Re-Discovering Tsallis distribution from High-Energy Physics data using Symbolic Regression

Nour Makke, Sanjay Chawla

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

Abstract

Symbolic regression is an AI-based technique whose objective is to learn concise mathematical expressions directly from data. Mathematical expressions are directly interpretable and are not only good predictors but can also be used for inferring causal behavior. Nearly all symbolic regression methods are evaluated on synthetic datasets, which do not necessarily have any hidden structure, and noise is generated from a normal distribution and added to the simulated dataset. In this paper, we present the application of symbolic regression to a real physics problem using data collected in high-energy physics experiments. We show that the symbolic regression method based on transformer neural networks could learn a model similar to the so-called Tsallis distribution, a well-known empirical law usually used to fit these datasets.

Introduction

It is widely acknowledged that the majority of deep learning-based models are black boxes (BB), meaning the relationship between a model's prediction and its inputs is complex and entirely opaque. Existing methods, such as DeepLIFT (Shrikumar, Greenside, and Kundaje 2017) and LIME (Ribeiro, Singh, and Guestrin 2016), attempt to explain predictions of BB models. However, this approach is often deemed inefficient in many application domains, and some of these methods are BB, meaning a BB model is used to explain the predictions of another BB model. A more direct and effective strategy is to build interpretable models. This is particularly pertinent in scientific research, where underlying phenomena are typically expressed as mathematical formulae. Symbolic regression (SR) perfectly aligns with this philosophy, allowing the extraction of mathematical equations directly from data, in contrast to prevailing practices where a pre-defined model is fit to data to learn its parameters. This perspective, as outlined in (Cava et al. 2021; Makke and Chawla 2023), aims at learning both model structure and parameters.

Physics case study

The field of high-energy physics (HEP) aims to understand and discover the universe by investigating the structure and the formation of matter through particle collisions. In an example HEP process, e.g., deep inelastic scattering, an elementary particle, such as an electron (e^-), scatters off a composite particle (with substructure), such as a proton. The electron interacts with the proton by exchanging a photon[1] with one quark inside of it. The struck quark subsequently hadronises into charged hadrons[2] in the final state through the hadronization mechanism, as illustrated in Fig. 1. The collision is described by kinematic variables that are defined from the energies of the incoming and the scattered electrons. Final-state hadrons are described by two variables: the fractional energy (z) and the transverse momentum (p_T) which is defined by the projection of the hadron's momentum ($\vec{p}_h$) onto the direction of the exchanged photon. Whereas the theory of quantum electrodynamics fully describes the elementary interaction between the electron and the struck quark (eq), the hadronization mechanism can not be determined by theoretical calculations and has to be learned from data. A way to quantify this process is to measure the distribution of charged hadrons produced in the scattering events, commonly called hadron spectra. These are accessible in different HEP processes and are either independently fit or appropriately fit altogether in a so-called "global fit" using a pre-defined functional form, where the

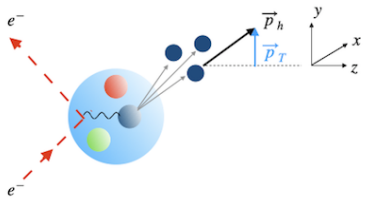

Figure 1: A sketch of the deep inelastic scattering process, where an electron (e^-) scatters off a proton (blue circle). The struck quark (blue ball inside the circle) hadronises into hadrons (h) that are detected in the final state. The hadron transverse momentum ($\vec{p}_T$) is the projection of the hadron's momentum ($\vec{p}_h$) with respect to the direction of the exchanged photon (z-axis).

XAI4Sci: Explainable machine learning for sciences, AAAI-24 (xai4sci.github.io)

[1]The photon is the messenger of the electromagnetic force.

[2]A hadron refers to a composite particle in the Standard Model (SM) of particle physics and consists of two or three quarks.

