their expressions cover a wide range of energy.

A differential cross section of a scattering process in high-energy physics (14) is usually expressed in terms of kinematical variables such as the energy (E) and the scattering angle (θ), which are are defined using the four-momenta of incoming and outgoing particles. A more generalized manner of expressing the cross section uses the so-called Mandelstam variables $\{s, t, u\}$. In the center-of-mass (CM) frame, and assuming massless particles in the ultra-relativistic limit, these variables are functions of the energy (E_{CM}) and the scattering angle (θ) defined between the direction of the scattered particle with respect to the direction of the incoming particles. They are expressed as follows:

$$\left. s = E_{\mathrm{CM}}^2,\ t = -\frac{E_{\mathrm{CM}}^2}{2}\left(1 - \cos\theta\right),\ u = -\frac{E_{\mathrm{CM}}^2}{2}\left(1 + \cos\theta\right) \right\} \frac{\mathrm{d}\sigma}{\mathrm{d}\Omega} = g(E_{\mathrm{CM}}, \theta) = g''(s, t, u) \quad (1)$$

The choice of the input vector of features and the target expression is flexible, yielding a total of 18 $\{\mathrm{x}, y\}$, as summarized in Table 1. There are the kinematic variables $\{E_{\mathrm{CM}}, \theta/\cos\theta\}$ which can be regarded as the original features, the Mandelstam variables which are themselves functions of kinematic variables, and the cross sections which could be expressed either in terms of kinematic or Mandelstam variables. Therefore, one could build different sets of input/output using the columns in Tab 1, and multi-dimensional distributions could be generated and used to infer ground-truth expressions using SR, namely, one dimension using$(\theta/\cos\theta)$ by fixing E_{CM}'s value, two dimensions $(E_{\mathrm{CM}}, \theta/\cos\theta)$ by varying E_{CM}'s value, and three dimensional configurations using (s, t, u). The advantage of expressing the cross-section in terms of different but equivalent variables resides in the different sampling ranges of these variables. This represents a robustness challenge, e.g., both $\mathrm{d}\sigma(\cos\theta)$ and $\mathrm{d}\sigma(\theta)$ exhibit the same angular dependence but with slightly different ground-truth expressions whose skeletons are equivalent up to a variable changing.

Assuming that measurement for particles are restricted with a scattering angle greater than 10°, then the sampling range for θ would be $[10^{\circ}, 170^{\circ}]$ whereas the sampling range of $\cos\theta$ would ideally

Table 1: Summary of possible input vector of variables and target expressions related to cross-section topic in HEP. The Mandelstam variables are defined in terms of $(E_{\mathrm{CM}}^2, \theta/\cos\theta)$ in the CM frame. The cross sections (3$^{\mathrm{rd}}$ column) could be expressed in terms of either $(E_{\mathrm{CM}}^2, \theta/\cos\theta)$ or $\{s, t, u\}$.

Kinematic variables			Mandelstam variables: {s,t,u}		Cross-sections: $\mathrm{d}\sigma$		
E_{CM}	θ	$\cos\theta$	$f(E_{\mathrm{CM}}^2, \theta)$	$f'(E_{\mathrm{CM}}^2, \cos\theta)$	$g(s, \theta)$	$g'(s, \cos\theta)$	$g''(s, t, u)$

Table 2: Expressions of the differential cross section for various scattering processes in high-energy physics (14) in terms of the Mandelstam variables (1$^{\mathrm{st}}$ column) and kinematic variables (2$^{\mathrm{nd}}$ column), such as energy and scattering angle. For a given scattering process, the expressions of $\mathrm{d}\sigma$ have different forms but the same numerical values. "O" refers to the complexity of an equation, which is the number of nodes in the expression-tree representation (7).

