ISBN: 978-3-95450-231-8 ISSN: 2673-5490 doi: 10.18429/JACoW-IPAC2023-THPL004

REAL-TIME BAYESIAN OPTIMIZATION WITH DEEP KERNEL LEARNING AND NN-PRIOR MEAN FOR ACCELERATOR OPERATIONS*

J. Martinez-Marin and B. Mustapha, Argonne National Laboratory

Abstract

The use of artificial intelligence (AI) has the potential to significantly reduce the time required to tune particle accelerators, such as the Argonne Tandem Linear Accelerator System (ATLAS). Bayesian optimization with Gaussian processes is a suitable AI technique for this purpose, it allows the system to learn from past observations to make predictions without explicitly learning representations of the data. In this paper, we present a Bayesian optimization method with deep kernel learning that combines the representational power of neural networks with the reliable uncertainty estimates of Gaussian processes. The kernel is first trained with physics simulations, then the model is deployed online in a real machine, in this case a subsection of the ATLAS linac, to perform the optimization. In addition to the kernel, we also modelled the mean of the Gaussian process using a neural network trained with simulation data and later with experimental data. The results show that the model not only converges quickly to an optimal tune, but it also requires very little initial data to do so. These approaches have the potential of significantly improving the efficiency of particle accelerator tuning, and could have important applications in a wide range of settings.

INTRODUCTION

The Argonne Tandem Linear Accelerator System (ATLAS) [1] is a Department of Energy (DOE)/Nuclear Physics (NP) User Facility for studying low-energy nuclear physics with heavy ions. It operates ~6000 hours per year. The facility (see Fig. 1), uses three ion sources and serves six target areas at beam energies from ~1-15 MeV/u. To accommodate the total number of approved experiments and their wide range of beam-related requirements, ATLAS reconfigures once or twice per week over 40 weeks of operation per year. The start-up time varies from ~12 – to 48 hours depending on the complexity of the tuning.

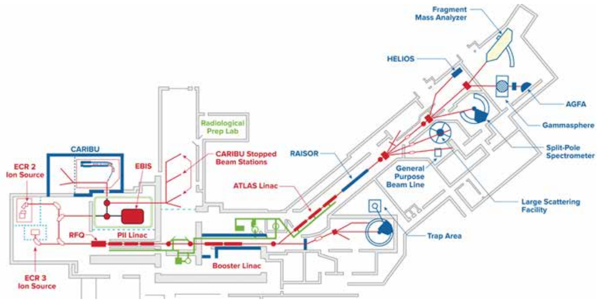

Figure 1: ATLAS Layout.

The procedure of tuning such an accelerator system is time-consuming and relies heavily on the intuition and experience of the operators. The uncertainties involved in tuning are in part due to unknown misalignments of the beamline components and the limited number of diagnostic devices to properly characterize the beam. The use of artificial intelligence (AI) has the potential of filling the information gap and significantly reducing the time needed to tune the accelerator. By reducing the time for beam tuning, more beam time will be available to help relieve the overbooked experimental nuclear physics program at ATLAS. In addition to beam tuning, AI models can be used to improve beam quality with the installation of new diagnostics and real-time data acquisition. These improvements will increase the facility's scientific throughput and the quality of the data collected.

This work is part of a DOE/NP project [2] that aims to apply artificial intelligence techniques to support the operations of ATLAS. The main goal of the project is to use AI to streamline beam tuning and improve the performance of the machine. The ultimate goal is to develop an AI model that can tune the machine in real-time while also providing insights that can help improve its overall performance. The experiments described in this paper represent an important step forward in achieving this goal.

The ultimate goal of this study is to optimize the transmission and quality of the beam in a subsection of ATLAS as quickly as possible. The authors achieve this by using Bayesian optimization with deep kernel learning to find the optimum settings of the linac, and also by using neural networks to model the mean of the Gaussian process. These approaches allow for a more efficient optimization of the accelerator's performance when the models are trained offline with simulation data and then transferred to the real machine.