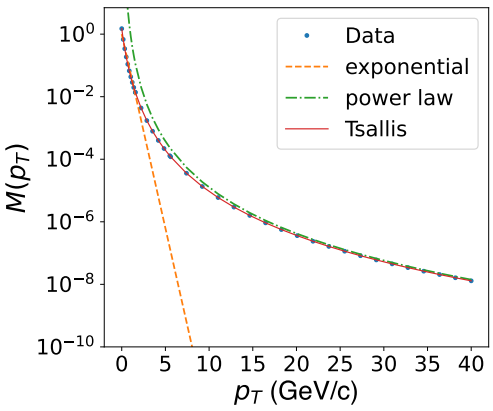

Figure 2: Sketch of a p_T spectra of charged hadrons (blue markers) in comparison with (i) a pure exponential function (orange dashed line), (ii) a power-law function (green dash-dotted line), and (iii) the Tsallis distribution (red solid line).

numerical values of the function's parameters are fit using data. Despite the significant progress made in the last decade, the hadronization mechanism remains poorly understood, particularly when considering hadron transverse momentum. This study investigates whether an analytical model can be learned directly from hadron p_T-spectra and how it compares with the functional forms traditionally used to fit these observables.

Shape and Functional form

The p_T spectra of charged hadrons exhibit distinct p_T-dependence over the full p_T range covered by experimental data. This shape is an essential and intriguing aspect of hadron production in the study of particle collisions and reflects the interplay between various production mechanisms. In the low-p_T region, the spectra exhibit an exponential form ($\exp(-p_T/\alpha)$), where α is a fit parameter, highlighting that hadrons are predominantly produced through thermal processes following a statistical distribution described by the Boltzmann-Gibbs statistics. In the high-p_T region, the spectra deviate from the exponential form and exhibit instead a power-law behavior (p_T^{-n}), n is often referred to as the "power-law index." This is typically associated with non-thermal hard processes, such as parton-parton[3] interactions. The curves of Fig. 2 illustrate both behaviors. Whereas neither one of these functions fully captures the data across the entire range of p_T, the so-called Tsallis distribution (Tsallis 1988) provides an exceptionally accurate description. The latter was introduced by C. Tsallis in an attempt to generalize the Boltzmann-Gibbs statistics to describe long-range correlations in collisions of high-energy particles. It is defined as:

$$f(p_T) = A\left(1-(1-q)\frac{p_T}{T}\right)^{1/(1-q)}. \qquad (1)$$

[3]A parton is either a quark or a gluon

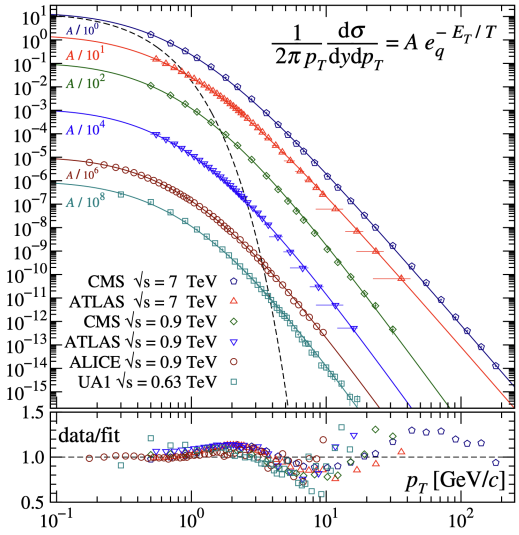

Figure 3: Comparison (Wong et al. 2015) of the Tsallis function (full line) with the experimental (markers) transverse momentum distributions of hadrons in pp collisions. The pure exponential fit is illustrated by the dashed curve.

where T can be physically interpreted as the temperature of a thermal distribution and q is the inverse slope parameter. This is equivalent to the power-law formula introduced by Hagedorn (Hagedorn 1983):

$$f(p_T) = Ap_T\left(1+\left(\frac{p_T}{p_0}\right)^2\right)^{-n} \qquad (2)$$

This is approximately $\exp(-n\frac{p_T}{p_0})$ at low-p_T and p_T^{-n} at high-p_T. Therefore, this function fully describes both the exponential decay in energy in the low-p_T region and the power-law tail in the high-p_T region. More importantly, it adequately describes over 14 decades of magnitude from the lowest to the highest p_T spanned by the transverse momentum spectra of charged hadrons measured at different energy scales, as shown in Fig. 3. Both equations (1,2) have been extensively used in phenomenological analyses of multiparticle production in high-energy proton-proton and heavy ion collisions at SPS, RHIC, and LHC experiments.

Symbolic regression

Symbolic regression (SR) is a subfield of machine learning aiming to learn analytical forms of underlying models from data, where both the model's structure and parameters are simultaneously learned. SR reduces to discovering a unary-binary tree[4] of mathematical symbols that are compatible with data. In such trees, internal nodes represent functions and leaf nodes represent variables or constants, as illustrated in Fig. 4. The importance of this representation is that any

[4]Encoding mathematical expressions as computational trees was introduced by Koza (Koza 1989) in his pioneering work on genetic programming.

