ISBN: 978-3-95450-248-6 ISSN: 2673-5490 doi: 10.18429/JACoW-IPAC2025-WEPM062

ADVANCES IN MACHINE LEARNING INFERENCE OF DYNAMIC APERTURE EVALUATION FOR THE LHC*

C.E. Montanari[1], D. Di Croce, M. Giovannozzi, S. Redaelli, F. Van der Veken
CERN, Geneva, Switzerland
R. B. Appleby, University of Manchester, Manchester, United Kingdom
T. Pieloni, EPFL, Lausanne, Switzerland
[1]also at University of Manchester, Manchester, United Kingdom

Abstract

Dynamic aperture (DA) is a crucial metric for understanding nonlinear beam dynamics and particle stability in circular accelerators such as the Large Hadron Collider (LHC) and its future upgrade, the High-Luminosity LHC (HL-LHC). Traditional methods for DA evaluation are computationally intensive and require extensive tracking of large particle ensembles over many turns. Recent advances in machine learning (ML) have demonstrated that models, particularly architectures such as Bidirectional Encoder Representations from Transformers (BERT), can significantly accelerate DA predictions while achieving accuracies comparable to those of traditional simulations. In addition, improved uncertainty quantification techniques have improved the reliability of these models, providing a foundation for more robust active learning frameworks. This work presents the latest progress in DA inference, with a focus on architectural advances, data set preparation, and optimised training techniques. Applied to LHC tracking data, these improvements underscore the importance of high-quality data generation and customised training strategies to increase model performance and uncertainty management, paving the way for future HL-LHC studies.

INTRODUCTION

The dynamic aperture (DA) [1] quantifies the extent of the stable region of the phase space. It affects critical parameters such as beam lifetime and luminosity in high-energy accelerators, such as the LHC [2] and the HL-LHC [3]. Accurate DA estimation is essential to optimise machine performance and guide design decisions for future projects such as the Future Circular Collider (FCC) [4]. However, its evaluation is computationally demanding, relying on extensive tracking simulations across multiple machine configurations.

Standard numerical approaches typically involve tracking particles over $10^5 - 10^6$ turns (see, e.g. [5, 6]), incorporating magnet imperfections and uncertainties through Monte Carlo methods. However, even these extensive simulations cover operational timescales significantly shorter than realistic physics runs. For instance, tracking particles for 10^6 turns corresponds to merely 89 s of the actual LHC operation time, whereas the physics fills typically extend for several hours. This discrepancy has been effectively addressed by proposing scaling laws for the time evolution of the DA [7, 8] that are based on the Nekhoroshev stability time theorem [9]. An alternative approach consists of looking for faster predictive models, especially since simulations can grow in complexity when additional factors are included, such as beam-beam interactions [10].

Surrogate models that leverage machine learning (ML) methods have emerged as efficient alternatives for parameter scans, significantly speeding up DA predictions with reliable performance [11–13]. However, beyond predictive accuracy, a critical aspect of ML-based DA modelling is the assessment and management of uncertainty in predictions [14]. Robust uncertainty quantification is essential to establish confidence in model outputs and optimise active learning strategies [15] to improve the quality of training data sets.

In this study, we detail recent developments in DA inference methodologies, focussing on improved model architectures, data preparation strategies, and advanced uncertainty quantification approaches. We further describe integrating these enhancements into Kubeflow pipelines [16], which facilitate greater scalability, adaptability, and automation in the ML-driven workflow for DA prediction.

METHODOLOGY

Data Generation and Preparation

Numerically evaluating DA is computationally demanding, particularly when considering the full four- or six-dimensional phase space. As a result, exhaustive sampling is often impractical. A common alternative is angular sampling [1], where the DA is determined by measuring the maximum stable radius (from now on called angular DA) along selected angular directions in the $x - y$ transverse plane.

The data set analysed in this work corresponds to a 2024 operational lattice at the LHC at its injection energy of 450 GeV. Simulated configurations were generated using MAD-X [17, 18], and particle tracking was carried out using XSuite [19]. Each simulated scenario incorporated essential machine parameters, such as betatron tunes, chromaticities, strengths of Landau octupoles, and a set of 60 standard magnetic error distributions commonly employed in previous studies. These parameters form the input set for the ML model. For each lattice configuration, the angular DA was computed along 44 polar angles. Additionally, the polar angle and the number of tracked turns were included as supplementary input, while the computed angular DA rep-

* Work supported by the HL-LHC project.

