Learning functional forms of fragmentation functions for hadron production using symbolic regression

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

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

Abstract

Hadron production is involved in all high-energy physics experiments and is a key element to understanding the formation of the visible matter in the universe. Fragmentation functions (FFs) are required to quantitatively describe hadron production; the key limitation is that they can not be calculated in theory and have to be extracted from experimental data. FFs are traditionally determined by fitting experimental observables with pre-assumed functional forms. This study is the first to infer, directly from data, a functional form of fragmentation functions using a machine learning technique, namely symbolic regression. The learned function significantly describes data, resembles the Lund string FF, and is generalizable; thus, it could be a potential candidate for use in global fits of FFs.

1 Introduction

The traditional approach in physical sciences relies upon the complementarity between theory and experiments, meaning that theoretical predictions shall be validated by experimental observations and vice versa. Physical sciences aim to understand the universe at different scales, ranging from celestial objects and galaxies (i.e., astronomy and cosmology) to elementary particles (e.g., quarks in particle physics). High-energy physics (HEP) investigates the universe at the lowest length scale. It focuses on studying different aspects of the strong nuclear interaction by colliding different types of elementary and composite particles, as well as nuclei, such as lepton-proton, proton-proton, and nucleus-nucleus. In such collisions, hadrons (e.g., particles composed of three quarks and their antimatter counterparts, antiquarks) are produced in the final state and thus used to measure various physical observables, such as particle distributions, differential cross sections, etc. Measured physical observables are then compared to theoretical calculations to either confirm or reject a given assumption or theory. For an exemplary process, say electron-proton (ep) scattering, the interaction consists of two parts. First is the interaction between elementary particles, i.e., the incoming electron emits a photon, W or Z boson that interacts with one quark (q) inside the proton, subsequently followed by the hadronization of the scattered quark into final-state hadrons. Analogously, the cross section of electron-proton scattering can be written as:

$$\mathrm{d}\sigma^{ep\to hX} \propto \mathrm{d}\sigma^{eq\to e'q} \times D_q^h \tag{1}$$

where $\mathrm{d}\sigma^{eq\to e'q}$ is the elementary cross section, i.e., describes the interaction between elementary particles (e, q), D_q^h are the so-called fragmentation functions (FFs) that quantitatively describe the transition from quarks to final-state hadrons (h). Whereas the elementary cross section can be calculated in Quantum Electrodynamics (QED) theory, FFs can not, and their determination fully counts on experimental measurements. The current methodology relies on global FF fits [1, 2], where a pre-assumed functional form of FFs is fit to a wide range of observables (e.g., differential cross sections and multiplicities) measured in various high-energy physics processes (ep, pp, etc.), to learn its parameters by involving the so-called DGLAP evolution equations [3] to take into account the different energy scales of the experimental measurements (e.g., $\sqrt{s} \approx 10^9 - 10^{12}$ electron volts).

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

In the fast-evolving era of AI, a basic question would be whether a functional form of FFs could be learned directly from data using machine learning (ML) instead of assuming a function and, most importantly, if the learned function could be interpretable and human-understandable to consider it in global FF fits. This is fortunately possible with symbolic regression (SR) [4–6]. The latter has shown remarkable potential in learning succinct mathematical equations directly from data and is proving to be a potential candidate for an automated scientific discovery tool. SR is particularly suited for physics applications since physical phenomena are described by mathematical equations. Its application to experimental data, however, is very limited [7, 8], in particular in high-energy physics [9], and was generally deployed on synthetic datasets in the majority of SR methods and applications.