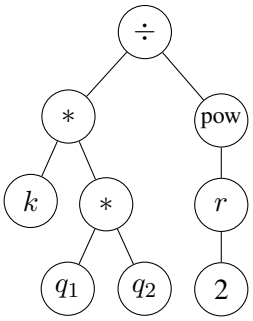

Figure 4: (a) Expression-tree structure of the Coulomb force $F_e = kq_1q_2/r^2$, which measures the interaction between two electrically charged particles q_1 and q_2 distant by r.

tree can be traversed into a unique sequence of symbols[5], allowing for its use in sequence-to-sequence models. The optimization problem in SR is defined over the space of mathematical expressions, composed from a set of allowable mathematical operators commonly called the library, e.g., $\mathcal{L} = \{x, \text{add}, \text{sub}, \text{mul}, \text{div}, \text{sqrt}, \text{pow}, \cos, \exp, \text{etc.}\}$, defined in accordance with the problem. The SR problem is very challenging given the discrete nature of the search space and its size which grows exponentially with model complexity. To further demonstrate this point, consider the Coulomb force problem, $F_e = kq_1q_2/r^2$, which consists of eight symbols, and a library consisting of 20 mathematical operations. Fitting the data set with a naive brute-force search will have to consider up to $20^8 = 2.5 \times 10^{10}$ candidate solutions without accounting for the optimization of numerical constants. The number of trials increases with model complexity (i.e., longer formula), making SR an "NP-hard" problem (Virgolin and Pissis 2022). SR can be tackled with various approaches based on genetic algorithms and deep learning among others, as summarized in (Cava et al. 2021; Makke and Chawla 2023). The deployment of genetic algorithms allows the exploration of a large set of mathematical equations through multiple generations. Each generation of models is evaluated based on a fitness function and those demonstrating better performance in terms of this metric are retained and used as a starting point for the next iteration, from which new models are derived by using traditional genetic operations such as mutation, copy, and cross-over.

Analysis

This analysis applies SR to experimentally measured p_T spectra of charged hadrons. The primary dataset used in this analysis is the semi-inclusive measurement of deep-inelastic scattering, where a lepton scatters off a target proton, and a final-state hadron is detected in coincidence with the scattered lepton. Measured observables are the differential multiplicities of charged hadrons, measured as a function of p_T^2 across simultaneous intervals of three additional variables: x, Q^2, and z. Part of this dataset is presented in Fig. 5 in a specific z range. Each group within the plot represents a distinct x range and encompasses multiple spectra corresponding to various Q^2 intervals within that x range. The full dataset consists of a total of 81 kinematic intervals, resulting in 4918 data points, as reported in (Phy 2018). The observed patterns reveal that the shape of the hadron multiplicities exhibits substantial sensitivity to variations in x, while its dependence on Q^2 is comparatively weaker. In light of these observations, various dataset configurations are considered by selecting distinct sets of variables (i.e., features). Finally, the transformer-based symbolic regression method, NeSymReS (Biggio et al. 2021), is selected and applied to the data. NeSymReS is the (first) SR method using a transformer neural network, which consists of an encoder-decoder structure. The open source code of NeSymReS includes a pre-trained model that was trained on 100 millions equations, which is used in the present analysis. The use of this SR method is driven by the fact that transformers learns causal relationships between the input features.

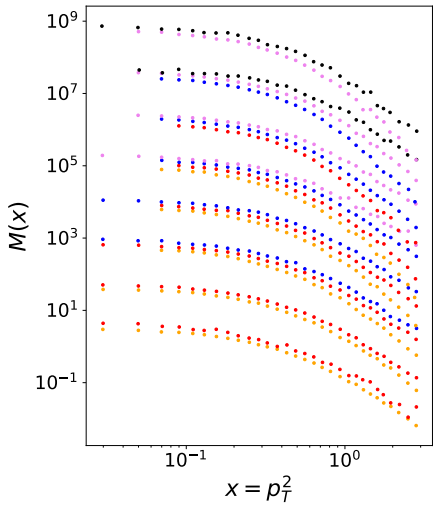

Figure 5: Charged hadron multiplicities as a function of p_T^2 in various x and Q^2 ranges. Each group represents a specific x range and comprises various Q^2 ranges denoted by different colors. The color code for Q^2 ranges is: {1: orange, 2: red, 3: blue, 4: violet, 5: black}. The same color code is applied to all x ranges.