ISBN: 978-3-95450-248-6 ISSN: 2673-5490 doi: 10.18429/JACoW-IPAC2025-WEPM062

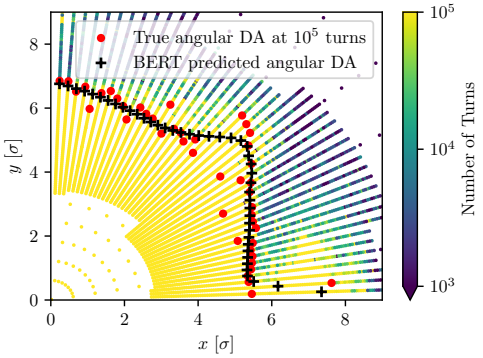

Figure 1: Example of angular DA at 10^5 turns in the $x-y$ transverse plane for an LHC lattice at injection. The first unstable particle along a radius marks the angular DA for the given direction. The BERT prediction is also shown.

resents the prediction target. An illustrative example that compares numerical angular DA results and predictions from our BERT-based model is provided in Fig. 1.

Initially, a baseline data set containing 5000 accelerator configurations was created through a parameter grid search. Subsequently, this data set was expanded to more than 10,000 configurations using active learning methods [15] that specifically targeted scenarios that exhibit high predictive uncertainty. Data quality was further enhanced through preprocessing steps, such as input parameter standardisation and balancing the DA distributions. The resulting data set, which contained approximately 50 million samples, was partitioned into 80% training, 10% validation, and 10% testing subsets.

Model Architecture and Training

We adopted a transformer-based neural network inspired by BERT [20] to perform DA predictions, exploiting its bidirectional self-attention capability to effectively capture intricate parameter interactions in accelerator data sets. Although BERT was initially proposed for natural language applications, its structure has proven to be effective in regression tasks involving complex relations between parameters, making it suitable for DA prediction starting from lattice parameters. The implementation contains 12 transformer encoder layers, each employing multi-head self-attention with 8 attention heads, followed by a feed-forward layer consisting of 512 units. To improve generalisation, dropout regularisation (rate of 0.5) and layer normalisation were included. A global average pooling layer was used to aggregate the output sequence into a single regression prediction.

The performance of the BERT-based model was benchmarked against other established neural networks [21] and our initial Multi Layer Perceptron (MLP) implementation. All tested architectures were implemented using TensorFlow [22], with optimised ReLU activations and hyperparameters through random search performed using Keras Tuner [23]. A schematic diagram illustrating the model structure is provided in Fig. 2.

Uncertainty Quantification

Uncertainty in ML-based inference arises from various sources [14], with epistemic uncertainty being particularly relevant for our case, as it reflects model confidence given limited training data. Reliable uncertainty estimation is essential for both guiding active learning strategies and assessing the reliability of DA predictions.

Our initial active learning implementation considered uncertainty estimates from a baseline Monte Carlo (MC) dropout [24, 25] configuration, where a 0.1 dropout rate was applied after the first hidden layer during inference, and the predictions were averaged over 256 stochastic forward passes. In a recent study [26], we compared different techniques for estimating epistemic uncertainty on our BERT-based deep learning model. Specifically, we evaluated MC dropout with different architectures and dropout rates and bootstrap aggregation (bagging) [27]. Bagging consists in training multiple models on resampled subsets of the training data and combining their predictions to assess uncertainty. Additionally, we also introduced a mixed technique that integrates both methods, where an MC dropout is performed on all networks obtained from bagging. All of these approaches were tested on the validation and testing data set, and the logarithm of the estimated relative error was compared with the actual relative errors observed in the DA predictions.

Pipeline Integration on Kubeflow

All simulations, machine learning training, and inference tasks described here were performed using CERN's GPU computing resources, with data sets stored and managed on the EOS storage system [28], suitable for handling large-scale accelerator data. Although the current infrastructure effectively supports these operations, further improvements in flexibility and automation are essential, especially for retraining and updating models for various LHC, HL-LHC, and FCC lattice configurations in future optimisation studies.