Fragmentation functions represent an interesting and challenging application of ML simply because they can not be calculated in theory, and are essential to describe hadron production in all HEP processes.This study aims to infer a functional form of FFs from experimental (naturally noisy) data, without explicit assumptions like in the traditional approach, and compare it with established ones. From a physics perspective, it is very intriguing to check what data reveals about FFs. From a technical perspective, this study could be regarded as a test of the credibility of SR as a scientific discovery tool in noisy data environments, such as in physical sciences and more generally natural sciences. FFs represent, from a phenomenological point of view, the probability of a particular parton q to transform into a charged hadron h carrying a fractional energy z. In global FF fits [1, 2], an FF is parameterized by $D_q^h \propto z^\alpha (1-z)^\beta$, where α and β are fit parameters. The component $(1-z)^\beta$ constraints the FFs at $z=1$ such that $D_q^h(z=1)=0$. This functional form is inspired by the Lund symmetric fragmentation function from the "Lund string model" [10] of hadronization and is given by:

$$f(z) \propto (1/z)(1-z)^\alpha \exp(-\beta m_h^2/z) \tag{2}$$

Where α and β are parameters that should be tuned to reproduce data, and m_h is a mass term. It is worth noting that this study does not replace or eliminate the need for global FF fits; it rather complements them by suggesting a functional form of FFs that originates from experimental data.

2 Datasets and Methods

The dataset comprises differential multiplicities of charged hadrons [11, 12] of different species measured in semi-inclusive deep inelastic scattering at the COMPASS experiment [13] at CERN. They are measured as a function of z, the hadron's fractional energy, in bins of the kinematic variables x, the Bjorken scaling variable, and y, the virtual photon transfer momentum. This dataset plays an instrumental role in constraining FFs in the global fits [14] thanks to its richness, where multiplicities are presented in a very detailed binning in the relevant kinematic variables. From a technical point of view, it encompasses multiple subsets that reveal a consistent fundamental structure while spanning diverse regions in the phase space. This mirrors multiple instances of SR to the same physics problem but with distinct data points. In addition, the effectiveness of the results can be easily verified for generalization within the same dataset and extended to other datasets, given the universality of FFs.

Symbolic regression aims to simultaneously learn both models' structure and parameters directly from data,in contrast to traditional deep learning methods where only models' parameters are optimized. A powerful representation of an equation is the unary-binary tree [15] of mathematical symbols, which in turn allows one to express any equation as a sequence of symbols by traversing its tree, referred to as the Polish notation [16]. The latter is a mathematical notation in which operators precede their operands, e.g., $F_e = kq_1q_2/r^2 \equiv \{/, *, k, *, q_1, q_2, \mathrm{pow}, r, 2\}$. The optimization problem in SR is defined over a discrete space of mathematical expressions, composed from a user-defined set of allowable mathematical operators, commonly referred to as a "library", e.g., $\mathcal{L} = \{\mathrm{add},\ \mathrm{sub},\ \mathrm{mul},\ \mathrm{etc.}\}$, and, in general, it has been shown to be an "NP-hard" problem [17].

SR methods have significantly evolved from traditional search-based approaches (e.g., heuristic search and evolutionary algorithms) to modern learning-based (e.g., transformer-based language models) and hybrid techniques, as reviewed in [4]. We specifically choose the NeSymReS [18] method based on an encoder-decoder transformer architecture [19]. Transformers were originally developed in natural language processing (NLP) to learn the context in text data by introducing the so-called attention blocks into a standard NN's architecture. The outstanding performance of transformers has quickly expanded their use beyond NLP to general sequential data, including time-series data. The choice of a transformer-based SR method is mainly driven by the fact that learning the context in data holds

