Symbolic Machine Learning for High Energy Physics Calculations

Abdulhakim Alnuqaydan
Department of Physics and Astronomy
University of Kentucky, USA
Department of Physics, College of Science
Qassim University, KSA
aal700@uky.edu

Sergei Gleyzer
Department of Physics and Astronomy
The University of Alabama, USA
sgleyzer@ua.edu

Harrison B. Prosper
Department of Physics
Florida State University, USA
hprosper@fsu.edu

Eric A. F. Reinhardt
Department of Physics and Astronomy
The University of Alabama, USA
eareinhardt@crimson.ua.edu

Neeraj Anand
IIT Dhanbad, India
neerajanandfirst@gmail.com

Francois Charton
Meta AI
fcharton@meta.com

Abstract

The calculation of cross sections is of paramount importance in high-energy physics. Among other steps, this process involves squaring the particle interaction amplitudes, which can be very computationally expensive. These lengthy calculations are currently done using domain-specific symbolic algebra tools. We demonstrate that a transformer model, when trained on symbolic sequence pairs, can predict correctly the squared amplitudes of the Standard Model processes, namely QED, QCD and EW with an accuracy of 98%, 97% and 95%, respectively, at a speed that is up to six orders of magnitude faster than current symbolic computation frameworks. We briefly note some limitations of the model and suggest possible future directions for this work.

1 Introduction

In high-energy physics, the calculation of a cross section Goldberg (2017)—a measure of the probability of a given particle-particle interaction—can be an exceedingly complicated task requiring many mathematical operations, including Lorentz index contractions, Dirac algebra, traces, and integrals. Moreover, the complexity of these operations increases dramatically with the number of final state particles and the number of loops in the interactions.

These calculations have been automated in domain-specific symbolic algebra tools such as `FeynCalc` Shtabovenko et al. (2020), `CompHEP` Boos et al. (2004), and `MARTY` Uhlrich et al. (2021)). In this paper, we explore the possibility of using a machine learning model to handle these complex calculations. We use a *sequence-to-sequence* (`seq2seq`) model, specifically, a *transformer* Vaswani et al. (2017), to symbolically compute the square of the particle interaction amplitude, a key element in calculating cross-sections. Our aim is to invest time initially in training these models so that, in the long run, we can perform symbolic calculations much faster than with traditional software tools.[1]

[1] Code Repository: `https://github.com/ML4SCI/SYMBA_Pytorch`

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

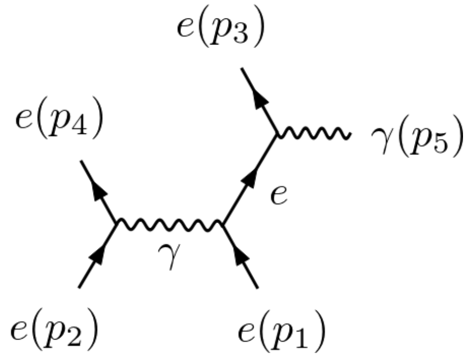

Figure 1: A Feynman diagram of two incoming electrons $e(p_1)$ and $e(p_2)$ scattering into two electrons and a photon.

- ***The amplitude*** $(e\ e \to e\ e\ \gamma)$:

$$i\mathcal{M} = \frac{\frac{1}{2}\,ie^3\big(p_{3\rho}\,\gamma^{\rho}_{\epsilon}\,\gamma_{\rho\eta}\,A^{\rho *}_{j}(p_5)\,\mathbf{e}^{*}_{i\eta}(p_4)\,\mathbf{e}^{*}_{l\epsilon}(p_3)\,\mathbf{e}_{k\delta}(p_2)\,\mathbf{e}_{i\delta}(p_1)\;-\;\frac{1}{2}\,p_{5\sigma}\,\gamma_{\rho\epsilon}\,\gamma^{\rho}_{\epsilon}\,\gamma_{\rho\eta}\,\gamma^{\sigma}_{\epsilon}\,A^{\rho *}_{j}(p_5)\mathbf{e}^{*}_{i\eta}(p_4)\,\mathbf{e}^{*}_{l\epsilon}(p_3)\,\mathbf{e}_{k\delta}(p_2)\,\mathbf{e}_{i\delta}(p_1)\big)}{((m_e^2 - \vec{p_2}\,.\vec{p_4})\;*\;\vec{p_3}.\,\vec{p_5})}$$