TUNING MODEL

Bayesian optimization with Gaussian processes is a well-suited technique for tuning particle accelerators [3] because it has several appealing properties. It is a powerful and efficient method for optimizing black-box functions that are expensive or time-consuming to evaluate. The key advantage is its capability to balance exploration and exploitation in a way that is guided by the model's uncertainty, which allows it to efficiently find the global optimum of the objective function. This method allows the system to learn from past observations and to make predictions without explicitly learning representations of the data. This property makes this technique well-suited for problems where the data may be noisy or limited, as is often the case when working with particle accelerators.

The Gaussian process is defined by a mean function and a covariance function (also called a kernel), which together

* This work was supported by the U.S. Department of Energy, under Contract No. DE-AC02-06CH11357. This research used the ATLAS facility, which is a DOE Office of Nuclear Physics User Facility.

determine the shape and behavior of the model. In Bayesian optimization, the mean function is typically set to zero, and the kernel is chosen based on the characteristics of the data. Because Bayesian optimization with Gaussian processes is based on statistical models of the system, the model may not accurately reflect the true behavior of the accelerator in all cases. This can lead to limitations in the performance of the model, particularly in cases where the system exhibits complex or non-linear behavior. Moreover, the large number of tunable parameters in a particle accelerator generally requires a large number of observations to find the optimum which is problematic for BO due to the computational complexity of scaling the GP.

There are several strategies that people have used to address the challenges of applying Bayesian optimization with Gaussian processes to tune particle accelerators. In the context of tuning particle accelerator is common to incorporate prior physics information into the GP to improve the accuracy and efficiency of the optimization. It allows the optimization process to be guided by the known physical relationships in the system, rather than relying solely on data-driven learning.

This physics knowledge is usually introduced by constraints or training offline classical kernels such as the radial basis function (RBF) kernel [4], one of the most common choices for Gaussian process models. However, the performance of the model may be limited by the expressiveness of the kernel used to define the surrogate model. Alternatively, modelling the Gaussian Process's prior mean function is another way to incorporate prior physics-based knowledge about the target function [5].

This work explores the concept of using a more generalized kernel based on neural network and on the other hand the modelling of the prior mean with neural network.

DEEP KERNEL MODEL

By using a neural network as the kernel in the GP model, it may be possible to capture complex relationships between the machine parameters and the performance of the accelerator, which could allow the model to find the optimal settings more efficiently.

In this work, the kernel is built by interpolating between two kernels following the structured kernel interpolation method. The approach chosen was a deep kernel learning with a SKI kernel that includes an RBF basis kernel. The deep kernel learning has been shown to be effective at capturing complex relationships in the data [6], which can be particularly useful in particle accelerators where the relationship between the input and output variables is nonlinear, complex and highly structured. By combining deep kernel learning with a structured interpolation kernel that includes an RBF basis kernel, it may be possible to capture these complex relationships more effectively than using a single kernel function alone. The SKI allows the model to flexibly adjust the combination of both kernels during training.

The deep-learning kernel used is composed of a fully connected network with the layered architecture 500-250-125-2 and acts as a neural network feature extractor in the kernel of the Gaussian process. Rectified linear activation function (ReLU) is used as the activation function and the features are scaled to be between 0 and 1 and remain in the grid bounds expected by SKI.

Because of the data needs for the neural network and the potential benefit of training a model offline that would work in the real machine, the data used was generated from TRACK simulations, particularly 4k samples for the training set and 1k samples for the validation set. The ATLAS subsection used in this work is the new material irradiation station, AMIS. The input parameters involved are the current settings of a triplet and a doublet, the objective is to maximize the transmission through the beamline using an upper confidence bound acquisition (ß=2). See Figure 2 for a schematic of the involved section.

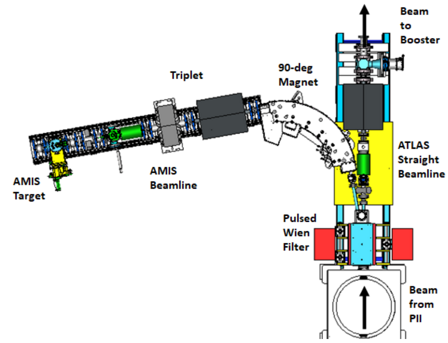

Figure 2: AMIS beamline.