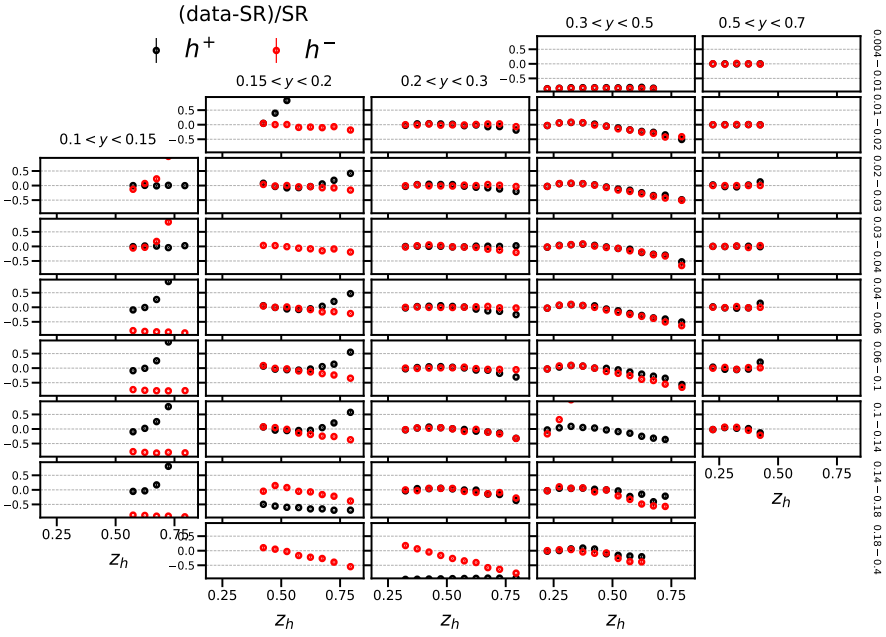

Figure 1: "(Data-SR)/SR" comparison for $h^{\pm}$ as a function of z. Data is experimental $M^h(z)$ [11], and SR denotes predicted $M^h(z)$ using models learned by SR in individual kinematic bins. Columns represent the five y bins, and rows represent the nine x bins as indicated on the right-hand side.

significant meaning in physics, particularly in light of the causal nature of physical phenomena, where capturing correlations among variables is crucial. We use the NeSymReS model that is pre-trained on 100 million equations. In the pre-training phase [18], equation skeletons (with numerical constant replaced by placeholders) and the inputs are generated from a sampling distribution $P_{e,X}$, where $e \equiv f_e : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ and $X \equiv \{x_i\}_{i=1}^n$, such that a training input is (X, y, e). The encoder maps a data set (X, y) into a latent vector z which then passes through the decoder. The latter creates a sequence of symbols in an auto-regressive manner, i.e., each generated symbol is appended to the input such that the next symbol is generated based on the new context. The model is trained to reduce the average loss between the predicted skeleton and the original one. At the end, the predicted skeleton equation is converted into a "proper" equation by replacing the constant tokens ("o") with their numeric counterparts using non-linear optimization.

3 Results and Discussion

We perform the SR analysis using differential multiplicities of positively (h^+) and negatively (h^-) charged hadrons. Each dataset comprises distributions of points as a function of z, i.e. $M^h(z)$, in nine x bins and up to five y bins within each x-bin. Figure 1 illustrates the performance of the models learned by SR in terms of a "(data-model)/model" comparison as a function of z in individual (x, y) bins for h^+ and h^-. In each kinematic bin, a model is independently inferred from the set of experimental data points using SR. Different equations are repetitively learned across the y bins. The top performing functions that remarkably describe the z dependence of $M^{h^{\pm}}$ and quantitatively match the data are learned in the third y bin, i.e., $0.2 < y < 0.3$, with loss values of 10^{-5} and up to 10^{-3}, except for the last x bin where $a = 1$. These functions are $f_1(z) = a\exp(-bz)/z^2$ and $f_2(z) = a\exp(-bz)/(z - c)$. The bins where the ratio is partially shown or fully missing refer to cases for which normalization factors are missing in the inferred models, i.e., $a = 1$.

