Improving Compute-Optimal Scaling for Jet Classification with BSM Pretraining

Jan-Lucas Uslu
Department of Applied Physics
Stanford University
Stanford, CA 94305, USA
uslu@stanford.edu

Kevin Greif
Department of Physics and Astronomy
University of California, Irvine
Irvine, CA 92697, USA
kgreif@uci.edu

Benjamin Nachman
SLAC National Accelerator Laboratory
Stanford University
Menlo Park, CA 94025, USA
nachman@stanford.edu

Abstract

Neural scaling laws describe how model performance improves as a power law in compute, model size, and dataset size. While well-established for large language models, these relationships remain largely unexplored in physics applications. In this paper, we investigate how the composition of pretraining data affects compute-optimal scaling behavior for jet classification at the Large Hadron Collider. We pretrain transformer models on subsets of the JetClass-II dataset with varying fractions of Beyond the Standard Model (BSM) processes and fine-tune on the JetClass benchmark. We find that including BSM processes in pretraining progressively shifts the compute-optimal balance from the classic $\sim$50/50 allocation between model size and data toward a strongly data-favoring regime, with the magnitude of the shift governed by the BSM fraction in the pretraining corpus.

1 Introduction

Neural scaling laws Kaplan et al. [2020], Hoffmann et al. [2022] have become a central tool for predicting the performance of large neural networks and allocating compute efficiently. Deep learning is now a standard tool in many fields of physics. As frontier applications move toward large pretrained foundation models Mikuni and Nachman [2024], understanding how to allocate compute efficiently becomes critical. In this paper, we consider scaling laws in jet classification, a prototypical supervised training task in particle physics Qu et al. [2022], Kasieczka et al. [2019]. While scaling laws have been studied in this context Batson and Kahn [2025], Vigl et al. [2026], the role of *pretraining data composition* has received little attention.

Particle physics offers a unique advantage for studying this question. Unlike most domains, pretraining data can be generated through high-fidelity Monte Carlo simulations, allowing for fine control over the pretraining corpora composition. This unique capability allows us to study a general question about how pretraining data diversity shapes scaling behavior, with implications for applications of large pretrained models in other sub-fields.

Our hypothesis is that including structurally diverse BSM processes in pretraining should improve downstream scaling efficiency by exposing the model to a richer set of jet topologies during pretraining. BSM resonance decays produce multi-prong jet signatures (e.g. $X \to WW$, $X \to t\bar{t}$) that overlap with the downstream signal classes, potentially reducing the data bottleneck at fine-tuning time.

Conference on Physics and AI at Stanford University (PAI 2026)

Table 1: Compute-optimal scaling exponents from two independent methods. Uncertainties from bootstrap resampling (100 samples, 80% subsample fraction).

	Approach 1 (iso-compute)		Approach 2 (parametric)	
Pretraining subset	a (N^*)	b (D^*)	a (N^*)	b (D^*)
Scratch	0.517 ± 0.002	0.483 ± 0.002	0.548 ± 0.001	0.452 ± 0.001
QCD	0.526 ± 0.004	0.474 ± 0.004	0.425 ± 0.002	0.575 ± 0.002
QCD + res2p	0.332 ± 0.005	0.668 ± 0.005	0.288 ± 0.009	0.712 ± 0.009
QCD + res34p	0.309 ± 0.003	0.691 ± 0.003	0.289 ± 0.004	0.711 ± 0.004
QCD + res2p + res34p	0.206 ± 0.004	0.794 ± 0.004	0.249 ± 0.005	0.751 ± 0.005

This is consistent with prior work showing that pretraining effectively multiplies the fine-tuning dataset Hernandez et al. [2021], that training distribution changes scaling exponents Cherti et al. [2023], and that pretraining-downstream alignment determines whether scaling is beneficial Isik et al. [2025].

We test this hypothesis by pretraining transformers on subsets of the JetClass-II dataset Li et al. [2024] (188 classes spanning SM and BSM processes) with varying BSM fractions, and fine-tuning on the JetClass benchmark Qu et al. [2022]. We extract Chinchilla-style scaling exponents Hoffmann et al. [2022] using two independent methods and find that BSM pretraining progressively shifts the compute-optimal balance from the classic 50/50 allocation toward a strongly data-favoring regime.