Before experimental testing, a first test was done with the TRACK code comparing both a BO with RBF kernel over 100 different BO simulations starting from a bad transmission configuration. Figure 3 shows how BO-DKL tends to optimize better the beamline than BO with RBF, even though the problem might not be complex enough for leveraging the whole potential of BO-DKL.

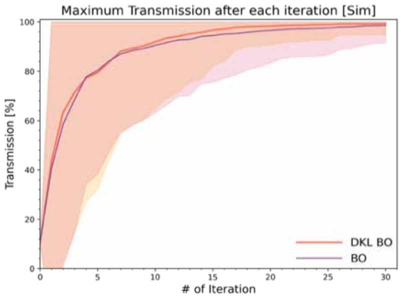

Figure 3: BO-DKL versus BO with RBF kernel mean and 2-sigma maximum transmission after each iteration.

The next step was to evaluate the performance of both methods in the real machine at the AMIS beamline. The results for the BO-GP and BO-DKL approaches are compared in Figure 4, where the blue line represents the results for the BO-GP approach and the green line represents the results for the BO-DKL approach after 100 iterations and using 3 configurations as warm start. The BO-DKL approach was able to achieve a maximum transmission of 56% in less than 50 iterations, surpassing the maximum transmission achieved by the operators on the same day (~53%).

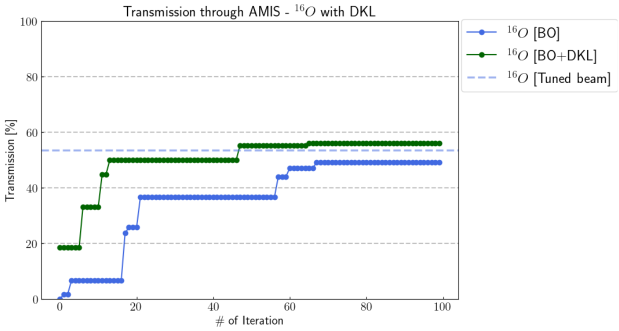

Figure 4: Comparison of the evolution of the maximum transmission achieved after each iteration for a ^{16}O beam using traditional BO-GP and BO-DKL.

The BO-DKL approach was also applied to a different beam, ^{22}Ne. For this beam, the data obtained for the ^{16}O beam was scaled and used to train the surrogate model. In this case, the BO-DKL approach was able to achieve a maximum transmission of 56% in less than 20 iterations, surpassing the maximum transmission achieved by the operators on the same day (~48%). The results for the BO-DKL approach on the ^{16}O and ^{22}Ne beams are shown in Fig. 5, where the green line corresponds to the BO-DKL approach on the ^{16}O beam and the orange line corresponds to the BO-DKL approach on the ^{22}Ne beam. The switching from ^{16}O beam to a ^{22}Ne beam demonstrate the transfer learning of the BO-based beam tuning model from one ion beam to another.

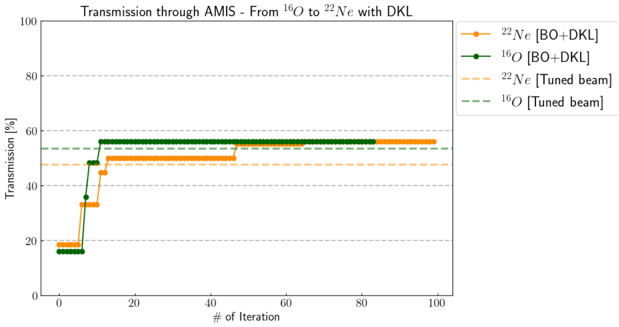

Figure 5: Comparison of the evolution of the maximum transmission achieved after each iteration using BO-DKL for a ^{16}O and the same model and scaled data for a ^{22}Ne beam.