- ***The squared amplitude*** $(e\ e \to e\ e\ \gamma)$:

$$|\mathcal{M}|^2 = -\frac{e^6}{((\vec{p_3}.\vec{p_5})^2 * (m_e^2 - \vec{p_2}.\vec{p_4})^2)}(2m_e^6 + m_e^4 * (-\vec{p_1}.\vec{p_3} - \vec{p_1}.\vec{p_5} - \vec{p_2}.\vec{p_4} + 2\vec{p_3}.\vec{p_5}) + m_e^2 * (\vec{p_1}.\vec{p_2} * \vec{p_3}.\vec{p_4} +$$
$$\vec{p_1}.\vec{p_2} * \vec{p_4}.\vec{p_5} + \vec{p_1}.\vec{p_4} * \vec{p_2}.\vec{p_3} + \vec{p_1}.\vec{p_4} * \vec{p_2}.\vec{p_5} + \vec{p_1}.\vec{p_5} * \vec{p_3}.\vec{p_4} - \vec{p_2}.\vec{p_4} * \vec{p_3}.\vec{p_5}) - \vec{p_1}.\vec{p_2} * \vec{p_3}.\vec{p_5} * \vec{p_4}.\vec{p_5} - \vec{p_1}.\vec{p_4} * \vec{p_2}.\vec{p_5} * \vec{p_3}.\vec{p_5})$$

Figure 2: Amplitude and squared amplitude of the $ee \to ee\gamma$ scattering process.

2 Related Work

`Seq2seq` transformer-based models Vaswani et al. (2017) have been successfully deployed on a wide range of tasks including natural language processing OpenAI (2023); Devlin et al. (2018); Ott et al. (2018); Guan et al. (2020), other applications such as image captioning and medicine Liu et al. (2021); Devlin et al. (2018); Karpathy and Fei-Fei (2014); Li et al. (2019); Hatamizadeh et al. (2021); Ji et al. (2020) the current `seq2seq` state-of-the-art. These models have achieved excellent results in symbolic calculations in calculus and the symbolic solution of ordinary differential equations Lample and Charton (2019). Transformers have been used to infer the recurrence relation of underlying sequences of numbers d'Ascoli et al. (2022) and for symbolic regression Valipour et al. (2021). In physics, symbolic regression has been applied to classical mechanics problems Cranmer et al. (2019); Lemos et al. (2022); Udrescu and Tegmark (2019) and to extract optimal observable in the context of the Standard Model Effective Field Theory (SMEFT) Butter et al. (2021). Furthermore, transformer model have been used to simplify polylogrithmic functions, which appears in loop calculations in high energy physics Dersy et al. (2022). In the current work, we take a further step and show that transformers can accurately encode the mapping from amplitudes to their square averaged and summed over the particle degrees of freedom.

3 Background

The Standard Model, one of the great intellectual achievements of the 20th century, describes all known elementary particles and their interactions through three of the four known fundamental forces, namely, electromagnetic, weak, and strong forces (see, for example, Ref. Goldberg (2017) Schwartz (2013)). The Standard Model is a Quantum Field Theory (QFT) specified in a mathematical expression called a Lagrangian. In QFT, the elementary particles are described in terms of quantum fields in space-time, where each type of particle is associated with a different field. The interactions among these particles are governed by fields, which are sometimes referred to as force carriers, whose details are precisely determined by the symmetries imposed on the Lagrangian.

