Predictable scaling for amplitude surrogates

Anonymous Author(s)
Affiliation
Address
email

Abstract

Scaling laws describing the dependence of neural network performance on the amount of training data, the spent compute, and the network size have emerged across a huge variety of machine learning tasks and datasets. In this work, we systematically investigate these scaling laws in the context of amplitude surrogates for particle physics. We show that the scaling coefficients are connected to the number of external particles of the process. Our results demonstrate that scaling laws are a useful tool to achieve desired precision targets.

1 Introduction

High-energy physics experiments at the LHC produce petabyte-scale datasets, with volumes expected to grow by an order of magnitude at the High-Luminosity LHC. Exploiting these datasets requires fast and accurate tools for event simulation and inference. Machine learning (ML) has emerged as a powerful framework for this purpose [7, 25], with applications to phase-space integration, event generation, detector simulation, inference, and unfolding. One particularly active area of research is the development of amplitude surrogates [5, 4, 24, 27, 14, 15, 28, 17, 10], where neural networks are trained to regress the squared matrix element for a given process from a phase-space point. These surrogates offer the prospect of replacing costly loop-level amplitude evaluations with inexpensive network calls.

An important practical question for amplitude surrogates is how much training data, network size, and compute are required to reach a given precision. *Scaling laws* [1, 19, 21, 18, 26] answer this question empirically: the test loss typically decreases as a power law in these quantities over several orders of magnitude. Scaling laws are well established for language and vision models [21, 20, 2] and have recently also been observed for particle physics tasks [13–15, 28].

In this paper, we systematically investigate scaling laws for amplitude surrogates across eleven LHC processes and two Lorentz-invariant architectures: a multilayer perceptron using Lorentz invariants as inputs (MLP-I) and the Lorentz Local Canonicalization Transformer (LLoCa-Transformer) [28, 17]. Our main finding is that the scaling exponents have a lower bound that is determined by the intrinsic dimension of the amplitude manifold, which is fixed by the number of external particles. This yields a simple recipe to estimate the minimal resources needed to reach any desired precision target, using only a single low-cost training as an anchor. A comprehensive discussion of these results, including further details on the methodology and additional analyses, is presented in Ref. [11].

2 Scaling laws in deep learning

The test loss of a neural network is empirically well described by a power law [21, 26],

$$L(X, Y, Z) = (X_c/X)^{\alpha_X} + K_X(Y, Z) \ , \tag{1}$$

where (X, Y, Z) denotes any ordering of the network size N, the training-dataset size D_{train}, and the spent compute C. For a fixed X, the loss saturates at the plateau $K_X(Y, Z)$, which is set by

Submitted to the 2026 Conference on Physics and AI (PAI26). Do not distribute.

whichever of the other two variables is the bottleneck. For amplitude surrogates, the data are nearly noise-free, so $K_X \to 0$ in the large-(Y, Z) limit.

Following Refs. [26, 12], the scaling exponents can be predicted from the intrinsic dimension d of the data manifold. A ReLU network with N parameters partitions its input into roughly N linear pieces; if the data live on a d-dimensional manifold, the average piece side length scales as $s \propto N^{-1/d}$. The MSE loss is then upper-bounded by $N^{-4/d}$ assuming that the network learns the linear dependence of the target function up to a negligible error. An analogous argument can be applied to the scaling in dataset size considering the distance between data points. Additionally, we can conjecture the same scaling the compute by considering the dependence $C(N, D_{\text{train}}) \propto N \cdot D_{\text{train}}$. We emphasize that this is a rough estimate that assumes that the ratio between N and D_{train} does not affect performance. Summarizing all of these predicted scalings, we find

$$\alpha_X \gtrsim \frac{4}{d} \,. \tag{2}$$

For amplitude surrogates, d is the number of physical degrees of freedom in the scattering process, which depends on the number of particles in the final state n_f as $d = 3n_f - 4$. We check these predictions numerically below and cross-validate them using the twoNN method [22, 16, 3], which estimates the intrinsic dimension of the data from the last hidden layer representations.