The function $f_2(z)$ outperforms $f_1(z)$ in describing the z dependence of $h^{\pm}$ multiplicities, particularly at high z ($z > 0.5$). The best description of the data is obtained for h^- in the range $0.02 < x < 0.03$,

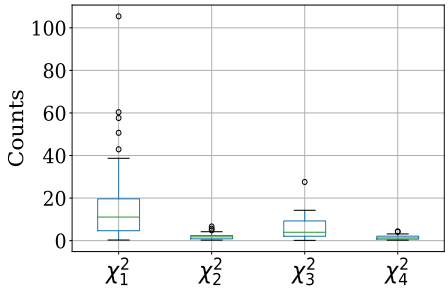

Figure 2: Comparison of the data fits using the functions listed in Eq. 3 for positive hadron multiplicities [11]. χ^2_i denotes the χ^2/ndf values obtained using $g_i(z)$. The box delimits the first and the third quartiles, whereas the middle line represents the median. The bottom and top lines represent, respectively, the minimum and maximum values in the χ^2/ndf values. Markers show the outliers (values significantly smaller or larger than median values).

with a loss of 5.8×10^{-5}, and parameters $a \approx -9.8$, $b = -7.7$, and $c \approx 1$. This result is equivalent to the equation $f(z; x_3, y_3) = a(1-z)^{-1}\exp(-bz)$, which we will refer to as $f_3(z)$. The latter significantly resembles the Lund FF (Eq. 2) in the exponential component and the term $(1-z)^\alpha$ with $\alpha = -1$. The multiplicities in the corresponding bin, i.e., $0.2 < y < 0.3$ and $0.02 < x < 0.03$, represent the "training" data in the context of this study, where a pre-trained transformer network is used. The multiplicities in the remaining bins of the phase space can be considered "test" data. They allow us to evaluate the overall performance of the models learned during training on "unseen" data, i.e., data points that were not part of the "training" set, and to check the generalizability of the model across the whole phase space (i.e., other (x, y) bins), and to other related measurements (e.g., $\pi^\pm$ and $K^\pm$ multiplicities, measurements at different energy, various processes, etc.).

The traditional ML approach achieves this goal by using $f_3(z) \equiv f(z; x_3, y_3)$ with the same numerical values of the constants a and b to produce predictions on test data. For physics data, however, numerical constants in the models are expected to be physical constants, which do not necessarily have the same values across the whole phase space and may exhibit weak or strong dependence upon the kinematic variables. In fact, an important aspect of such multi-dimensional experimental measurements in physics is to investigate the dependence of physical constants upon the kinematic variables of interest in the measurement. Therefore, to evaluate the performance of the learned models on test data, we performed fits of $M^h(z)$ in individual kinematic bins. We consider the most frequently learned functional forms ($g_1 \equiv f_1, g_2 \equiv f_2$) and the top-learned function ($g_3 \equiv f_3$) associated with the lowest error. In addition, we consider a general form of f_3 by taking the power exponent in the term $(1-z)$ as a free fit parameter, which we refer to to as g_4. This choice is mainly driven by the existence of a power "2" exponent in the learned function f_1. Thus, merging f_1 and f_3 into a general form requires the freeing of the exponent parameter. The list of fit functions includes:

$$\begin{aligned} g_1(z) &= a\exp(-bz)/z^2 & \qquad g_3(z) &= a\exp(-bz)/(1-z) \\ g_2(z) &= a\exp(-bz)/(c-z) & \qquad g_4(z) &= a(1-z)^c\exp(-bz) \end{aligned} \tag{3}$$

The best fits of $h^\pm$ multiplicities are obtained using $g_2(z)$ and $g_4(z)$, both having an extra parameter c and an overall better description across the phase space obtained using $g_4(z)$. This is illustrated in Fig. 2, where a comparison of the range of χ^2/ndf values obtained using different fit functions (Eq. 3) is shown for h^+. Figure 3 (left) illustrates the quality of the fits using g_4 in terms of the distribution of the relative errors of the predictions, $(y_{\text{true}} - y_{\text{pred}})/y_{\text{pred}}$, where y denotes the multiplicity (M^h), "true" denotes experimental data, and "pred" denotes predictions produced using $g_4(z; \{a, b, c\}_{\text{fit}})$ in individual bins, for $h^\pm$ test data. Each dataset consists of about 280 data points. The quality of the fits reflects a remarkable description of the data and is comparable between positively and negatively charged hadrons. In addition, the fits describe remarkably well the z dependence of the multiplicities in all kinematic bins for h^+ and h^-. No systematic effect is observed in the z dependence of the relative errors in individual kinematic bins, as illustrated in Fig. 3 (right) for $h^\pm$, in contrast to global fits where this ratio shows a systematic slope in the z dependence at high z particularly in the range $0.2 < y < 0.5$. Figure 4 compares the z dependence of experimental data and predictions obtained using g_4 for h^+ in two exemplary x bins and five y bins within each. The fits significantly describe $M^h(z)$ in terms of numerical values and slope, and this observation is common across all (x, y) bins.