Scattering Process	$\mathrm{d}\sigma(s, t, u)$	O	$\mathrm{d}\sigma(E_{\mathrm{CM}}, x = \cos\theta)$	O
Moeller $e^- e^- \to e^- e^-$	$\frac{\alpha^2}{2s}\left(\frac{s^2+u^2}{t^2} + \frac{s^2+t^2}{u^2} + \frac{2s^2}{tu}\right)$	42	$\frac{\pi\alpha^2}{E_{\mathrm{CM}}^2}\left(\frac{2(x^2+3)^2}{(x^2-1)^2}\right)$	23
Bhabha $e^- e^+ \to e^- e^+$	$\frac{\alpha^2}{2s}\left(\frac{s^2+u^2}{t^2} + \frac{t^2+u^2}{s^2} + \frac{2u^2}{ts}\right)$	42	$\frac{\pi\alpha^2}{E_{\mathrm{CM}}^2}\left(\frac{(x^2+3)^2}{2(x-1)^2}\right)$	23
emu $e^- \mu^- \to e^- \mu^-$	$\frac{\alpha^2}{2s}\left(\frac{s^2+u^2}{t^2}\right)$	18	$\frac{e^4}{8\pi^2 s}\left(\frac{1+1/4(1+x)^2}{(1-x)^2}\right)$	24
mupp $e^+ e^- \to \mu^+ \mu^-$	$\frac{\alpha^2}{2s}\left(\frac{t^2+u^2}{s^2}\right)$	18	$\frac{\alpha^2}{4s}(1 + x^2)$	13

be $[-1, 1]$ and more realistically $[-0.95, 0.95]$. The set of angular data points is valid for a specific value of the energy E_{CM}, which in turn could take any value with $E_{\mathrm{CM}} > 0$. The numerical ranges for t, u would be $-s/2 < t, u < 0$. The ideal θ-range is $[10^{\circ}, 170^{\circ}]$; however, at the practical level, the angular coverage depends on experimental setups and therefore could be truncated to much smaller intervals. When dealing with synthetic data, one could truncate the overall θ range into different intervals, towards smaller or larger values, to imitate experimental conditions, and check the performance of the SR methods in this regard. In other terms, if θ could only be measured in a restricted range, would SR be able to recover the correct expression from that truncated range, and how far could it extrapolate? This step is equivalent to systematic studies in physical observables' measurement and could be considered as a fundamental task in the research workflow of developing SR algorithms.

The expressions of the cross-section for various HEP processes are listed in Tab. 2, in terms of $\{s, t, u\}$ (1$^{\mathrm{st}}$ column) and $(s, \cos\theta)$ (2$^{\mathrm{nd}}$ column). The cross-sections have similar expressions in the "Mandelstam variable" representation; however, they exhibit different angular dependencies, i.e, $\mathrm{d}\sigma_{\mathrm{emu}}$ versus $\mathrm{d}\sigma_{\mathrm{mupp}}$. The expressions in terms of (s, θ) are obviously the same as those shown in the second column by replacing x with $\cos\theta$. α is a physical constant and is equivalent to $1/137$.

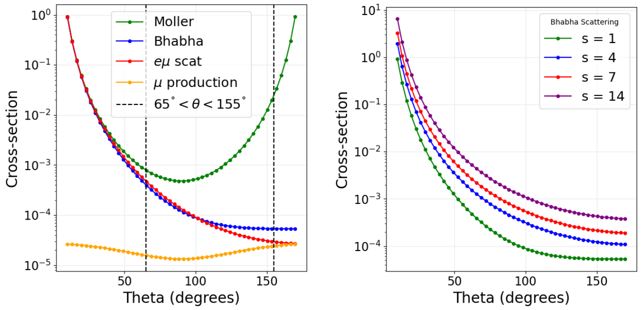

Figure 2: Left: 1D angular dependence of $\mathrm{d}\sigma$ for various HEP processes (expressions summarized in Tab. 2), generated at $s = 1$ (GeV)^2. Right: 2D angular dependence of the cross-section of the Bhabha scattering at different values of the center-of-mass energy (s). Shown on logarithmic scales.