3 Training setup

We work with two architectures that predict the squared matrix elements of a process from the momenta of the external particles. Both networks are Lorentz invariant, which is a property that has been observed to improve surrogate performance significantly. The MLP-I [15] ensures this by using momentum invariant pairs as inputs. The LLoCa-Transformer [28, 17] preserves the Lorentz symmetry by transforming each particle four-momentum into a learned local frame. The local features are then processed with a standard transformer backbone, which also enables permutation equivariance.

Our goal is to extract scaling laws in each of the variables N, D_{train}, and C. We use the maximal update parametrization [29, 30] to transfer the MLP-I learning rate across network sizes and a cosine-annealing schedule [23]. We scan $N \in [10^2, 10^6]$, $D_{\text{train}} \in [10, 10^6]$, and $C \in [10^{10}, 10^{17}]$ FLOPs in an effort to cover the widest range possible within a practical computing budget. When studying the scaling law in a given variable, we treat one of the remaining two as a secondary variable and scan over a few representative values, while fixing the third to the maximum value within its range. The datasets for all our trainings consist of unweighted samples that follow the physical distribution of the process.

4 Results

4.1 Case study: $q\bar{q} \to t\bar{t}H$

We first study $q\bar{q} \to t\bar{t}H$, a relevant process for Higgs sector analyses whose amplitude is expensive to compute on the multi-loop regime. Amplitude surrogates trained on this process have been observed to saturate past a certain threshold in the dataset size [15]. This process has $n_f = 3$ final-state particles, so the theoretical lower bound on all three scaling exponents is $\alpha_X = 0.8$. Fig. 1 shows the MSE test loss for the MLP-I as a function of the compute (left) and the dataset size (right). Both exhibit a clear power-law regime terminated by a plateau set by the complementary variable. The dataset size is by far the dominant performance driver: for $D_{\text{train}} \lesssim 10^4$ the loss saturates regardless of compute or network size, whereas for $D_{\text{train}} \simeq 8 \times 10^5$ the loss keeps decreasing up to 10^{17} FLOPs. This also explains the saturation reported in Ref. [15], whose compute budget was below the D_{train}-limited plateau. The scaling laws on N depend very much on D_{train}, with the performance quickly saturating for $D_{\text{train}} \leq 10^4$, so we choose not to show them. Fitting Eq. (1) yields the ranges $\alpha_C \simeq 2$–3, $\alpha_D \simeq 1.7$–1.9, and $\alpha_N \simeq 0.6$–1.9 over the considered secondary variables, all above the predicted lower bound. The last hidden layers also recover the correct $d = 5$ once $N \gtrsim 10^4$ and $D_{\text{train}} \gtrsim 10^2$.

Table 1 collects the fitted exponents together with the critical scale X_c and the plateau K_X for a representative subset of (Y, Z) choices. The plateau is indistinguishable from zero whenever the complementary variable is large, as expected for noise-free targets.

Table 1: Representative fitted scaling parameters for $q\bar{q} \to t\bar{t}H$ with the MLP-I trained on MSE loss.

param. X	2nd param.	X_c	α_X	K_X
C [FLOPs]	$D_{\text{train}} = 10^2$	$\left(2.58^{+0.64}_{-0.18}\right) \times 10^{11}$	3.67 ± 0.43	$\left(3.47^{+0.54}_{-0.21}\right) \times 10^{-1}$
	$D_{\text{train}} = 8.3 \times 10^5$	$\left(6.66^{+0.59}_{-0.31}\right) \times 10^{11}$	2.08 ± 0.032	$\left(19.81^{+0.21}_{-1.8}\right) \times 10^{-8}$
D_{train}	$N_{\text{MLP}} = 10^4$	$\left(6.51^{+0.51}_{-0.29}\right) \times 10^{1}$	1.811 ± 0.027	$\simeq 0$
	$N_{\text{MLP}} = 10^6$	$\left(5.36^{+0.40}_{-0.23}\right) \times 10^{1}$	1.928 ± 0.027	$\simeq 0$
N	$D_{\text{train}} = 10^2$	$595\ ^{+360}_{-5.1}$	0.601 ± 0.040	$\left(1.64\ ^{+0.72}_{-0.13}\right) \times 10^{-2}$
	$D_{\text{train}} = 10^6$	$\left(2.06\ ^{+2.0}_{-0.19}\right)$	1.850 ± 0.070	$5.97\ ^{+0.80}_{-0.34} \times 10^{-8}$

