Validating physics-based (ASTRA/TRANSP), data-driven (D3D+AUG), and physics+data hybrid models for quantitatively accurate yet generalizable guidance for ITER operators

J. Abbate[1,2], E. Fable[3], G. Tardini[4], A. Pankin[1], B. Grierson[4], I. Char[5], E. Kolemen[1,2]

[1]Princeton Plasma Physics Laboratory, Princeton, NJ, USA; [2]Princeton University, Princeton, NJ, USA; [3]Max Planck Institut für Plasmaphysik, Garching, DEU; [4]General Atomics, San Diego, CA, USA; [5]Carnegie Mellon University, Pittsburgh, PA, USA
Email: jabbate@princeton.edu

ITER operators will leverage predictive models to plan upcoming operational phases, rundays, and discharges. These models are developed based on experience with previous devices, and will ideally be continuously updated as ITER data is collected. However, the reliability of even state-of-the-art simulations for predicting and extrapolating to a fusion-grade plasma is an open question; and it is not obvious how to rigorously update models with data.

Here we present the results from two recent publications [1,2]. A first-of-its kind large-database verification of state-of-the-art integrated modeling (ASTRA[3]/TRANSP[4] with TGLF[5]) demonstrates that quantitative predictions of **transport models perform no better than simple benchmarks** involving no physics and few to no fit parameters. In contrast, **a fully data-driven neural network approximation yields >20% better results** than these benchmarks. **Nonetheless, the fully data-driven approach lacks generalizability**: for predicting sufficiently far out of the distribution of training data, the neural network fares no better than the integrated physics modeling. What is more, **training on data from multiple machines (DIII-D+AUG) to attempt to increase generalizability yields no improvement**. Therefore, a variety of other techniques for combining the generalizability of physics models and the accuracy of data-driven models are tested, and **a "meta-learning" model trained for the explicit task of extrapolation demonstrates a >10% improvement in performance beyond physics or data alone**. This approach can be retrained in milliseconds and is interpretable, which could yield not only improvements in models used by ITER operators, but also a robust mechanism for continuously updating models as ITER data is collected.

Large-database verification and validation of ASTRA/TRANSP with TGLF: profile predictions no better than benchmark two-parameter regression: As defined by the ITER expert group [6], a popular metric for performance of integrated modeling is

$$Error_{ITER\ 1999}(T_{prediction}) = \sqrt{\sum_{j=1}^{N}(T_{prediction}(R_j) - T_{truth}(R_j))^2} \Big/ \sqrt{\sum_{j=1}^{N} T^2_{truth}(R_j)}$$

for $T_{prediction}(R_j)$ the temperature predicted by a model at spatial location R_j , and $T_{truth}(R_j)$ the experimental temperature measurement. Many studies have evaluated this error for core temperature predictions made by state-of-the-art transport models like TGLF on a few hand-picked shots [7]. By contrast, this work evaluates both ASTRA and TRANSP on more than 100 DIII-D discharges to ascertain statistical significance, and contextualizes these numbers with reasonable benchmarks. This work employs a "Profile Consistency" benchmark model in which profiles are assumed to exponentially grow between the edge (normalized flux coordinate $\rho = 0.8$) and the core ($\rho = 0.2$); and assumed to be flat at the very center ($\rho = 0 - 0.2$). Only one parameter for each of the two (electron and ion) temperatures must be fit to predict the entire core profile given the edge value: an estimate for $T(\rho = 0.2)$ given $T(\rho = 0.8)$. Figure 1 shows an example case for these predictions. For the standard validation exercise of core profile prediction, ASTRA and TRANSP have no statistically significant advantage over the Profile Consistency benchmark.

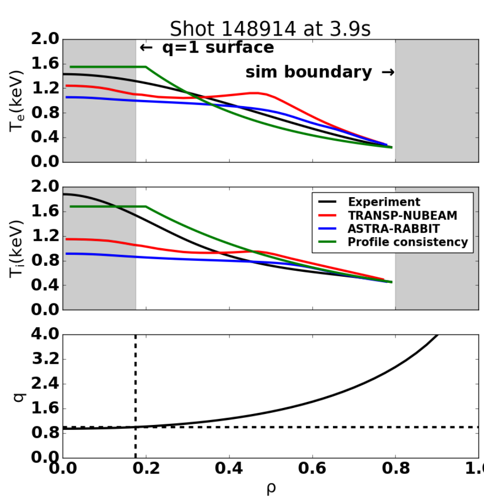

Figure 1: Timeslice 20ms after neutral beam power is rapidly increased from 0 to 4MW in DIII-D discharge 148914, wherein temperature begins to increase. Consistent with most core profile validation exercises, an experimentally measured value at the $\rho = 0.8$ flux surface is used for the boundary condition; and error is evaluated only outside the safety factor q=1 surface.