2 Methods

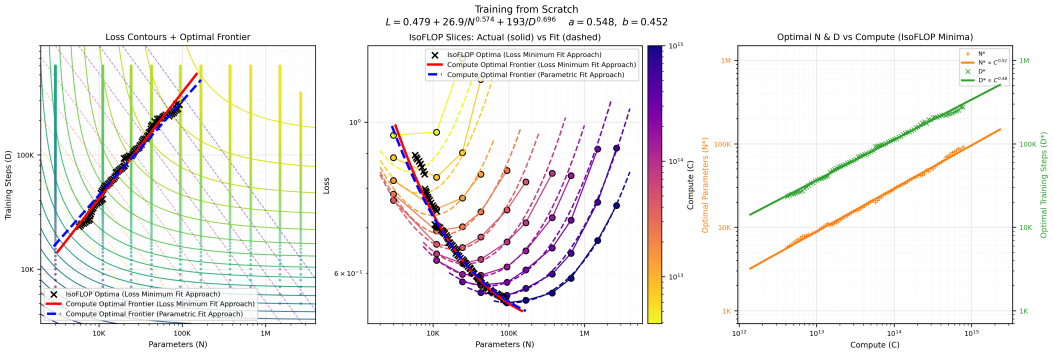
Model. We use a transformer for variable-length particle jet point clouds. Each jet is a set of constituent particles with 9 kinematic features. The model uses a learnable class token, pre-norm transformer blocks with SwiGLU Shazeer [2020] feed-forward networks, and no positional encodings. We fix depth at 4 layers and head dimension at 8, varying the embedding dimension from 8 to 256 to yield 12 model sizes from approximately 4K to 4.2M parameters.

Datasets. For pretraining, we use JetClass-II Li et al. [2024], grouping its 188 classes into three buckets: QCD jets (light quarks/gluons), 2-prong BSM resonance decays, and 3/4-prong BSM resonance decays. We define four pretraining subsets with increasing BSM content: **QCD** (QCD only), **QCD + res2p** ($\sim$40/60 QCD/res2p), **QCD + res34p** ($\sim$20/80 QCD/res34p), and **QCD + all** ($\sim$15/20/65 QCD/res2p/res34p). The downstream task is 10-class jet classification on JetClass Qu et al. [2022], following the single-epoch protocol of Vigl et al. Vigl et al. [2026].

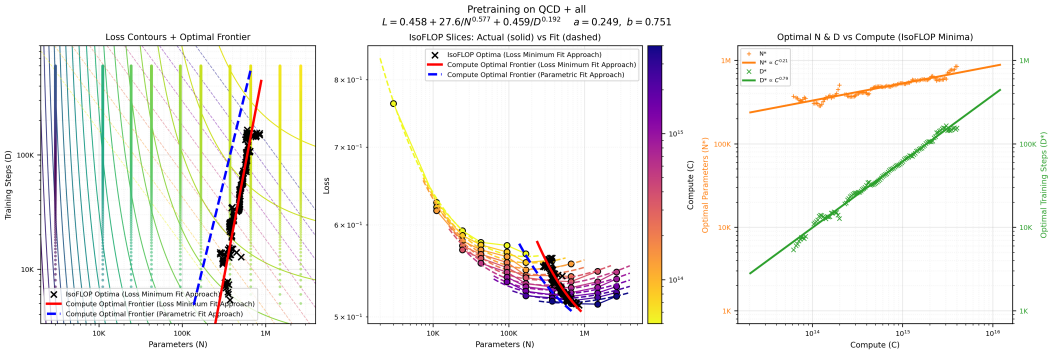
Training. Models are pretrained for 200K iterations (AdamW, lr 10^{-3}, batch size 128, 1K warmup steps), ensuring identical total token counts across configurations. Fine-tuning uses the same hyperparameters for 600K iterations with 5K warmup steps. Each configuration is trained across 5 independent seeds.