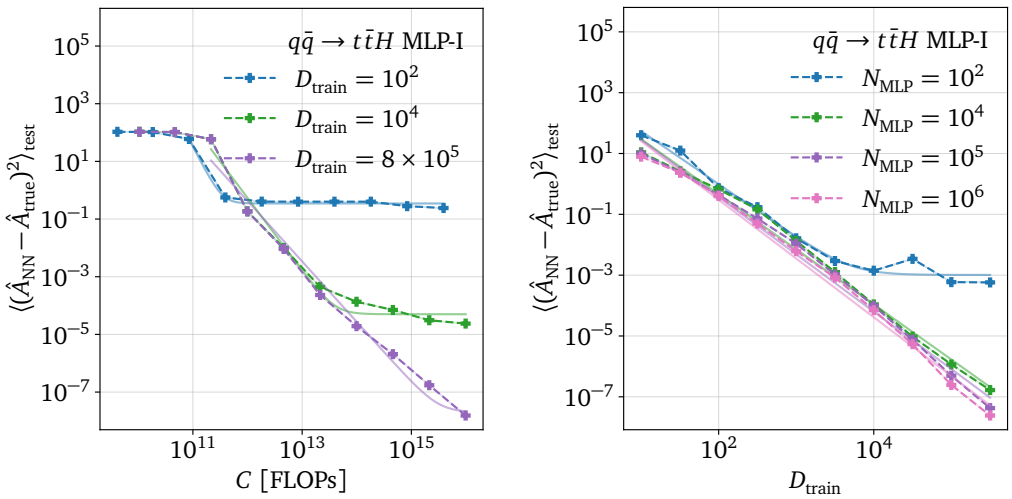

Figure 1: MSE test loss for $q\bar{q} \to t\bar{t}H$ with the MLP-I as a function of compute (left) and training dataset size (right). Power-law fits are overlaid as shaded lines.

We also trained a LLoCa-Transformer surrogate on this process and found a comparable asymptotic loss but a less predictable scaling, which we attribute to sub-optimal hyperparameter transfer in the absence of μP.

4.2 Scaling across processes

We repeat the analysis on ten additional processes that span different particle multiplicities, final-state species, and symmetry patterns: jet-associated Z production $q\bar{q} \to Z + ng$ with $n \in \{1, \dots, 4\}$ [14, 27, 17, 28], loop-induced jet-associated di-photon production $gg \to \gamma\gamma + ng$ with $n \in \{1, 2\}$ [4, 6, 8, 9], $q\bar{q} \to WZ + ng$ ($n \in \{0, 1, 2\}$) and $q\bar{q} \to WWZ$. The last two processes are studied here for the first time in this context. Fig. 2 illustrates the typical behaviour on the $q\bar{q} \to Z + ng$ family: the loss follows a power law, with slopes that clearly decrease as n grows. Identical qualitative behaviour is found for all other processes, with D_{train} again being the dominant factor.

Fig. 3 collects the extracted exponents α_D and α_C as a function of the number of final-state particles n_f, which determines $d = 3n_f - 4$. The exponents track the theoretical lower bound of Eq. (2): higher multiplicity yields smaller exponents, confirming that the intrinsic dimension is the main control parameter. Exponents fall below the bound only for $2 \to 2$ processes, which we attribute to the enhanced sensitivity to forward/backward scattering singularities at low n_f, where the Lipschitz assumption underlying Eq. (2) is locally violated [10]. We observe that switching to the LLoCa-Transformer on the $q\bar{q} \to Zgggg$ surrogate reduces X_c but leaves α_X essentially unchanged, in line with previous observations [28, 17]. We further verify that both architectures recover the expected intrinsic dimension $d = 3n_f - 4$ with as few as 10^4 parameters and 10^4 training points.