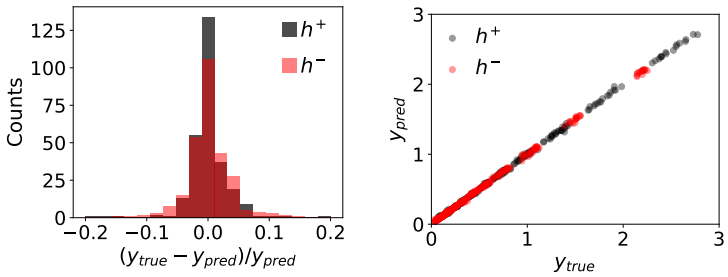

Figure 3: Left: distribution of relative errors of the fitted functions on the test set for $h^{\pm}$. Right: "true" vs. "prediction" scatter plot for positively (black) and negatively (red) charged hadrons.

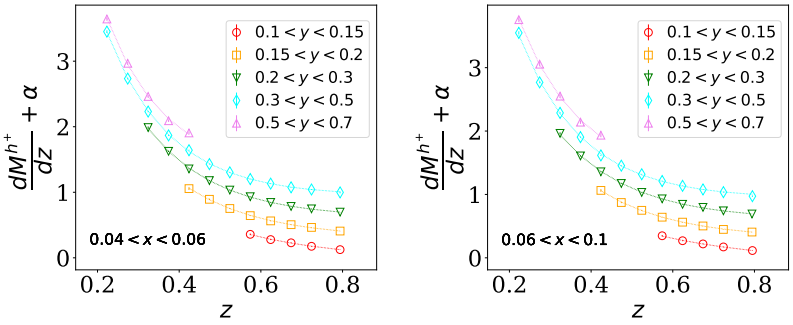

Figure 4: Comparison between experimental data (markers) and fits (curves) for positive hadron multiplicities [11], displayed as a function of z in five y bins (staggered vertically by $\alpha = 0.3$ for clarity) in two exemplary x bins, i.e., $x \in [0.04, 0.06]$ and $x \in [0.06, 0.1]$. Fits are performed in individual kinematic bins using the generalized SR model $g_4(z)$ (cf., Eq. 3). Statistical uncertainties are considered in the individual fits. The fits significantly agree with data.

To test the credibitily of the generalized learned model g_4 across all hadron species, we fit $\pi^{\pm}$ and $K^{\pm}$ multiplicities in individual (x, y) bins. The same observations made for h are made for identified hadrons, i.e., the fits describe remarkably well the z dependence of the multiplicities in all kinematic bins for both $\pi^{\pm}$ and $K^{\pm}$. The quality of the data fits is better for h and π ($\sigma_{\mathcal{N}} \approx 5\%$) compared to K ($\sigma_{\mathcal{N}} \approx 7 - 13\%$). This observation is understandable given that the unidentified hadron sample includes all hadron species, i.e., $h = \{\pi, K, p\}$ where p denotes protons, whereas pions represent around 70% and kaons only 20% of the total hadron sample.