Scaling law extraction. Following Kaplan et al. Kaplan et al. [2020] and the Chinchilla framework Hoffmann et al. [2022], we estimate compute as $C = 6Nn_pBS$, where N is the number of model parameters, n_p=40 is the average number of constituents per jet, B=128 is the batch size, and S is the number of training steps. We use cross-entropy loss over the test set and extract compute-optimal scaling exponents a and b, defined by $N^* \propto C^a$ and $D^* \propto C^b$ with the consistency condition $a + b \approx 1$, via two independent approaches.

Approach 1 (iso-compute) is a non-parametric, bottom-up extraction. For each compute budget C, we construct an IsoFLOP profile by evaluating the loss achieved by each model size at the number of training steps corresponding to that budget. We fit a parabola in log-log space to each profile to locate the optimal model size N^* and dataset size D^* that minimize loss at that compute level. This yields a set of (C, N^*, D^*) triples from which we extract the scaling exponents by fitting power



(a) Scratch training.



(b) Full BSM pretraining (QCD + all resonances).

Figure 1: Scaling analysis for scratch training (a) and full BSM pretraining (b) on JetClass. **Left:** Loss contours $L(N, D)$ with the compute-optimal frontier from Approach 2 (blue dashed line) and extracted Approach 1 optima (crosses). The frontier tilts from a balanced N–D allocation (a) to a strongly data-favoring regime (b). **Center:** IsoFLOP slices comparing the parametric fit (dashed) to the observed loss (solid). **Right:** Power-law fits to the optimal N^* and D^* as a function of compute.

laws $N^* \propto C^a$ and $D^* \propto C^b$ using RANSAC regression for robustness to outliers. This approach makes no assumption about the functional form of the loss surface.

Approach 2 (parametric) is a top-down, global fit. We fit the parametric loss model

$$L(N, D) = L_\infty + \frac{A}{N^\alpha} + \frac{B}{D^\beta}, \tag{1}$$

to the full (N, D, L) dataset simultaneously, where L_∞ is the irreducible loss, A/N^α captures the model-capacity bottleneck, and B/D^β captures the data bottleneck. The compute-optimal scaling exponents are then derived analytically by minimizing L subject to the constraint $C = 6ND$, yielding $a = \beta/(\alpha{+}\beta)$ and $b = \alpha/(\alpha{+}\beta)$. Because this approach fits a global surface rather than pointwise optima, it can extrapolate beyond the observed compute range but depends on the assumed functional form.

The two approaches are complementary: Approach 1 is assumption-free but noisy, while Approach 2 is smoother but model-dependent. Agreement between them provides confidence that the extracted exponents are robust.

3 Results

Loss curves are smoothed, binned into 1,000-iteration windows, averaged across seeds, and monotonized via a running minimum following the Chinchilla protocol Hoffmann et al. [2022]. The first 5,000 iterations are discarded to exclude warmup transients.

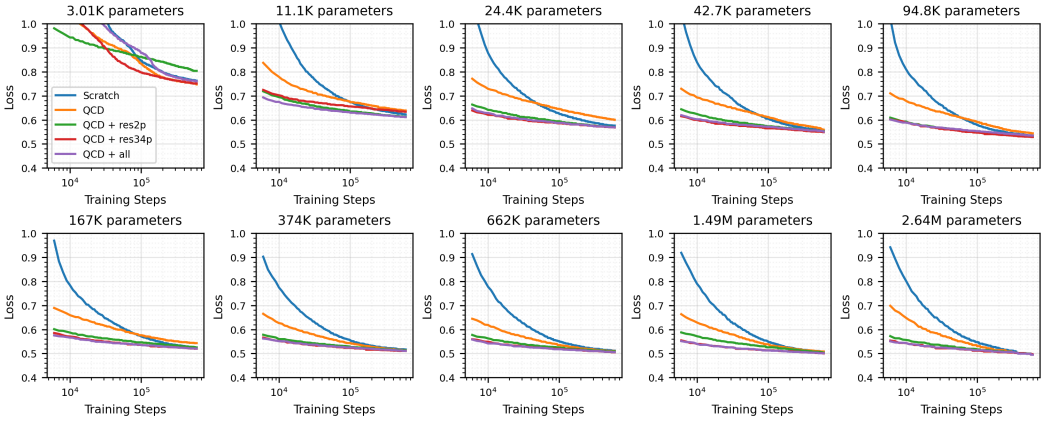


Figure 2: Fine-tuning loss curves as a function of training steps for each model size, comparing scratch training and all four pretraining configurations. BSM-enhanced pretraining consistently lowers the loss across all model sizes, with the benefit increasing for larger models.