The Standard Model combines all the quantum field theories that describe the three forces, namely, the quantum electrodynamics (QED) which describe the electromagnetic interaction, quantum chromodynamics (QCD) which describe the strong interaction and electroweak (EW) which unifies electromagnetic and the weak interactions. Furthermore, the Standard Model provides an explanation for how the elementary particles acquire masses through a mechanism called "Higgs Mechanism". There is a well-defined procedure to extract from the Lagrangian all the possible particle interactions of interest, each associated with a mathematical quantity called an amplitude. These amplitudes can be represented by Feynman diagrams.

Calculating the cross section, for example of the process depicted in Fig. 1 and represented symbolically in Fig. 2, requires computing the squared amplitude and averaging and summing over the internal degrees of freedom of the particles.

Figure 2 shows one of the "shortest" $2 \to 3$ tree-level quantum electrodynamics (QED) squared amplitudes after simplification, where in this process two incoming electrons $e(p_1)$ and $e(p_2)$ scatter into two electrons $e(p_3)$ and $e(p_4)$ and a photon $\gamma(p_5)$. Typical expressions can have hundreds of terms and the computational time can become a major challenge, especially if higher-order (with loops) amplitudes are included to render predictions more precise.

The key insight in all the current uses of machine learning for symbolic applications is that many tasks can be viewed as a natural language processing problem. For example, a system that maps images to textual summaries of them can be viewed as translating from the language of images to a natural language. Likewise, algebraic manipulation such as the mapping of amplitudes to their squared form can be conceptualized as a language translation task. Since this task maps one sequence of symbols to another it is natural to consider `seq2seq` models.

4 Dataset

The symbolic sequence pairs, the amplitude and its square averaged and summed over initial and final particle degrees of freedom, respectively, are generated with `MARTY` Uhlrich et al. (2021) for interactions in quantum electrodynamics (QED), quantum chromodynamics (QCD) and electroweak (EW) Seq. 3. Here, we restrict the scope to 2-to-2 and 2-to-3 particle processes in all the Standard Model theories at tree-level, 2-to-2 in QED and QCD at one loop order. We consider two approaches: 1) mapping the amplitude to the squared amplitude as shown in Fig. 2, and 2) using the Feynman diagrams (written as a sequence) as input into the model to predict the squared amplitude. A notable advantage of the second approach is the Feynman diagram can be written by hand, if desired, without the need for a domain-specific tool to construct the amplitude.

All expressions are simplified with the `Python` symbolic mathematics module `SymPy` Meurer et al. (2017). We perform the tokenization using `torchtext` Paszke et al. (2019), that is, the assignment of an integer to each symbol and the padding of sequences to make them of equal length. Each sequence is then converted to a vector built from these integers. The amplitudes are tokenized by operator (tensor) and its indices, while for squared amplitudes we tokenize them by each mass, product of momenta, weak mixing angle for EW, and numerical factor (for example, $4 * m_e^2 * p_1.p_2$ is three tokens) as there are a finite number of terms consistent with the physical dimension (in powers of mass) and conservation laws.

For squared amplitudes coming from loop interactions, there are additional symbols corresponding to the n-point functions integrals that can be evaluated numerically with other tools such as `LoopTools` Hahn and Pérez-Victoria (1999). For practical computational reasons, we exclude expressions longer than 264 tokens after the simplification which excludes 5%, 26% and 12% of all QED, QCD and EW (2-to-3) tree-level expressions, respectively. For loop interaction, we exclude expressions longer than 500 which excludes 56%, 51% of all QED, QCD loop expressions. The data are split into three sets: training, validation and test, 70%, 15% and 15%, respectfully, and we choose a random sample of 500 expression pairs (from the test set) to evaluate the performance of the trained model.

5 Model and Training

The transformer model (see Ref. Vaswani et al. (2017)) consists of an *encoder* and a *decoder*. The encoder embeds the input sequence as a vector in a high-dimensional vector space and encodes the position of each token in the sequence. Next a mechanism called multi-head attention computes the degree to which a given token is related to other tokens. During the training of the model, the decoder takes the output vector from the encoder together with the target sequence one token at a time and predicts a sequence, also one token at a time. The transformer without structural modifications is implemented using `Pytorch` Paszke et al. (2019). The model has $1-2$ layers and 8 attention-heads, with 512 embedding dimensions and 2048 latent dimensions. *Cross-entropy* is used as the loss function and the `Adam` optimizer Kingma and Ba (2014) is used with a learning rate of 10^{-4} and a batch size of $64-512$. The training was performed for $(50-100)$ epochs on four NVIDIA A100 Tensor Core GPUs which took about $2-12$ hours.

