Double Descent and Overparameterization in Particle Physics Data

Matthias Vigl
Technical University of Munich
matthias.vigl@tum.de

Lukas Heinrich
Technical University of Munich
lukas.heinrich@tum.de

Abstract

Recently, the benefit of heavily overparameterized models has been observed in machine learning tasks: models with enough capacity to easily cross the *interpolation threshold* improve in generalization error compared to the classical bias-variance tradeoff regime. We demonstrate this behavior for the first time in particle physics data and explore when and where 'double descent' appears and under which circumstances overparameterization results in a performance gain.

1 Introduction

Particle physics aims to understand the smallest constituents of matter and their interactions and to search for new fundamental particles to explain unsolved phenomena such as the nature of dark matter. Due to the high-dimensional, heterogeneous, and multimodal data particle detectors produce, the computational analysis of the data relies on increasingly large neural network components for a wide range of tasks, such as jet-tagging [1, 2], particle reconstruction [3, 4], detector simulation [5, 6] and many more. Recently, with the advent of foundation models [7–12], the field explores larger-scale, higher capacity models in order to push towards ultimate experimental sensitivity.

A key question in scaling neural networks is to understand their generalization behavior. Perhaps most strikingly, it has been observed that heavily overparameterized model can outperform models in the "classical regime" due to the "double descent" [13] phenomenon. Thanks to *implicit bias* in the training dynamics [14] the generalization of neural networks improves again after initially worsening due to heavy overfitting close to the interpolation regime, where the training loss first vanishes. The detailed behavior, however, is not entirely universal: while model-wise double descent has been observed more universally across different architectures and learning settings, epoch-wise double descent tends to be highly dataset- and task-dependent. Some details, for example, whether or not the test-time performance in the overparameterized regime outperforms the best underparametrized model, or whether the double-descent phenomenon persists under early-stopping model selection, requires domain-specific investigation. The phenomenon has been observed in the domains of natural images [15], quantum systems [16] or protein folding [17] but was not yet studied in particle physics data.

In this paper, we investigate generalization and overparameterization for particle physics data. We show that generalization behavior is indeed not uniform for the domain itself but is quite sensitive to task and dataset, motivating more detailed future experiments. We show for the first time explicit instances of epoch-wise and model-wise double descent. Additionally, we show that indeed instances can be found in data-constrained regimes, where overparameterized models outperform models in the classical regime. The experiments were performed on publicly available data to facilitate future investigations of these phenomena in particle physics.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

2 Datasets and training setup

We investigate overparameterization on typical tasks in particle physics using two representative public datasets from the ATLAS collaboration [18]:

JetSet This dataset [19] consists of approximately 200M anti-k_{T} R=0.4 [20] "small-R" jets from simulated top quark pair production. The dataset contains event-level, jet-level, track-level and truth hadron information used for training the GN2 jet flavor tagger [21]. A detailed description of the dataset content can be found in the dedicated repository [1]

SUSY Wh1Lbb Channel Open Data Set This dataset [22] consists of approximately 12M simulated events of signal chargino decay and 14 Standard Model (SM) background processes. The dataset contains object-level and event-level variables with systematic uncertainties.

To disentangle the effect of model size on training dynamics and test set performance from the choice of optimizer parameters and learning rate schedulers, we use Adam optimizer [23] with no weight decay across all experiments and model sizes, set a common batch size of 128 and 10^{-4} learning rate: the learning rate is kept constant when training multilayer perceptrons (MLPs), while a learning rate warmup for 5% of the total straining steps is applied when training Transformer models.

3 Jet p_{T} Regression: Model and Epoch-wise Double Descent

Particle jets emerge from cascading interactions in quantum chromodynamics (QCD) and produce a large number of collimated particles. For data analysis, these particles are clustered using agglomerative clustering [20] into "jets" which are strongly related to the underlying generative process. Due to their high dimensionality, machine-learning approaches for particle jets are a major area of research in particle physics. Here, we demonstrate overparameterization behavior in the task of regressing the "transverse momentum", p_T, of the jet given the set of its constituent objects, i.e. the clustered tracks.