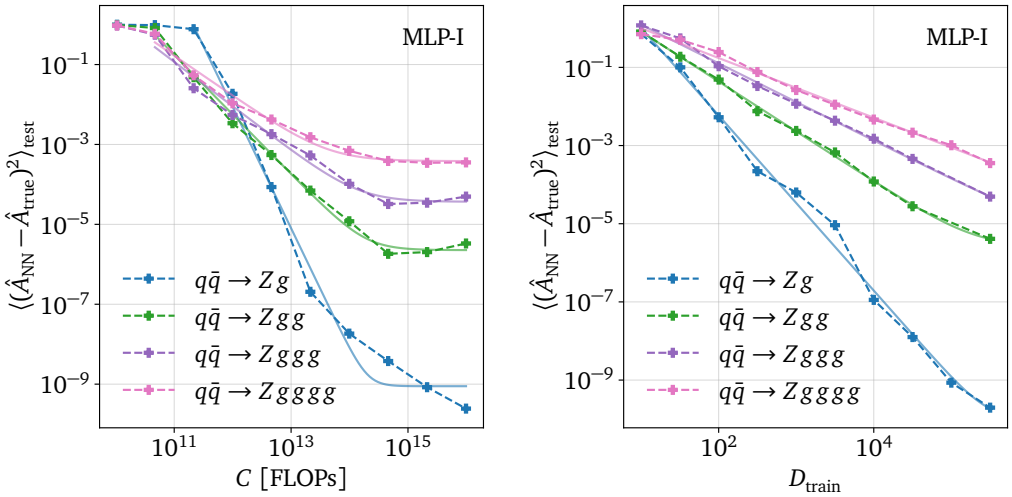

Figure 2: MLP-I MSE test loss for the $q\bar{q} \to Z + ng$ family as a function of compute (left) and training dataset size (right). The slopes decrease as the multiplicity n grows, reflecting the larger intrinsic dimension $d = 3n_f - 4$.

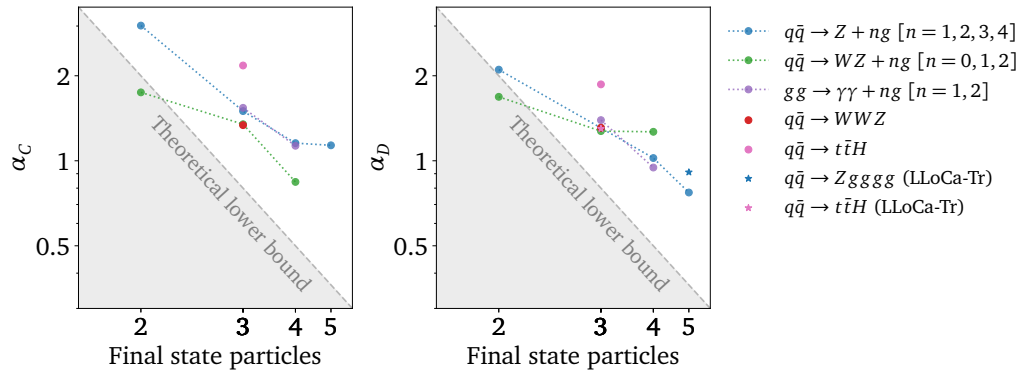

Figure 3: Fitted scaling exponents α_D (left) and α_C (right) as a function of the number of final-state particles n_f. Error bars are fit uncertainties; dashed lines show the theoretical lower bound $4/(3n_f - 4)$.

These observations translate into a simple recipe for planning amplitude-surrogate trainings. Since $\alpha_X \gtrsim 4/d$ is fixed by the known particle multiplicity, a single low-cost training at small $(N, D_{\rm train}, C)$ is sufficient to fix X_c provided that the other two training variables are fixed at sufficiently large values; the predicted loss at any larger budget then gives a conservative estimate of the resources needed to reach a target precision.

5 Conclusions

We established that amplitude surrogates follow clean power-law scaling in the training-dataset size, network size, and compute. Across eleven LHC processes and two architectures, the scaling exponents track the theoretical lower bound set by the intrinsic dimension of the particle phase space. A single low-cost training thus determines a worst-case estimate for the scaling laws. We emphasize that additional features of the amplitude datasets beyond the phase-space dimension are needed to predict the exact value of the scaling exponents. As a by-product, the dimension of the learned last-layer representations matches the real intrinsic dimension of the processes, suggesting that amplitude surrogates can double as estimators of the underlying data-manifold dimension. These results pave the way to systematic and predictable surrogate training for next-generation LHC event generators.