PRIOR MEAN NN-MODEL

On the other hand, by using a neural network to model the prior mean in the GP model based on historical data is another way of incorporating prior physics-based knowledge about the target function, which tends to make the optimization more efficient and accurate. Notice that choosing a bad prior can have detrimental consequences for the whole inference endeavour because of the nature of the GP. The prior mean for the GP was modeled using a neural network of a couple of layers and 20 hidden nodes each using the TRACK simulation data mentioned in the DKL section. Figure 6 shows how the BO with NN-prior mean GP is capable of instantaneously optimizing the transmission by varying the 5 input settings during the 100 simulations while the BO with no prior knowledge takes more time. However, when transferring this model into the real machine, the results were not so clear, and we have not included them in this study. The different beams and the differences between the simulation lattice and the real lattice played a key role when transferring the prior mean model so far.

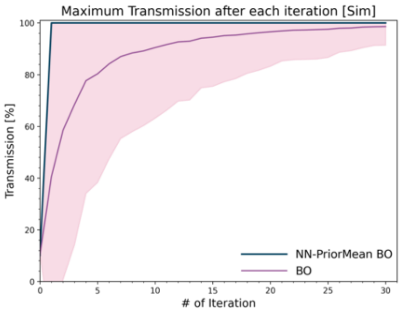

Figure 6: NN-Prior Mean BO versus BO with no prior knowledge and 2-sigma maximum transmission after each iteration.

Then, we trained the same model with experimental data based on a ^{14}N beam experiment and used it for a later ^{16}O beam experiment with promising results. Figure 7 shows how the BO with NN-prior mean based on previous experiment data was capable of optimizing better and faster than a BO with no prior knowledge, showing not only the power of using a model prior mean for the GP but again the potential of transfer learning techniques in tuning particle accelerators.

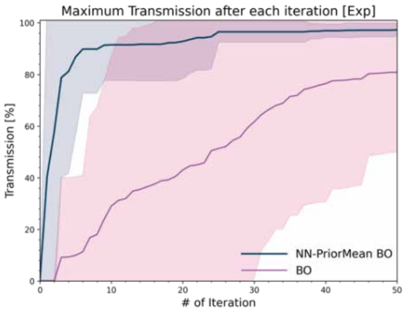

Figure 7: NN-Prior Mean BO versus BO with no prior knowledge and 2-sigma maximum transmission after each iteration.

CONCLUSIONS

In conclusion, this study highlights the effectiveness and accuracy of Bayesian optimization with deep kernel learning and NN-prior mean in optimizing the performance of real particle accelerators. These approaches enable transfer learning from simulation to machine and from one beam to another, even with limited data. However, further research is required to fully investigate the potential and limitations of these methods. Overall, the findings suggest that these approaches hold significant promise for optimizing particle accelerator performance in a more efficient and accurate manner.

REFERENCES

[1] P. N. Ostroumov *et al.*, “Completion of Efficiency and Intensity Upgrade of the ATLAS Facility”, in *Proc. LINAC’14*, Geneva, Switzerland, Aug-Sep. 2014, paper TUPP005, pp. 449-451.

[2] B. M. Mustapha, B. R. Blomberg, C. Dickerson, J. L. Martinez Marin, and C. E. Peters, “AI-ML Developments for the ATLAS Ion Linac Facility”, in *Proc. IPAC'21*, Campinas, Brazil, May 2021, pp. 4122-4125. doi:10.18429/JACoW-IPAC2021-THPAB181

[3] M. W. McIntire, T. M. Cope, D. F. Ratner, and S. Ermon, “Bayesian Optimization of FEL Performance at LCLS”, in *Proc. IPAC'16*, Busan, Korea, May 2016, pp. 2972-2975. doi:10.18429/JACoW-IPAC2016-WEPOW055

[4] A. Hanuka et al., “Physics model-informed Gaussian process for online optimization of particle accelerators”, *Physical Review Accelerators and Beams*, vol. 24, no. 7, 2021. doi:10.1103/physrevaccelbeams.24.072802

[5] C. Xu, R. Roussel, and A. Edelen, “Neural Network Prior Mean for Particle Accelerator Injector Tuning”. arXiv, 2022. doi:10.48550/ARXIV.2211.09028

[6] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, “Deep Kernel Learning.” arXiv, 2015. doi:10.48550/ARXIV.1511.02222