To address this, we are exploring the integration of our workflow into Kubeflow, an open source Kubernetes-based platform designed to automate and scale complex ML pipelines. Kubeflow allows a streamlined deployment of the various stages of active learning within a single reproducible framework. Given Kubeflow's availability at CERN [16], our objective is to establish an integrated end-to-end pipeline capable of efficiently updating and retraining models as new accelerator configurations are studied and optimised.

RESULTS AND DISCUSSION

The BERT-based neural network demonstrated good predictive accuracy for dynamic aperture estimation, outperforming the baseline MLP architecture, while maintaining acceptable computational efficiency. Quantitative and visual evaluations of the model predictions are provided in Table 1 and Fig. 3, respectively.

For uncertainty quantification, MC dropout with optimised architecture and a dropout rate of 0.005 per layer (see Fig. 2 for dropout layer positions) proved most effective,

ISBN: 978-3-95450-248-6 ISSN: 2673-5490 doi: 10.18429/JACoW-IPAC2025-WEPM062

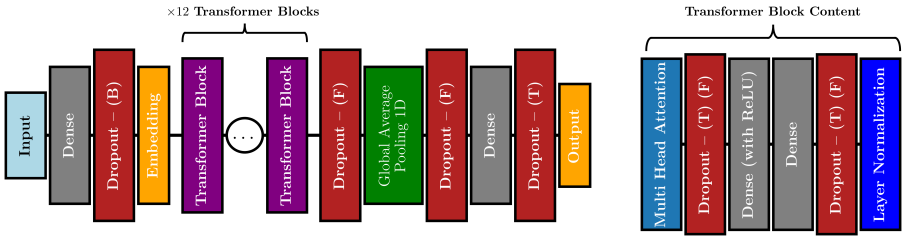

Figure 2: BERT-based ML architecture used for angular DA prediction. The dropout layers are used either in training (T), in the baseline architecture for uncertainty estimation (B), and/or in the final architecture for uncertainty estimation (F).

offering both accuracy and computational efficiency. The more expensive MC dropout-bagging method performed similarly. Results, compared to the previous baseline, are summarised in Table 2 and Fig. 4, using Pearson correlation, RMSE, and MAPE to assess the agreement between predicted and true uncertainties (lower values indicate better performance).

Table 1: Model performance comparison for angular DA prediction on test data set and inference time for sampling a full lattice configuration on an NVIDIA V100

Model	MAPE [%]	RMSE [σ]	Time [ms]
MLP (baseline)	18.57	0.621	3
BERT	14.37	0.536	51

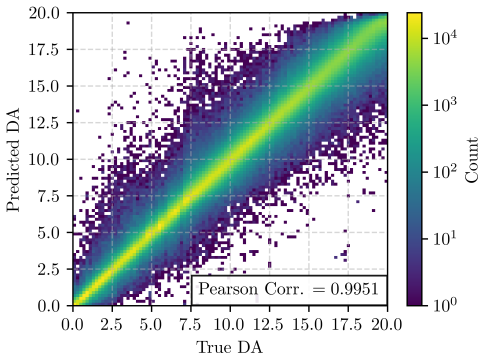

Figure 3: Test data set angular DA predictions from BERT compared with the true values.

Table 2: Uncertainty Estimation Performance on Test Data Set of Different Techniques

Technique	Pearson correlation	RMSE
MC dropout (baseline)	0.381	0.92
Bagging	0.518	0.581
Mixed technique	0.560	0.523
MC dropout (final)	0.579	0.525