Large-database validation of ASTRA + data-driven models for time-dependent predictions: physics and extrapolated machine learning underperform; combining data+physics yields improvement: Like any experimental campaign, ITER will have a staged set of operations during which plasma parameters are increased. For example, the 2024 ITER research plan describes the anticipated steps in plasma current from 500kA to 15MA over 2.5 years of pre-fusion operation [8]. Neural networks exclusively trained on experimental data have been demonstrated to perform well within distribution [9], but for extrapolation to new regimes it is usually assumed physics-based models are required. This work gives some quantitative sense of this by separating DIII-D discharges into regimes of plasma current. A machine learning model trained for <0.9MA is tested for >1.3MA ("far extrapolation"); a model trained for <1.2MA is tested for >1.3MA ("near extrapolation"); and a model trained for all cases is tested for >1.3MA ("interpolation"). The task of the model, analogous to transport models, is full-profile (electron and ion temperature) time-dependent predictions: given initial profiles and a trajectory over time for actuators, predict the trajectory of profiles. The "Constant" benchmark employed in this case is to assume profiles remain fixed at their initial-time values. It is found that physics models and "far extrapolation" machine learning models both perform quantitatively worse (red) than this simple "Constant" benchmark. Meanwhile, "near extrapolation" machine learning performs neither statistically significantly worse nor better than the benchmark (yellow). A variety of mechanisms to improve the generalizability of models are therefore attempted: (1) supplementing DIII-D training data with AUG data (including at higher plasma current), (2) transfer learning from simulations to experimental data, (3) adding contextual simulation information (such as estimated deposited power profiles) as additional model inputs, and (4) "meta-learning" for the explicit task of extrapolation. Of the four methods, meta-learning is uniquely able to provide a statistically significantly better (green) result than the "Constant" benchmark. Based loosely on "stacked generalization" [10], in this method the training set is split into pieces such that extrapolation can be simulated: the training set of <1.2MA is split into <0.9MA and 1.0-1.2MA. A variety of models can be trained on <0.9MA data, then a "meta" model takes as inputs the predictions from each model and outputs a single prediction. This meta model is trained on 1.0-1.2MA, and in the simplest case (shown here) is represented by a weighted average of all models, i.e. weights $\alpha_{predictor}$ are learned such that $\widehat{T}_{meta}(\rho, t) = \sum_{predictors} \alpha_{predictor} T_{predictor}(\rho, t)$. In this particular case, 70% of the weight is given to the machine learning model, and the other 30% is distributed among physics predictors like a TGLF-nn-like [11] model and a GyroBohm-like [12] model. The machine learning model is then retrained on the entire <1.2MA training set, and the weighted sum is taken with this new model to predict >1.3MA.

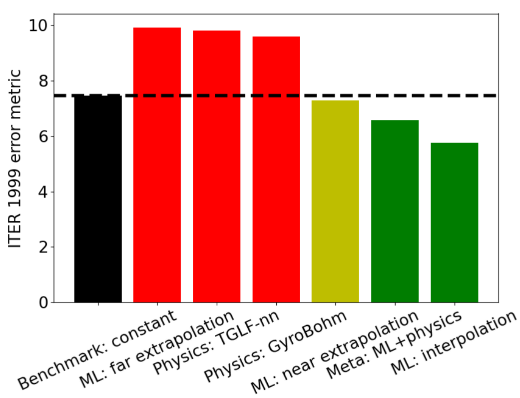

Figure 2: ITER 1999 "sigma" error metric for various machine learning, physics, and meta predictors. Physics-based predictors and machine learning extrapolating far out of distribution (from plasma current <0.9MA to >1.3MA) perform no better than the simple benchmark of holding profiles fixed at their initial values; extrapolating machine learning slightly (from plasma current <1.2MA to >1.3MA) does better; but combining a near-extrapolated machine learning prediction with physics predictions using a meta-learned model yields significantly better results, closer to performance of a machine learning model predicting within the distribution upon which it was trained (all plasma current).

ACKNOWLEDGEMENTS This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Fusion Energy Sciences, using the DIII-D National Fusion Facility, a DOE Office of Science user facility, under Award(s) DE-FC02-04ER54698; and PPPL contract DE-AC02-09CH11466.

REFERENCES [1] Abbate J. et al. (2024) Phys. of Plasmas 31, 042506 [2] Abbate J. et al. (in review) Nucl. Fusion [3] Pereverzev G. et al. (1991) ASTRA - an automatic system for transport analysis in a tokamak [4] Hawryluk R. (1981) An Empirical approach to Tokamak Transport [5] Kinsey J. et al. (2008) Phys. of Plasmas 15, 055908 [6] ITER Physics Expert Group (1999) Nucl. Fusion 39, 2175 [7] Kinsey J. et al. (2010) Phys. of Plasmas 17, 122315 [8] ITER research plan within the staged approach (2024) [9] Abbate J. et al. (2021) Nucl. Fusion 61, 046027 [10] Wolpert D. et al. (1992) Neural Networks 5, 08936080. [11] Meneghini O. et al. (2017) Nucl. Fusion 8, 086034 [12] Erba M. et al. (1998) Nucl. Fusion 38, 00295515.