Results

In the following, the dataset $\mathcal{D}$ will be defined in each section, and reported results are obtained using SR, namely NeSymReS, by independently applying it to each dataset, unless explotherwise stated. We will represent p_T^2 by x for simplicity.

Full p_T range

1. 1D-configuration, $\mathcal{D} \equiv \{p_T^2, M^h\}$: In this case, the full dataset includes 81 separate subsets. The most frequently learned functions in these subsets are summarized in Tab. 1. Although none of the learned functions provide

[5]This is called "polish form"

a fair and full description of the datasets, they commonly share a basic structure that may be written as:

$$f(\mathrm{x}) \propto (1 + c\mathrm{x}^n)^{-1}, \quad n = 2, 3 \tag{3}$$

2. 2D-configuration $\mathcal{D} \equiv \{Q^2, p_T^2, M^h\}$: This case includes 32 separate data subsets. The top performing functions, which provide a fair description of data, are two, but only one is valid in terms of dimensional analysis and is given by:

$$f(\mathrm{x}) = (1 + c_0\mathrm{x}^3)^{-1} \tag{4}$$

Truncated p_T range

In this section, we consider the full dataset $\mathcal{D} \equiv \{p_T^2, M^h\}$, i.e., 81 data subsets. The full p_T^2 range covered in the hadron spectra is decomposed into low-p_T^2 and large-p_T^2 ranges, and SR (NeSymReS) is independently applied on each dataset with truncated p_T range. Results are summarized in the following:

1. $p_T^2 < 0.5$: The top learned function is an exponential function such as $\exp(-\mathrm{x})/\mathrm{x}^2$, $c_0 \exp(c_1\mathrm{x})/\mathrm{x}$. Notably, an exponential form was not frequently learned, although expected, and a combination of exponential and trigonometric functions, e.g., $\exp(\sin(c\mathrm{x}))/\mathrm{x}$ was learned in numerous cases with a good data-fit match. However, these functions are discarded because data do not exhibit any periodic behavior.
2. $p_T^2 > 0.5$: The top performing functions obtained using SR by discarding the low-p_T^2 region (i.e., $p_T^2 < 0.5$) are reported in table 2 for different ranges of z. The function that describes most datasets across various z ranges is:

$$f(x) = c_0(1 + c_1\mathrm{x}^3)^{-1} \tag{5}$$

for which the values of the loss function are significantly better (from 2 up to 4 orders of magnitude difference) than those obtained for $f(\mathrm{x}) = (1 + c_0\mathrm{x}^3)^{-1}$. This function was also learned by considering the full p_T^2 range (f_3 in table 1), except that it describes enormously better data with $p_T^2 > 0.5$. This observation is interesting; SR could correctly learn the model's basic structure regardless of the data-fit match quality.

3. $p_T^2 > 1$: The most commonly learned functions in this range are summarized in table 3. The first four functions describe at best datasets with $p_T^2 > 1$. It's noteworthy that the first function that is explicitly a power-law function is expected to be the underlying function in the high p_T region, and the third function was already learned in the previous extractions (cf. tables 1, 2). Finally, the last function is learned in more than 50% of the datasets, however, it does not correctly describe the shape of the hadron spectra, mostly because of the absence of constants to be optimized.

	$f(\mathrm{x})$	Loss range	nb. of finding
f_1	$1/(1 + c\mathrm{x}^3)$	[1.17-3]	17
f_2	$1/(1 + c\mathrm{x}^2)$	[1.9-4.35]	9
f_3	$c_0/(1 + c_1\mathrm{x}^3)$	[0.018-0.025]	12

Table 1: Functional forms learned using SR on 81 data subsets where the full p_T range is considered ($\mathrm{x} \equiv p_T^2$).

$f(\mathrm{x})$	$z_1(18)$	$z_2(22)$	$z_3(22)$
$1/(1 + c_0\mathrm{x}^3)$	8	11	-
$c_0/(1 + c_1\mathrm{x}^3)$	8	4	2
$c_0/(1 + \mathrm{x}^3)$	-	-	12

Table 2: Functional forms learned by applying SR on 81 data subsets with a truncated p_T range ($p_T^2 > 0.5$). The numerals represent the number of findings for each function, and the total number of subsets in each z range is indicated in parentheses in the heading of the table. ($\mathrm{x} \equiv p_T^2$).