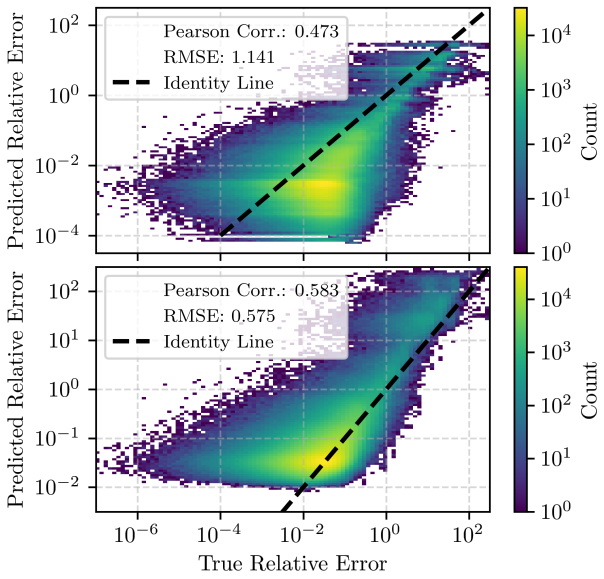

Figure 4: Uncertainty estimates of test data set compared with actual relative error for the baseline MC dropout (top) and the improved final MC dropout architecture (bottom).

CONCLUSION AND OUTLOOK

We presented recent advances in ML-based inference for dynamic aperture predictions at the LHC, highlighting improvements in model architecture, data set preparation, and uncertainty estimation. The BERT-based neural network achieved high accuracy without strongly impacting the overall computational cost, which is still extremely low compared to traditional tracking simulations. Uncertainty quantification through optimised MC dropout and combined methods improved the overall estimation, supporting effective active learning. Finally, we presented our ongoing efforts to integrate the workflow into Kubeflow pipelines to enhance automation, scalability, and flexibility, laying the essential groundwork for forthcoming LHC, HL-LHC, and FCC optimisation studies.

ACKNOWLEDGMENTS

This work was carried out under the auspices and with the support of the Swiss Accelerator Research and Technology programme (CHART) and the Swiss Data Science Centre project grant C20-10.

ISBN: 978-3-95450-248-6 ISSN: 2673-5490 doi: 10.18429/JACoW-IPAC2025-WEPM062

REFERENCES

[1] E. Todesco and M. Giovannozzi, "Dynamic aperture estimates and phase-space distortions in nonlinear betatron motion", *Phys. Rev. E*, vol. 53, no. 4, pp. 4067–4076, 1996. doi:10.1103/PhysRevE.53.4067

[2] O. S. Brüning *et al.*, *LHC Design Report*. CERN, 2004. doi:10.5170/CERN-2004-003-V-1

[3] G. Apollinari *et al.*, *High-Luminosity Large Hadron Collider (HL–LHC)*. CERN, 2017, vol. 4. doi:10.23731/CYRM-2017-004

[4] A. Abada *et al.*, "FCC–hh: The Hadron Collider: Future Circular Collider Conceptual Design Report Volume 3. Future Circular Collider", *Eur. Phys. J. Spec. Top.*, vol. 228, no. 4, pp. 755–1107, 2019. doi:10.1140/epjst/e2019-900087-0

[5] J. Barranco Garcia and T. Pieloni, "Global compensation of long-range beam-beam effects with octupole magnets: dynamic aperture simulations for the HL-LHC case and possible usage in LHC and FCC", CERN, Tech. Rep. CERN-ACC-NOTE-2017-0036, 2017, https://cds.cern.ch/record/2263347.

[6] K. Skoufaris *et al.*, "Numerical optimization of dc wire parameters for mitigation of the long range beam-beam interactions in High Luminosity Large Hadron Collider", *Phys. Rev. Accel. Beams*, vol. 24, no. 7, p. 074 001, 2021. doi:10.1103/PhysRevAccelBeams.24.074001

[7] M. Giovannozzi, W. Scandale, and E. Todesco, "Dynamic aperture extrapolation in presence of tune modulation", *Phys. Rev.*, vol. E57, no. 3, p. 3432, 1998. doi:10.1103/PhysRevE.57.3432

[8] A. Bazzani, M. Giovannozzi, E. H. Maclean, C. E. Montanari, F. F. Van der Veken, and W. Van Goethem, "Advances on the modeling of the time evolution of dynamic aperture of hadron circular accelerators", *Phys. Rev. Accel. Beams*, vol. 22, no. 10, p. 104 003, 2019. doi:10.1103/PhysRevAccelBeams.22.104003