As a set-level prediction task, we choose a permutation equivariant two-layer transformer model, with embedding dimension d_{model}, here referred to as *model width*, used as the scaling parameter. The transformed set of tokens is then pooled and the final p_{T} regression is computed using an MLP head network. The models are trained on 40k jets and we use mean-squared error (MSE) as the standard regression loss function. The loss is monitored for up to 4k epochs at increasing model width values ranging from $d_{\mathrm{model}} \in [4, 2048]$.

The achieved train and test loss are shown in Figures 1a and 1b as a function of model capacity (i.e. model width) and the training epoch. Here, the typical features of the double descent phenomenon are for the first time visible in particle physics data. In the training loss, we can observe that for sufficiently large models the *interpolation threshold* can be reached with sufficiently many training epochs. That is, the train loss vanishes and the models "memorize" the 40k training samples.

The test loss exhibits model-wise double descent as shown in Fig. 1d and is expected from generalization theory. In the model-wise direction the test loss is worst at the interpolation threshold, where the model capacity is just big enough to interpolate the data but does not have capacity for the implicit bias of the optimization procedure to select well-generalizing functions among the set of interpolating functions. For large models, the test-time performance decreases again.

Notably, the task also exhibits epoch-wise double descent, as shown in Fig. 1c. The test performance first improves over the course of training before the model starts overfitting but at sufficiently high epoch times, after the training data is "memorized", the generalization performance gradually improves again, a phenomen different from *grokking* [24] which appears as a delayed phase transition after perfect memorization. The emergence of epoch-wise double descent is theoretically less understood but is hypothesized to be related to differing speeds at which task-relevant features are learned [15]. Understanding which aspects of particle jet data contribute to this phenomenon is an interesting future research direction.

For both types of double descent, the second descent *does not* outperform the test performance of the "classical regime", underlining again the dataset dependence of the benefit of overparameterization.

[1] JetSet repository: `https://gitlab.cern.ch/atlas/open-data/transforming-jet-flavor/-/blob/main/vars_open.md`

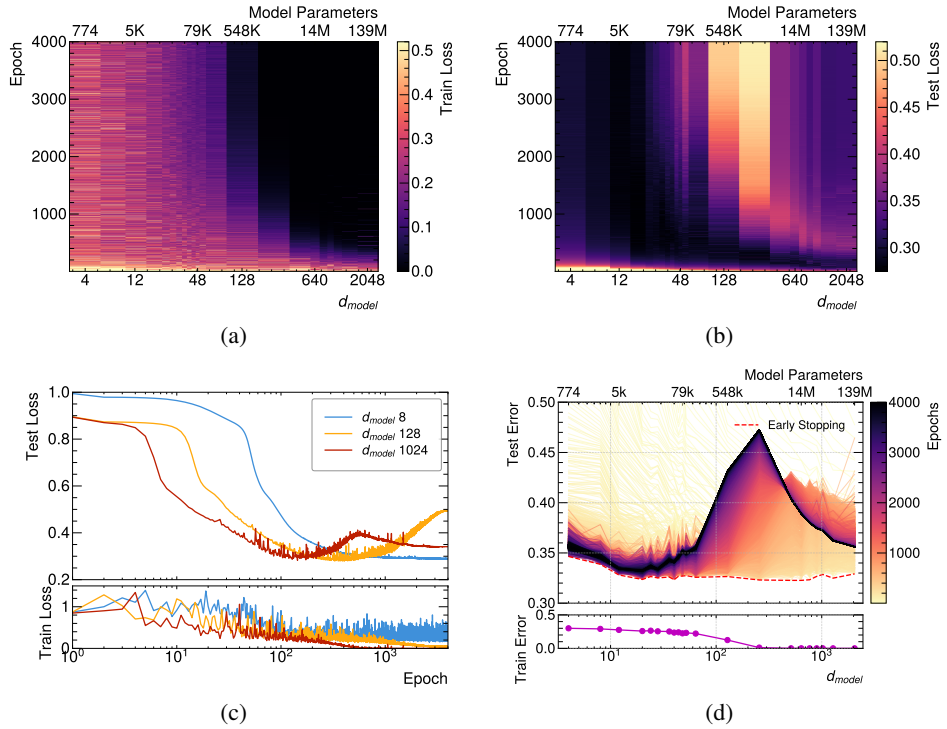

Figure 1: (a,b): Train and test jet p_{T} regression loss as a function of model size and epoch, model- and epoch-wise double descent can be observed looking at horizontal and vertical slices. (c): Vertical slices of plot (b) showing epoch-wise double descent. (d): Model-wise double descent.