Scratch baseline. Training from scratch, Approach 1 yields $a = 0.517 \pm 0.002$ and $b = 0.483 \pm 0.002$, reproducing the approximately equal allocation found by Hoffmann et al. Hoffmann et al. [2022]. Approach 2 gives consistent values ($a = 0.548 \pm 0.001$, $b = 0.452 \pm 0.001$). Figure 1a shows the analysis: the Approach 1 optima agree well with the Approach 2 frontier, and the parametric fit accurately reproduces the observed IsoFLOP profiles.

Effect of BSM pretraining. Comparing Figures 1a and 1b reveals the impact of pretraining on compute-optimal scaling. Table 1 summarizes the exponents: BSM pretraining progressively shifts the balance away from model size and toward data. QCD-only pretraining produces a marginal shift ($a = 0.526$ vs. 0.517), while adding 2-prong resonances shifts a to 0.332, and including all resonances pushes a down to 0.206, a factor-of-2.5 reduction. Both extraction methods agree to within ~ 0.05 for all BSM-containing subsets, confirming the shift is robust. Figure 2 shows that BSM pretraining consistently lowers the fine-tuning loss across all model sizes, with the benefit increasing for larger models.

4 Discussion and Conclusions

The key finding is that beyond simply lowering the loss, BSM pretraining reshapes the compute-optimal frontier. The shift from $a \approx 0.52$ (scratch) to $a \approx 0.21$ (full BSM) means that pretrained models favor investing additional compute in data rather than parameters. Notably, QCD-only pretraining barely shifts the exponents despite using the same total number of pretraining jets, indicating that it is the structural diversity of BSM resonance topologies that drives the effect rather than dataset size. This is consistent with our initial hypothesis: the multi-prong decay structures of BSM resonances overlap with the downstream signal classes, effectively reducing the data bottleneck at fine-tuning time. Further study is needed to disentangle the contributions of pretraining diversity versus objective alignment.

These results demonstrate that pretraining corpus design is important for scaling law behavior in physics applications. This is especially true in particle physics where accurate Monte Carlo simulations can be used to build the corpus. When diverse pretraining data is available, downstream compute budgets should be invested in training smaller models on more data rather than scaling up model size. Future work should vary the pretraining compute budget and test whether these findings generalize across different fine-tuning tasks beyond jet classification.

References

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

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii, and Sanmi Koyejo. Scaling laws for downstream task performance of large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

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.

Gregor Kasieczka et al. The machine learning landscape of top taggers. *SciPost Physics*, 7:014, 2019. doi: 10.21468/SciPostPhys.7.1.014.

Congqiao Li, Antonios Agapitos, Jovin Drews, Javier Duarte, Dawei Fu, Leyun Gao, Raghav Kansal, Gregor Kasieczka, Louis Moureaux, Huilin Qu, Cristina Mantilla Suarez, and Qiang Li. Accelerating resonance searches via signature-oriented pre-training. *arXiv preprint arXiv:2405.12972*, 2024.

Vinicius Mikuni and Benjamin Nachman. OmniLearn: A method to simultaneously facilitate all jet physics tasks. *arXiv preprint arXiv:2404.16091*, 2024.

Huilin Qu, Congqiao Li, and Sitian Qian. Particle transformer for jet tagging. In *International Conference on Machine Learning (ICML)*, 2022.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Matthias Vigl, Nicole Hartman, Michael Kagan, and Lukas Heinrich. Neural scaling laws for boosted jet tagging. *arXiv preprint arXiv:2602.15781*, 2026.