Discussion

The learned function that best describes datasets across multiple selections of variables and in spanning different regions of p_T by decomposing the full p_T range was found to be:

$$f(\mathrm{x}) = c_0 \left(1 + c_1\mathrm{x}^3\right)^{-1} \tag{6}$$

which can be written as:

$$f(\mathrm{x}) = c_0 \left(1 + \left(\frac{\mathrm{x}}{c_1}\right)^3\right)^{-1} \tag{7}$$

The form of this learned function closely resembles the Tsallis distribution (Eq. 1), where c_0 serves as the normalization constant analogous to A, c_1 corresponds to the Tsallis parameter resembling T, and $-n$ corresponds to the exponent parameter equivalent to $1/(1-q)$. The key distinction with the Tsallis distribution lies in the placement of the exponent parameter, which is associated with the variable itself rather than the sum. This finding is very important; we could learn, directly from data, a fundamental functional form that (i) describes the full p_T range with appropriate values of its free parameters, (ii) is similar to Tsallis distribution, which was theoretically developed from Boltzmann-Gibbs statistics, and (iii) can be regarded as a generalization of the fundamental structure $(1 + \mathrm{x}^n)^{-1}$, which has been learned in the majority of the cases.

As a final remark, this analysis did not consider experimental uncertainties of measured data (statistical and systematic uncertainties), and the check for dimensional analysis was performed after the learning procedure, which solidifies the result because considering experimental uncertainties constantly improves the findings.

Conclusion

This paper discusses the application of symbolic regression on a real physics problem, yet under investigation in multiple high-energy physics experiments and by theorists. Two goals were aimed: (i) to check if a mathematical expression

	$f(\mathrm{x})$	nb. of finding
1	$c_0\mathrm{x}^{c_1}$	8
2	$1/(1+c_0\mathrm{x}^3)$	1
3	$c_0/(1+c_1\mathrm{x}^3)$	1
4	$c_0\mathrm{x}^{c_2}/(1+c_1\mathrm{x})^{c_2}$	2
5	$(1+\mathrm{x}^n)^{-1}, \quad n=3,4$	32

Table 3: Functional forms learned by applying SR on 81 data subsets with a truncated p_T range ($p_T^2 > 1$). ($\mathrm{x} \equiv p_T^2$).

can be inferred from data and (ii) how it would compare to existing formulae. The result is promising and highlights that interpretable machine learning, through symbolic regression, can be used to boost scientific discovery by learning simple mathematical expressions directly from data. This result also emphasizes that interpretable machine learning would be used to guide research in theoretical physics.

References

2018. Transverse-momentum-dependent multiplicities of charged hadrons in muon-deuteron deep inelastic scattering. *Phys. Rev. D*, 97: 032006.

Biggio, L.; Bendinelli, T.; Neitz, A.; Lucchi, A.; and Parascandolo, G. 2021. Neural Symbolic Regression that Scales. *CoRR*, abs/2106.06427.

Cava, W. G. L.; Orzechowski, P.; Burlacu, B.; de França, F. O.; Virgolin, M.; Jin, Y.; Kommenda, M.; and Moore, J. H. 2021. Contemporary Symbolic Regression Methods and their Relative Performance. *CoRR*, abs/2107.14351.

Hagedorn, R. 1983. Multiplicities, p_T Distributions and the Expected Hadron $\rightarrow$ Quark - Gluon Phase Transition. *Riv. Nuovo Cim.*, 6N10: 1–50.

Koza, J. R. 1989. Hierarchical Genetic Algorithms Operating on Populations of Computer Programs. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'89, 768–774. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Makke, N.; and Chawla, S. 2023. Interpretable Scientific Discovery with Symbolic Regression: A Review. arXiv:2211.10873.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 1135–1144. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning Important Features through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 3145–3153. JMLR.org.

Tsallis, C. 1988. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52(1).

Virgolin, M.; and Pissis, S. P. 2022. Symbolic Regression is NP-hard. arXiv:2207.01018.

Wong, C.-Y.; Wilk, G.; Cirto, L. J. L.; and Tsallis, C. 2015. From QCD-based hard-scattering to nonextensive statistical mechanical descriptions of transverse momentum spectra in high-energy pp and $p\overline{p}$ collisions. *Phys. Rev. D*, 91: 114027.