References

[1] Subutai Ahmad and Gerald Tesauro. Scaling and generalization in neural networks: A case study. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 1. Morgan-Kaufmann, 1988.

[2] Ibrahim M Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting vit in shape: Scaling laws for compute-optimal model design. *Advances in Neural Information Processing Systems*, 36:16406–16425, 2023.

[3] Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *arXiv e-prints*, art. arXiv:1905.12784, May 2019. doi:10.48550/arXiv.1905.12784.

[4] Joseph Aylett-Bullock, Simon Badger, and Ryan Moodie. Optimising simulations for diphoton production at hadron colliders using amplitude neural networks. *JHEP*, 08:066, 6 2021. doi:10.1007/JHEP08(2021)066.

[5] Simon Badger and Joseph Bullock. Using neural networks for efficient evaluation of high multiplicity scattering amplitudes. *JHEP*, 06:114, 2020. doi:10.1007/JHEP06(2020)114.

[6] Simon Badger, Anja Butter, Michel Luchmann, Sebastian Pitz, and Tilman Plehn. Loop amplitudes from precision networks. *SciPost Phys. Core*, 6:034, 2023. doi:10.21468/SciPostPhysCore.6.2.034.

[7] Simon Badger et al. Machine learning and LHC event generation. *SciPost Phys.*, 14(4):079, 2023. doi:10.21468/SciPostPhys.14.4.079.

[8] Henning Bahl, Nina Elmer, Luigi Favaro, Manuel Haußmann, Tilman Plehn, and Ramon Winterhalder. Accurate Surrogate Amplitudes with Calibrated Uncertainties. 12 2024.

[9] Henning Bahl, Nina Elmer, Tilman Plehn, and Ramon Winterhalder. Amplitude Uncertainties Everywhere All at Once. 8 2025.

[10] Henning Bahl, Jens Braun, Gudrun Heinrich, Tilman Plehn, and Rebecca Revelli. How to Trust Learned Loop Amplitudes. 1 2026.

[11] Henning Bahl, Victor Bresó-Pla, Anja Butter, and Joaquín Iturriza Ramirez. Scaling laws for amplitude surrogates. 1 2026.

[12] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024. doi:10.1073/pnas.2311878121.

[13] Joshua Batson and Yonatan Frederick Kahn. Scaling laws in jet classification. *SciPost Phys. Core*, 8:034, 2025. doi:10.21468/SciPostPhysCore.8.1.034.

[14] Johann Brehmer, Víctor Bresó, Pim de Haan, Tilman Plehn, Huilin Qu, Jonas Spinner, and Jesse Thaler. A Lorentz-Equivariant Transformer for All of the LHC. 11 2024.

[15] Víctor Bresó, Gudrun Heinrich, Vitaly Magerya, and Anton Olsson. Interpolating amplitudes. 12 2024.

[16] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7:12140, September 2017. doi:10.1038/s41598-017-11873-y.

[17] Luigi Favaro, Gerrit Gerhartz, Fred A. Hamprecht, Peter Lippmann, Sebastian Pitz, Tilman Plehn, Huilin Qu, and Jonas Spinner. Lorentz-Equivariance without Limitations. 8 2025.

[18] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

[19] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[20] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models. *arXiv e-prints*, art. arXiv:2203.15556, March 2022. doi:10.48550/arXiv.2203.15556.

[21] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[22] Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004.

[23] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

[24] Daniel Maître and Henry Truong. A factorisation-aware Matrix element emulator. *JHEP*, 11:066, 7 2021. doi:10.1007/JHEP11(2021)066.

[25] Tilman Plehn, Anja Butter, Barry Dillon, Theo Heimel, Claudius Krause, and Ramon Winterhalder. Modern Machine Learning for LHC Physicists. 11 2022.

[26] Utkarsh Sharma and Jared Kaplan. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34, 2022.