4 Conclusion

Fragmentation functions are determined by global FF fits using pre-defined functional forms. This study is the first to infer a functional form of FFs directly from data using symbolic regression. The function inferred from the considered data set [11] is:

$$f^{\mathrm{SR}}(z) = a(1-z)^c \exp(-bz) \tag{4}$$

$f^{\mathrm{SR}}(z)$ resembles the Lund symmetric fragmentation function (Eq. 2) but distinct. The learned function is fit to data and found to describe them very well for all hadron species and across the whole phase space covered in the measurement. The learned function could be considered a potential candidate for an FF's parameterization in global FF fits. This would be a departure from traditional methodology, wherein both the model and its parameters originate from data. It is worthy of note here that, despite the fact that the SR approach, while powerful, may be more complex to implement compared to traditional fitting methods, SR does not replace fitting methods. The ultimate goal of the application of SR is to enable the discovery, in a purely data-driven manner, of formulae that can be further tested using traditional statistical techniques such as fitting data. Finally, this result positions symbolic regression as a promising tool for automated scientific discovery in noisy data environments and thus represents a novel approach to follow in QCD phenomenology studies.

5 Acknowledgments

We thank the referees at the Machine Learning and the Physical Sciences (ML4PS) Workshop at the Thirty-eight Conference on Neural Information Processing Systems (NeurIPS 2024) for comments and feedback on this work.

References

[1] de Florian D, Sassot R, Stratmann M. Global analysis of fragmentation functions for pions and kaons and their uncertainties. Phys Rev D. 2007 Jun;75:114010. Available from: https://link.aps.org/doi/10.1103/PhysRevD.75.114010.

[2] Borsa I, Sassot R, de Florian D, Stratmann M. Pion fragmentation functions at high energy colliders. Phys Rev D. 2022 Feb;105:L031502. Available from: https://link.aps.org/doi/10.1103/PhysRevD.105.L031502.

[3] Altarelli G, Parisi G. Asymptotic freedom in parton language. Nuclear Physics B. 1977;126(2):298-318. Available from: https://www.sciencedirect.com/science/article/pii/0550321377903844.

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

[5] 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.

[6] Makke, Nour and Chawla, Sanjay. A living Review of Symbolic Regression; 2022. Available from: https://github.com/nmakke/SR-LivingReview.

[7] Lemos P, Jeffrey N, Cranmer M, Ho S, Battaglia P. Rediscovering orbital mechanics with machine learning. Mach Learn Sci Tech. 2023;4(4):045002.

[8] 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.

[9] 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.

[10] Andersson B, Gustafson G, Ingelman G, Sjöstrand T. Parton fragmentation and string dynamics. Physics Reports. 1983;97(2):31-145. Available from: https://www.sciencedirect.com/science/article/pii/0370157383900807.

[11] Adolph C, et al. Multiplicities of charged pions and charged hadrons from deep-inelastic scattering of muons off an isoscalar target. Phys Lett B. 2017;764:1-10.

[12] Adolph C, et al. Multiplicities of charged kaons from deep-inelastic muon scattering off an isoscalar target. Phys Lett B. 2017;767:133-41.

[13] Abbon P, et al. The COMPASS experiment at CERN. Nucl Instrum Meth A. 2007;577:455-518.

[14] Borsa I, Stratmann M, de Florian D, Sassot R. Charged hadron fragmentation functions at high energy colliders. Phys Rev D. 2024 Mar;109:052004. Available from: https://link.aps.org/doi/10.1103/PhysRevD.109.052004.

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

[16] Robinson R. Jan Łukasiewicz: Aristotle's Syllogistic from the Standpoint of Modern Formal Logic. Second edition enlarged. Pp. xvi 222. Oxford: Clarendon Press, 1957. Cloth, 305. net. The Classical Review. 1958;8(3-4):282–282.

[17] Virgolin M, Pissis SP. Symbolic Regression is NP-hard. Transactions on Machine Learning Research. 2022. Available from: https://openreview.net/forum?id=LTiaPxqe2e.

[18] 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.

[19] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All You Need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 6000–6010.