Table 3: Results obtained by applying GP-based (PySR (1), 1$^{\text{st}}$ sub-row) and transformer-based (NeSymReS (3), 2$^{\text{nd}}$ sub-row) SR methods on data generated from the equations presented in this paper, namely Eq. 1 and Tab. 2. "R" refers to the recovery status, and Y/N refers to Yes or No for recovering the ground-truth equation for each problem.

input	output	MSE	expression	R
(E_{CM}, x)	s	0.0	(1)	Y
		1.2224391e-14	E_{CM}*(E_{CM} - sin(0.0*E_{CM}))	Y
	t	0.0062	$-0.42\exp(-0.43x(E_{\text{CM}}+1.23))$	N
		1.1044069e-14	$-0.5*E_{\text{CM}}*(-E_{\text{CM}}*x+E_{\text{CM}})$	Y
	u	0.0032	$-0.65\cos(E_{\text{CM}}-x)-0.22$	N
		1.0132345e-14	$-0.5*E_{\text{CM}}*(1.0*E_{\text{CM}}*x+E_{\text{CM}})$	Y
(E_{CM}, θ)	g_{moller}	0.038	$1.26\exp(0.67E_{\text{CM}}-\theta)$	N
		0.5269817	$82.09788*\exp(-1.27896*E_{\text{CM}}-0.32175*\theta)$	N
	g_{bhabha}	3.410^{-05}	$7.72\exp(1.12E_{\text{CM}}-0.33*\theta)$	N
		1.0232418	$-0.21606*\exp(E_{\text{CM}}-\theta^2)$	N
	$g_{e\mu}$	0.019	$1.11\exp(E_{\text{CM}}-\theta^2)$	N
		1.0234939	$0.6197*\exp(E_{\text{CM}}-\theta^2)$	N
	$g_{\mu pp}$	6.13e-11	$1.20e-7*E_{\text{CM}}*(\theta+44.73)$	N
		0.51509655	$0.0*(E_{\text{CM}}+\theta+1.36933)/(E_{\text{CM}}$ - 0.4185)	N
(E_{CM}, x)	g'_{moller}	0.028	$0.18*E_{\text{CM}}^2*(-x-0.0064)^2$	N
		0.7218451	$0.0*\exp(-15.35508*\cos\theta*\sin(1.59189*E_{\text{CM}}))$	N
	g'_{bhabha}	0.0158	$0.0058*\exp(E_{\text{CM}}*(x+1.83))$	N
		0.8730789	$0.101*\exp(E_{\text{CM}}*(-E_{\text{CM}}+\cos\theta))$	N
	$g'_{e\mu}$	0.016	$3.58*\exp(E_{\text{CM}}*(x-4.58))$	N
		0.8718219	$0.10235*\exp(E_{\text{CM}}*(-E_{\text{CM}}+x))$	N
	$g'_{\mu pp}$	1.34e-10	$-1.33e-5*E_{\text{CM}}*(x-1)$	N
		0.12042928	$3e-05*E_{\text{CM}}/(E_{\text{CM}}^2+0.27527*E_{\text{CM}})^2$	N
(s, t, u)	g''_{moller}	0.028	$0.182\,s^2\,(0.99\,t-u)^2$	N
		0.515459	$-5e-05*s^2/(t^2*u)$	N
	g''_{bhabha}	0.0001	$0.0038\exp(179.6s(s-u-1.96))$	N
		1.0080006	$0.0*s^2*u^2/t^2$	N
	$g''_{e\mu}$	0.018	$0.15s\exp(-1.6*s^2)$	N
		1.0080122	$0.0*s^2*u^2/t^2$	N
	$g''_{\mu pp}$	3.26e-11	$-10^{-5}*s(t-1.26)$	N
		0.09380683	$-1e-05*(s+t+u-1.85774)/s$	N

Figure 2 (left) compares the cross sections for different processes as a function of θ, generated for a CM energy value of 1 TeV. The expressions exhibit different angular dependencies. The black-dashed line represents a possible experimental limitation of the angular coverage, which is arbitrarily chosen here and depends on the experimental setup. Figure 2 (right) compares the angular dependence of the cross-section for one scattering (cf. Tab. 2) evaluated at different values of s, namely, $1, 4, 7$, and 14 (TeV)^2.