[9] A. Bazzani, S. Marmi, and G. Turchetti, "Nekhoroshev estimate for isochronous non resonant symplectic maps", *Cel. Mech.*, vol. 47, no. 4, p. 333, 1990. doi:10.1007/BF00051010

[10] C. Droin *et al.*, "Status of beam-beam studies for the high-luminosity LHC", in *Proc. 15th Int. Particle Accelerator Conf. (IPAC'24)*, pp. 3213–3216, 2024. doi:10.18429/JACoW-IPAC2024-THPC77

[11] M. Schenk *et al.*, "Modeling Particle Stability Plots for Accelerator Optimization Using Adaptive Sampling", in *Proc. IPAC'21*, Campinas, Brazil, May 2021, pp. 1923–1926, 2021. doi:10.18429/JACoW-IPAC2021-TUPAB216

[12] M. Casanova, B. Dalena, L. Bonaventura, and M. Giovannozzi, "Ensemble reservoir computing for dynamical systems: prediction of phase-space stable region for hadron storage rings", *Eur. Phys. J. Plus*, vol. 138, no. 6, p. 559, 2023. doi:10.1140/epjp/s13360-023-04167-y

[13] D. Di Croce, M. Giovannozzi, T. Pieloni, M. Seidel, and F. F. Van der Veken, "Accelerating dynamic aperture evaluation using deep neural networks", in *Proc. IPAC'23*, Venice, Italy, May 2023, pp. 2870–2873, 2023. doi:10.18429/jacow-ipac2023-wepa097

[14] J. Gawlikowski *et al.*, "A survey of uncertainty in deep neural networks", *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 1513–1589, 2023. doi:10.1007/s10462-023-10562-9

[15] D. Di Croce *et al.*, "Optimizing dynamic aperture studies with active learning", *J. Instrum.*, vol. 19, no. 04, p. P04004, 2024. doi:10.1088/1748-0221/19/04/P04004

[16] D. Golubovic and R. Rocha, "Training and Serving ML workloads with Kubeflow at CERN", in *EPJ Web of Conferences*, p. 02 067, 2021. doi:10.1051/epjconf/202125102067

[17] R. D. Maria *et al.*, "Status of MAD-X V5.09", in *Proc. IPAC'23*, Venice, Italy, pp. 3340–3343, 2023. doi:10.18429/JACoW-IPAC2023-WEPL101

[18] *MAD - Methodical Accelerator Design*, https://mad.web.cern.ch/mad/.

[19] G. Iadarola *et al.*, "Xsuite: An integrated beam physics simulation framework", in *Proc. IPAC'24*, Nashville, TN, pp. 2623–2626, 2024. doi:10.18429/JACoW-IPAC2024-WEPR56

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019. https://arxiv.org/abs/1810.04805

[21] D. Di Croce, M. Giovannozzi, C. E. Montanari, T. Pieloni, S. Redaelli, and F. F. Van der Veken, "Assessing the performance of deep learning predictions for dynamic aperture of a hadron circular particle accelerator", *Instruments*, vol. 8, no. 4, 2024. doi:10.3390/instruments8040050

[22] Martín Abadi *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. https://www.tensorflow.org/

[23] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, *et al.*, *Kerastuner*, https://github.com/keras-team/keras-tuner, 2019.

[24] Y. Gal and Z. Ghahramani, *Bayesian convolutional neural networks with bernoulli approximate variational inference*, 2016. https://arxiv.org/abs/1506.02158

[25] Y. Gal, J. Hron, and A. Kendall, "Concrete dropout", in *Advances in Neural Information Processing Systems*, 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/84ddfb34126fc3a48ee38d7044e87276-Paper.pdf

[26] C. E. Montanari *et al.*, *Machine Learning Techniques for Uncertainty Estimation in Dynamic Aperture Prediction*, (under peer-review), 2025.

[27] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Springer New York, NY, 2006, vol. IV, pp. XX, 778.

[28] A. Peters, E. Sindrilaru, and G. Adde, "EOS as the present and future solution for data storage at CERN", *J. Phys. Conf. Ser.*, vol. 664, no. 4, p. 042 042, 2015. doi:10.1088/1742-6596/664/4/042042