An active learning pipeline for surrogate models of gyrokinetic turbulence.

J. Barr[1], T. Madula[1], L. Zanisi[2], A. Ho[3,4], J. Citrin[3,4], A. Spurio Mancini[5], V. Gopakumar[2], S. Pamela[2] and JET contributors *

[1] *UCL's Centre for Doctoral Training in Data Intensive Sciences, University College London;* [2] *Culham Centre for Fusion Energy, Abingdon, ;* [3] *Dutch Institute for Fundamental Energy Research, Eindhoven, Netherlands;* [4] *Science and Technology of Nuclear Fusion Group, Eindhoven University of Technology, Eindhoven, Netherlands*
[5] *Mullard Space Science Laboratory, University College London, Holmbury St. Mary, Dorking, Surrey*

One of the bottlenecks in integrated models of tokamak plasmas is the plasma turbulence module. Even quasilinear gyrokinetic reduced order models such as QualiKiz [1] still result in long runtimes that are undesirable for many applications. In previous work, QLKNN [2], a simple feed-forward neural network (NN) surrogate model of Qualikiz, has been shown to provide a speedup of up to a factor of 10,000 in the turbulence module of some integrated models. However, the Qualikiz runs needed to simulate extensive databases used to train the NN involved a considerable computational burden. Moreover, specific regions of the input parameter space turned out to be invalid for numerical or physical reasons. Lastly, only a minority of points in the input parameter space resulted in unstable turbulent modes. Similar brute-force approaches are not feasible for more computationally-intensive models, for which a more optimised methodology would be required. Overall, it is hypothesized that the amount of training points needed to obtain a performant surrogate can be reduced by orders of magnitude. The proposed solution is to train a surrogate using Active Learning (AL, e.g, [4]), a method that allows to sample the input space of Qualikiz only at the points that will most likely result in an improvement of the surrogate model by reducing its output uncertainty or a related metric. We start from a subsample of the existing jetexp-15D database [3] consisting of only a few thousands data points ($\ll 1\%$). These are used to pre-train a NN to predict the Qualikiz fluxes, as well as two NNs tasked with identifying regions in the parameter space that are invalid or that do not result in turbulent fluxes (preliminary work indicates that a 80% accuracy is achievable for the two latter NNs). The three networks are uncertainty-aware, which enables AL, and they are retrained after each time a sampling of the input space is performed. With this technique, we expect the size of the original dataset to be significantly reduced while retaining the performance of the original QLKNN surrogate. This method may be used to scale up the number of dimensions required to obtain an improved surrogate compared to the state of the art, and it may be applied to other, potentially more computationally expensive gyrokinetics simulations where a vast simulated database is unavailable.

References

[1] Stephens, C. et al., 2021, doi:10.1017/S0022377821000763

[2] K. L. van de Plassche et al., 2019, https://doi.org/10.1063/1.5134126

[3] A. Ho et al., 2021, https://doi.org/10.1063/5.0038290

[4] P. Ren et al., 2021. https://doi.org/10.1145/3472291

*See the author list of 'Overview of JET results for optimising ITER operation' by J. Mailloux et al. published in Nuclear Fusion Special issue: Overview and Summary Papers from the 28th Fusion Energy Conference (Nice, France, 10-15 May 2021)