We conduct several experiments to infer the target functions summarized in Tab.2. Two SR packages were used: a transformer-based SR, i.e., NeSymReS (3), and a GP-based SR, i.e., PySR (1). The choice of a transformer-based SR is mainly driven by the fact that learning the context in data holds significant meaning in physics, particularly in light of the causal nature of physical phenomena, where capturing correlations among variables is crucial. This study uses the NeSymReS model pretrained on 100 million datasets, and the default configuration of PySR. This choice is driven by the successful applications of SR using experimental data that use these two software. Results obtained using both SR methods are reported in Tab. 3. Except for the equations of the Mandelstam variable, none of the differential cross-section equations were correctly inferred, namely $g,\ g', g''$ in Tab. 1. In addition, it's evident that some of the learned models do not fulfill the dimensional requirements of physics equations, i.e., the units of measurements have to follow dimensional rules.

3 Conclusion

This paper presents an exemplary set of physics problems that could be used in the training and performance evaluation of SR algorithms, as part of an ongoing work on an SR benchmark consisting of physics equations. The equations presented and discussed here depend on multiple variables that are generated within specific physical ranges. In addition, they can be expressed in different analytical forms when the input variables are changed, which can be challenging for SR algorithms.

References

[1] Cranmer M. Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl; 2023.

[2] Petersen BK. Deep symbolic regression: Recovering mathematical expressions from data via policy gradients. CoRR. 2019;abs/1912.04871. Available from: `http://arxiv.org/abs/1912.04871`.

[3] Biggio L, Bendinelli T, Neitz A, Lucchi A, Parascandolo G. Neural Symbolic Regression that Scales. CoRR. 2021;abs/2106.06427. Available from: `https://arxiv.org/abs/2106.06427`.

[4] Kamienny PA, d'Ascoli S, Lample G, Charton F. End-to-end symbolic regression with transformers. arXiv; 2022. Available from: `https://arxiv.org/abs/2204.10532`.

[5] Landajuela M, Lee C, Yang J, Glatt R, Santiago CP, Aravena I, et al. A Unified Framework for Deep Symbolic Regression. In: Advances in Neural Information Processing Systems; 2022. .

[6] Makke N, Chawla S. Interpretable scientific discovery with symbolic regression: a review. Artificial Intelligence Review. 2024;57(1):2. Available from: `https://doi.org/10.1007/s10462-023-10622-0`.

[7] Makke N, Chawla S. Symbolic Regression: A Pathway to Interpretability Towards Automated Scientific Discovery. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '24. New York, NY, USA: Association for Computing Machinery; 2024. p. 6588–6596. Available from: `https://doi.org/10.1145/3637528.3671464`.

[8] Lemos P, Jeffrey N, Cranmer M, Ho S, Battaglia P. Rediscovering orbital mechanics with machine learning. arXiv; 2022. Available from: `https://arxiv.org/abs/2202.02306`.

[9] Reinbold PAK, Kageorge LM, Schatz MF, Grigoriev RO. Robust learning from noisy, incomplete, high-dimensional experimental data via physically constrained symbolic regression.

Nature Communications. 2021 May;12:3219. Available from: https://doi.org/10.1038/s41467-021-23479-0.

[10] Makke N, Chawla S. Data-driven discovery of Tsallis-like distribution using symbolic regression in high-energy physics. PNAS Nexus. 2024 10:pgae467. Available from: https://doi.org/10.1093/pnasnexus/pgae467.

[11] Makke N, Chawla S. Inferring Interpretable Models of Fragmentation Functions using Symbolic Regression. 2025 1.

[12] Matsubara Y, Chiba N, Igarashi R, Ushiku Y. Rethinking Symbolic Regression Datasets and Benchmarks for Scientific Discovery; 2024. Available from: https://arxiv.org/abs/2206.10540.

[13] Udrescu SM, Tegmark M. AI Feynman: a Physics-Inspired Method for Symbolic Regression; 2020. Available from: https://arxiv.org/abs/1905.11481.

[14] Halzen F, Martin AD. QUARKS AND LEPTONS: AN INTRODUCTORY COURSE IN MODERN PARTICLE PHYSICS; 1984.