[27] Jonas Spinner, Victor Bresó, Pim de Haan, Tilman Plehn, Jesse Thaler, and Johann Brehmer. Lorentz-Equivariant Geometric Algebra Transformers for High-Energy Physics. 5 2024.

[28] Jonas Spinner, Luigi Favaro, Peter Lippmann, Sebastian Pitz, Gerrit Gerhartz, Tilman Plehn, and Fred A. Hamprecht. Lorentz Local Canonicalization: How to Make Any Network Lorentz-Equivariant. 5 2025.

[29] Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

[30] Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer. *arXiv e-prints*, art. arXiv:2203.03466, March 2022. doi:10.48550/arXiv.2203.03466.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The abstract and introduction state that we extract empirical scaling laws for amplitude surrogates across eleven LHC processes and two Lorentz-invariant architectures, and that the scaling exponents are controlled by the intrinsic dimension; these claims are directly supported by the results in Sec. 4.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: The main text explicitly notes that $2 \to 2$ processes fall below the theoretical lower bound due to forward/backward scattering singularities, that the LLoCa-Transformer shows less predictable scaling in the absence of μP, and that the bound $\alpha_X \gtrsim 4/d$ is a conservative estimate. Further limitations are discussed in the companion reference [11].
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification: The derivation of the lower bound $\alpha_X \gtrsim 4/d$ in Sec. 2 is presented as a heuristic argument that relies on ReLU piecewise-linear partitioning and local Lipschitz continuity of the target; both assumptions are explicitly stated, and the original derivations are cited [26, 12].
4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper?
   Answer: [Yes]
   Justification: Section 3 specifies the two architectures, the use of μP, cosine-annealing, the scan ranges in N, D_{train}, and C, and the dataset composition. Additional implementation details are provided in the companion reference [11].
5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results?
   Answer: [No]
   Justification: Code and data are not released with this submission. The companion reference [11] provides detailed descriptions about both the dataset generation and the training setup to ensure reproducibility.
6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details necessary to understand the results?
   Answer: [Yes]
   Justification: Section 3 describes the architectures, learning-rate parametrisation, schedule, and scan ranges; Section 4 specifies how secondary variables are treated in each scaling fit.
7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [Yes]
   Justification: Fit uncertainties on the scaling exponents are reported in Table 1 and shown as error bars in Fig. 3; they are the standard one-sigma uncertainties returned by the non-linear power-law fit.
8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources needed to reproduce the experiments?
   Answer: [Yes]
   Justification: The scan covers $C \in [10^{10}, 10^{17}]$ FLOPs per training, which is the direct measure of the compute cost of each run; hardware details are given in the companion reference [11].
9. **Code of ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?
   Answer: [Yes]
   Justification: The work uses only synthetic Monte-Carlo samples of particle-physics processes and involves no human subjects, personal data, or other content covered by the Code of Ethics.
10. **Broader impacts**
    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
    Answer: [NA]
    Justification: The work develops surrogate models for evaluating scattering amplitudes in high-energy physics simulations; we see no direct societal impact beyond accelerating fundamental-physics research.

11. **Safeguards**
    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse?
    Answer: [NA]
    Justification: The trained surrogates regress squared matrix elements for specific LHC processes and pose no foreseeable misuse risk.
12. **Licenses for existing assets**
    Question: Are the creators or original owners of assets used in the paper properly credited and are the license and terms of use explicitly mentioned and properly respected?
    Answer: [Yes]
    Justification: All prior architectures, methods, and datasets re-used in this work are cited at the point of use, including the MLP-I [15], the LLoCa-Transformer [28, 17], the μP parametrisation [29, 30], and the twoNN method [22, 16, 3].
13. **New assets**
    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
    Answer: [NA]
    Justification: No new assets are released with this submission; planned releases are described in the companion reference [11].
14. **Crowdsourcing and research with human subjects**
    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation?
    Answer: [NA]
    Justification: The work does not involve crowdsourcing or human-subject research.
15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether IRB approvals were obtained?
    Answer: [NA]
    Justification: The work does not involve human subjects and requires no IRB approval.
16. **Declaration of LLM usage**
    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?
    Answer: [NA]
    Justification: LLMs were not used as part of the core methodology; their use was limited to editorial polishing, for which declaration is not required.