4 Event classification: Prior Dependence and Early-Stopping Behavior

The rich context dependence of generalization performance of neural networks can be further investigated by comparing the behavior observed in jet p_{T} regression to that in high-level tasks such as event classification. In high energy physics event classification, the neural network is tasked with classifying a small set of high-level summary statistics extracted from high-dimensional multi-modal detector data with respect to their theoretical origin. A typical tasks is differentiating events likely to originate from known physics phenomena due to interactions described by the Standard Model of Particle Physics and those compatible with new phenomena such as Supersymmetry (SUSY) [25] that could offer explanations for e.g. dark matter. For this task, we train a simple 3-layer MLP network to classify events from the SUSY dataset into multiple classes using a standard categorical cross-entropy loss. The scaling dimension in this case is the width of the hidden MLP layers.

4.1 Prior Dependence of Epoch-wise Double Descent

First, we investigate the nature of epoch-wise double descent more closely and can correlate the phenomenon empirically to prior choices. To this end, the network is tasked to classify summaries of collider events into one of four possible classes which represent the signal process (a supersymmetric particle model) and three major background processes (single top production, top-anti-top quark production and W-boson production associated with jets). As a posterior prediction task, the trained classifier depends strongly on the prior distribution chosen among the four classes, which we observe to also have an impact on generalization performance.

In Fig. 2, we compare test-loss trajectories for a balanced dataset of 800k events and an unbalanced dataset of 7M events, where processes are sampled proportionally to their occurrence in the full simulation, across the four target classes. We see that epoch-wise double descent is not a universal feature but rather is observed only in the imbalanced and not in the balanced training. We hypothesize that the class imbalance may adversely affect the speed at which individual features are learned and thus lead to non-monotonic training trajectories.

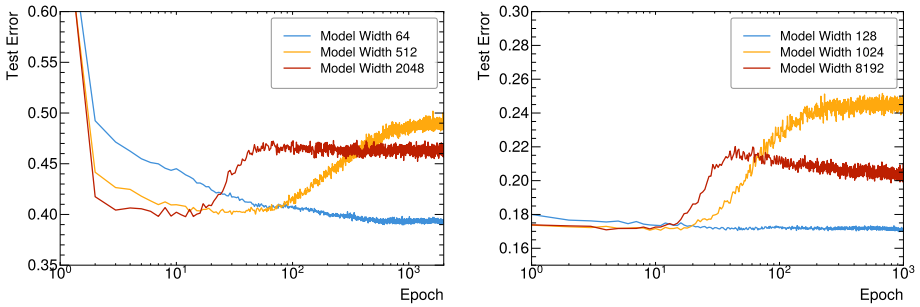

Figure 2: Large models trained on the 800k balanced dataset (left) do not exhibit epoch-wise double descent, whereas the same models trained on the 7M unbalanced dataset (right) do.

4.2 Early Stopping and Overparameterization Benefit

Second, we investigate model-wise generalization more closely and observe additional nuance emerge when considering early stopping, task complexity, and dataset size, i.e. when the number of classes and number of samples available per class is scaled up. We extend the balanced 4-way classification task to a 15-way classification taking into account a more complete set of Standard Model background processes. As in the previous cases, model-wise double descent is observed when training long past the interpolation threshold. In both cases, the high-capacity models do not improve over the classical regime when comparing the final test-time loss.