Table 1: Model performance: sequence accuracy.

Training Sample	process	Training Size	Sequence Accuracy
QED (amplitude)	2-to-2 & 2-to-3	251K	98.6%
QCD (amplitude)	2-to-2 & 2-to-3	205K	97.4%
EW (amplitude)	2-to-2	258K	95.4%
EW (amplitude)	2-to-3	7M	94.4%
QED (diagram)	2-to-2 & 2-to-3	258K	99.0%
QCD (diagram)	2-to-2 & 2-to-3	250K	87.7%
EW (diagram)	2-to-2	259K	96.6%
EW (diagram)	2-to-3	7M	82.3%
QED (diagram)	2-to-2 (Loop)	13K	68.9%
QCD (diagram)	2-to-2 (Loop)	5.5K	60.0%

6 Results and discussion

Table 1 summarizes the performance of the model trained with different data sets. The accuracy of the predicted symbolic expressions is assessed by taking a random sample of 500 amplitudes (or diagram) from the test set that have not been used in the training of the transformer model and predicting their squared amplitudes. The sequence accuracy (one of three measures we explored) is the percentage of predicted symbolic expressions that exactly match the targets. The results demonstrate that even for a mathematical problem as complicated as squaring an amplitude, averaging and summing of over the internal degrees of freedom of particles, and manipulating the result into a meaningful form, it is possible to encapsulate that domain knowledge in a transformer model. The prediction time is up to 6 orders of magnitudes faster than `MARTY` for some amplitudes. It is noteworthy that the accuracy is affected by three factors: data complexity, data size and sequence length. The performance on QCD amplitude is much higher than on diagrams, which indicates the fact that the complexity in QCD, which comes from color factors and gluon self-interactions—components not present in QED, is higher, so expressing the input in Feynman diagram is not sufficient. A similar pattern is observed in the case of (EW) for 2-to-3 processes, where accuracy is lower when compared to models that rely on amplitude sequence information. In loop dataset, all of these three factors manifest which explains the lower accuracy. As there is a strong dependence on data set size, we expect better performance using a larger training dataset.

There are several ways to address the issues of data complexity including adding more details about the interaction, so the input data should encompass a comprehensive range of features, taking into consideration all complexities of interactions. The data size issue can be solved by including adding more processes from theories beyond the Standard Model (BSM) which exposes the model to a greater variety of examples. Longer sequence length expressions can be addressed with variants of the basic transformer model that exhibit better scaling with sequence length Beltagy et al. (2020). We leave that to future work. The length of the sequence can be further tuned by adjusting the tokenization process.

7 Conclusion

Our results demonstrate that a basic transformer model can encode the mapping between a particle interaction amplitude and its squared amplitude to high accuracy despite the complexity of the mapping. The accuracy of the transformer model is currently data limited and depends on the data complexity and sequence length. The results obtained, however, are sufficiently promising to motivate further work.

8 Acknowledgement

This work was supported in part by the U.S. Department of Energy (DOE) under Awards No. DE-SC0012447, DE-SC0023713 (SG) and DE-SC0010102 (HP). This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231 using NERSC award HEP-ERCAP0025759. (ER) and (NA) were participants in the ML4SCI 2023 Google Summer of Code program.

9 Broader Impact

The use of transformer models in high energy physics can speed up discoveries. Importantly it doesn't raise any big ethical or social concerns, so it's positive development for advancing our understanding of high energy physics.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. URL https://arxiv.org/abs/2004.05150.

E. Boos, V. Bunichev, M. Dubinin, L. Dudko, V. Ilyin, A. Kryukov, V. Edneral, V. Savrin, A. Semenov, and A. Sherstnev. CompHEP 4.4: Automatic computations from Lagrangians to events, 2004.

Anja Butter, Tilman Plehn, Nathalie Soybelman, and Johann Brehmer. Back to the formula – lhc edition, 9 2021. URL http://arxiv.org/abs/2109.10414.

Miles D. Cranmer, Rui Xu, Peter Battaglia, and Shirley Ho. Learning symbolic physics with graph networks, 9 2019. URL https://arxiv.org/abs/1909.05862.

Stéphane d'Ascoli, Pierre-Alexandre Kamienny, Guillaume Lample, and François Charton. Deep symbolic regression for recurrent sequences, 2022. URL https://arxiv.org/abs/2201.04600.

Aurélien Dersy, Matthew D. Schwartz, and Xiaoyuan Zhang. Simplifying polylogarithms with machine learning, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL https://arxiv.org/abs/1810.04805.

D. Goldberg. *The standard model in a Nutshell*. Princeton University Press, 2017. ISBN 9780691167596.

Wang Guan, Ivan Smetannikov, and Man Tianxing. Survey on automatic text summarization and transformer models applicability, 2020. URL https://doi.org/10.1145/3437802.3437832.

T. Hahn and M. Pérez-Victoria. Automated one-loop calculations in four and d dimensions. *Computer Physics Communications*, 118(2):153–165, 1999. ISSN 0010-4655. doi: https://doi.org/10.1016/S0010-4655(98)00173-8. URL https://www.sciencedirect.com/science/article/pii/S0010465598001738.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation, 2021. URL https://arxiv.org/abs/2103.10504.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome, 2020. URL https://www.biorxiv.org/content/early/2020/09/19/2020.09.17.301879.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions, 2014. URL https://arxiv.org/abs/1412.2306.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

Guillaume Lample and François Charton. Deep learning for symbolic mathematics, 12 2019. URL https://arxiv.org/abs/1912.01412.

Pablo Lemos, Niall Jeffrey, Miles Cranmer, Shirley Ho, and Peter Battaglia. Rediscovering orbital mechanics with machine learning, 2 2022. URL http://arxiv.org/abs/2202.02306.

Yikuan Li, Shishir Rao, Jose Roberto Ayala Solares, Abdelaali Hassaine, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: Transformer for electronic health records, 2019. URL https://arxiv.org/abs/1907.09538.

Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. CPTR: full transformer network for image captioning, 2021. URL https://arxiv.org/abs/2101.10804.

Aaron Meurer et al. Sympy: symbolic computing in python, January 2017. ISSN 2376-5992. URL https://doi.org/10.7717/peerj-cs.103.

OpenAI. Gpt-4 technical report, 2023.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Scaling neural machine translation, 2018. URL https://arxiv.org/abs/1806.00187.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Matthew D. Schwartz. *Quantum Field Theory and the Standard Model*. Cambridge University Press, 2013. doi: 10.1017/9781139540940.

Vladyslav Shtabovenko, Rolf Mertig, and Frederik Orellana. Feyncalc 9.3: New features and improvements, 2020. ISSN 0010-4655. URL https://www.sciencedirect.com/science/article/pii/S001046552030223X.

Silviu-Marian Udrescu and Max Tegmark. Ai feynman: a physics-inspired method for symbolic regression, 5 2019. URL http://arxiv.org/abs/1905.11481.

Grégoire Uhlrich, Farvah Mahmoudi, and Alexandre Arbey. – modern artificial theoretical physicist a c++ framework automating theoretical calculations beyond the standard model, 2021. ISSN 0010-4655. URL https://www.sciencedirect.com/science/article/pii/S001046552100062X.

Mojtaba Valipour, Bowen You, Maysum Panju, and Ali Ghodsi. Symbolicgpt: A generative transformer model for symbolic regression, 6 2021. URL http://arxiv.org/abs/2106.14131.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 6 2017. URL https://arxiv.org/abs/1706.03762.