The picture changes when considering "early-stopping" in the sense of taking the epoch checkpoint with the best test-set performance. As shown in Fig. 3, for small datasets the generalization penalty at the interpolation peak disappears completely, and a monotonic improvement in test-time performance is observed for heavily overparameterized models. When increasing dataset size the monotonous behavior stops and a temporary increase in generalization error appears around the interpolation threshold, i.e. the double descent pattern emerges. However, unlike the previous examples, the high-capacity models clearly outperform models in the classical regime. In short, the recently observed benefit of training models with far more parameters than data samples can also be observed in particle physics data: a 1B parameter model outperforms classical models on a comparatively modest dataset.

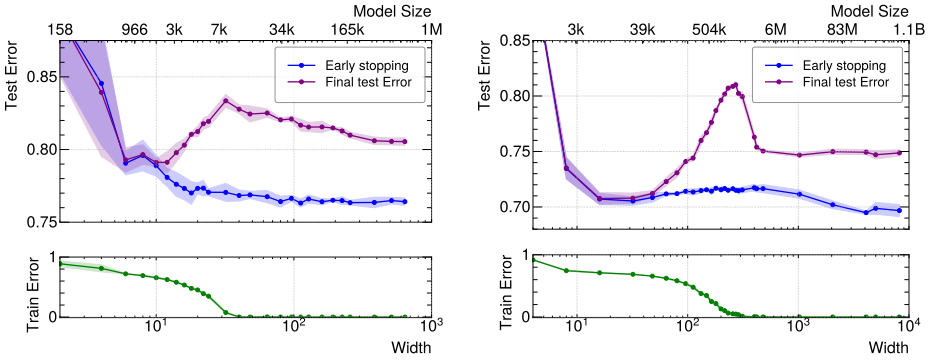

Figure 3: Model-wise double descent for a 3-layer MLP trained on 3k events (left) and a 16-layer MLP trained on 150k events (right). overparameterised models outperform underparameterised ones.

5 Conclusions

We investigated the behavior of under- and overparameterized models on high-energy physics data. We show that neural networks exhibit a rich range of behaviors: depending on context, both model-*and* epoch-wise double descent is observed both with and without early stopping. We identified cases where overparameterized models significantly outperform models from the classical regime. As the field moves towards higher-capacity models it is essential to further understand these phenomena.

References

[1] Gregor Kasieczka, Tilman Plehn, Anja Butter, Kyle Cranmer, Dipsikha Debnath, Barry M. Dillon, Malcolm Fairbairn, Darius A. Faroughy, Wojtek Fedorko, Christophe Gay, Loukas Gouskos, Jernej Fesel Kamenik, Patrick Komiske, Simon Leiss, Alison Lister, Sebastian Macaluso, Eric Metodiev, Liam Moore, Benjamin Nachman, Karl Nordström, Jannicke Pearkes, Huilin Qu, Yannik Rath, Marcel Rieger, David Shih, Jennifer Thompson, and Sreedevi Varma. The machine learning landscape of top taggers. *SciPost Physics*, 7(1), jul 2019. doi: 10.21468/scipostphys.7.1.014. URL https://doi.org/10.21468%2Fscipostphys.7.1.014.

[2] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. Energy Flow Networks: Deep Sets for Particle Jets. *JHEP*, 01:121, 2019. doi: 10.1007/JHEP01(2019)121.

[3] Joosep Pata, Javier Duarte, Jean-Roch Vlimant, Maurizio Pierini, and Maria Spiropulu. MLPF: efficient machine-learned particle-flow reconstruction using graph neural networks. *The European Physical Journal C*, 81(5), may 2021. doi: 10.1140/epjc/s10052-021-09158-w. URL https://doi.org/10.1140%2Fepjc%2Fs10052-021-09158-w.

[4] Francesco Armando Di Bello, Etienne Dreyer, Sanmay Ganguly, Eilam Gross, Lukas Heinrich, Anna Ivina, Marumi Kado, Nilotpal Kakati, Lorenzo Santi, Jonathan Shlomi, and Matteo Tusoni. Reconstructing particles in jets using set transformer and hypergraph prediction networks. *The European Physical Journal C*, 83(7), jul 2023. doi: 10.1140/epjc/s10052-023-11677-7. URL https://doi.org/10.1140%2Fepjc%2Fs10052-023-11677-7.

[5] Michela Paganini, Luke de Oliveira, and Benjamin Nachman. Calogan: Simulating 3d high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks. *Physical Review D*, 97(1), January 2018. ISSN 2470-0029. doi: 10.1103/physrevd.97.014021. URL http://dx.doi.org/10.1103/PhysRevD.97.014021.

[6] Baran Hashemi and Claudius Krause. Deep generative models for detector signature simulation: A taxonomic review. *Reviews in Physics*, 12:100092, December 2024. ISSN 2405-4283. doi: 10.1016/j.revip.2024.100092. URL http://dx.doi.org/10.1016/j.revip.2024.100092.

[7] Matthew Leigh, Samuel Klein, François Charton, Tobias Golling, Lukas Heinrich, Michael Kagan, Inês Ochoa, and Margarita Osadchy. Is Tokenization Needed for Masked Particle Modelling? 9 2024.

[8] Tobias Golling, Lukas Heinrich, Michael Kagan, Samuel Klein, Matthew Leigh, Margarita Osadchy, and John Andrew Raine. Masked particle modeling on sets: towards self-supervised high energy physics foundation models. *Mach. Learn. Sci. Tech.*, 5(3):035074, 2024. doi: 10.1088/2632-2153/ad64a8.

[9] Matthias Vigl, Nicole Hartman, and Lukas Heinrich. Finetuning foundation models for joint analysis optimization in High Energy Physics. *Mach. Learn. Sci. Tech.*, 5(2):025075, 2024. doi: 10.1088/2632-2153/ad55a3.

[10] Joschka Birk, Frank Gaede, Anna Hallin, Gregor Kasieczka, Martina Mozzanica, and Henning Rose. OmniJet-α_C: learning point cloud calorimeter simulations using generative transformers. *JINST*, 20(07):P07007, 2025. doi: 10.1088/1748-0221/20/07/P07007.

[11] Joschka Birk, Anna Hallin, and Gregor Kasieczka. OmniJet-α: the first cross-task foundation model for particle physics. *Mach. Learn. Sci. Tech.*, 5(3):035031, 2024. doi: 10.1088/2632-2153/ad66ad.

[12] James Giroux and Cristiano Fanelli. Towards Foundation Models for Experimental Readout Systems Combining Discrete and Continuous Data. 5 2025.

[13] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt, 2019. URL https://arxiv.org/abs/1912.02292.

[14] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (70):1–57, 2018.

[15] Cory Stephenson and Tyler Lee. When and how epochwise double descent happens, 2021. URL https://arxiv.org/abs/2108.12006.

[16] Marie Kempkes, Aroosa Ijaz, Elies Gil-Fuster, Carlos Bravo-Prieto, Jakob Spiegelberg, Evert van Nieuwenburg, and Vedran Dunjko. Double descent in quantum machine learning. 1 2025.

[17] Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, et al. Openfold: Retraining alphafold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature methods*, 21(8):1514–1524, 2024.

[18] ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3: S08003, 2008. doi: 10.1088/1748-0221/3/08/S08003.

[19] ATLAS Collaboration. ATLAS tt^ Simulation for ML-based Jet Flavour Tagging (JetSet). CERN Open Data Portal, 2025.

[20] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-k_t jet clustering algorithm. *JHEP*, 04:063, 2008. doi: 10.1088/1126-6708/2008/04/063.

[21] ATLAS Collaboration. Transforming jet flavour tagging at atlas, 2025. URL https://arxiv.org/abs/2505.19689.

[22] ATLAS Collaboration. ATLAS SUSY Searches in Wh1Lbb Channel Open Data Set. CERN Open Data Portal, 2024.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, art. arXiv:1412.6980, December 2014. doi: 10.48550/arXiv.1412.6980.

[24] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

[25] STEPHEN P. MARTIN. *A SUPERSYMMETRY PRIMER*, page 1–98. WORLD SCIENTIFIC, July 1998. doi: 10.1142/9789812839657_0001. URL http://dx.doi.org/10.1142/9789812839657_0001.