1. Cover Letter

Federal Agency	Department of Energy
Grant #	DE-SC0021380
Project Title	EFIT-AI: Machine Learning and Artificial Intelligence Assisted Equilibrium Reconstruction for Tokamak Experiments and Burning Plasmas
PI	Scott E. Kruger, kruger@txcorp.com (303) 996-2039
Submitting Official	Larry Nelson, lnelson@txcorp.com (720) 974-1856
Submission Date	12/31/2024
DUNS#	806486692
Recipient Organization	Tech-X Corporation, 5621 Arapahoe Ave, Suite A, Boulder CO 80303
Project Period	09/01/2021 - 08/31/2024
Reporting Period End Date	08/31/2024
Report Term	Final Report
Signature	

2. Executive Summary

The EFIT-AI project is creating a modern advanced equilibrium reconstruction code suitable for tokamak experiments of burning plasmas. EFIT [1,2] was the first and is the most extensively used equilibrium reconstruction code in the world. This project builds on the production-level experience and adds key elements as follows. 1. A Model Order Reduction (MOR) version of the two-dimensional (2D) Grad-Shafranov equation solver (EFIT-MORNN) using physics-informed neural networks. 2. Improved optimization and data analysis capabilities using a Bayesian framework enhanced with machine learning. 3. A MOR version of the three-dimensional (3D) perturbed equilibrium reconstruction tool.

EFIT-MORNN has three goals: 1. As an initial condition for the core EFIT solver to enable fast full-GS solutions, 2. As a stand-alone tool for real-time control of tokamak plasmas, and 3. As part of a Bayesian framework for equilibrium reconstruction. For all three goals, accuracy and speed are of primary importance. For accuracy, we are interested in understanding the implications of training on magnetics-only, magnetic + Motional Stark Effect (MSE), and kinetic EFITs as well as their impacts. One technique is to use high-resolution EFITs for the training database (high accuracy), but then train the neural network to learn sub-sampled equilibria for speed. This gives a neural network that is more accurate than if we had trained with lower-resolution EFITs. EFIT-MORNN learns to predict the poloidal flux function while satisfying force balance with the toroidal current density. Both the flux prediction and constraint projection are learned using two neural networks trained simultaneously. We employ a neural architecture search to discover neural networks that are optimal in terms of both accuracy and computational efficiency, while leveraging the DeepHyper package to train these models at-scale on leadership-class HPC systems. The preliminary success of our neural networks is shown in Figure 1 where our inferred solution is more accurate that the real-time version of EFIT.

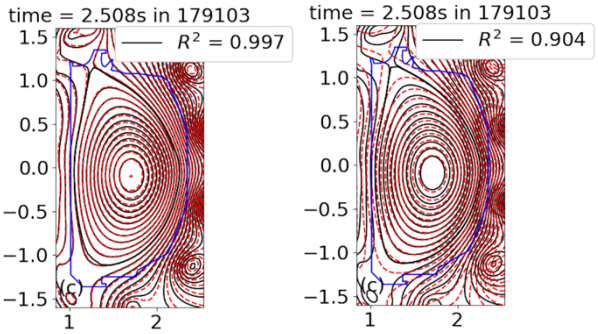


Figure 1. Comparison of EFIT-MORNN left versus real-time EFIT right for a Super-H mode shot (DIII-D discharge 179103) shows that the neural network is able to give more accurate flux values for approximately the same computational time (~2 msec).

The success of machine learning (ML) relies on large amounts of quality data. To enable this, an EFIT-AI database has been assembled and curated: a collection of multiple equilibrium reconstructions for a variety of tokamak discharges. The database features the entirety of the 2019 DIII-D campaign (approx. 2500 discharges) and contains three different types of EFIT reconstructions: 1. with only magnetics constraints; 2. with magnetics + MSE constraints; and 3. with bother user-generated and OMFIT-automated kinetic constraints (for a subset of shots). The database is currently being used for training of EFIT-MORNN. To incorporate Findable, Accessible, Interoperable, and Reusable (FAIR) data principles, the data was organized according to the ITER IMAS (Integrated Modeling and Analysis Suite) data schema (ontology), and then stored as self-descriptive HDF5 binary files that will be made publicly available.

The equilibrium reconstruction is an inverse problem that must deal with uncertainty and accuracy throughout the process. Bayesian methods offer a compelling approach for this problem. Here, we start by considering only the inference of Thomson temperature profiles using Gaussian

Process Regression (GPR). GPR is a Bayesian method for inferring profiles based on input data. Here, we present results investigating the use of a Student-T distribution of the likelihood distribution and show how it can accurately handle outliers in the data [4]. We then contrast these techniques with multiple traditional fitting methods. Bayesian methods are inherently a sampling-of-probability-distribution-functions method. For multiple diagnostics, the sampling over distribution functions can become expensive. We have demonstrated how machine learning can enable efficient sampling to make the Bayesian approach tractable.

Part of the uncertainty in traditional equilibrium reconstruction is in the model itself; the Grad-Shafranov equation assumes axisymmetry even though error fields, discrete coils, etc. are known to introduce perturbations to this. To address this, we have developed a reduced model of 3D perturbed equilibrium based on solutions from the 3D MHD code MARS-F. A database was created, and singular value decomposition was used to reduce the data in our training session. By training on this reduced data, we can efficiently and accurately capture three-dimensional effects [5].

To make all these techniques ready in the production environment of burning plasmas, we have made many improvements to the core EFIT solver. These improvements include clearly separating out the device-specific coding, improving code portability, and improving thread-safety in preparation for GPU-developments. To create the large database using automated tools while maintaining quality, we have improved quality-of-equilibrium checks and enabled their use. We have extensively improved the input and output to work with IMAS. Using extremely portable OpenMP-offload directives, we have been able to improve the performance of EFIT using GPU hardware. The most expensive computational kernel was made 65 times faster for an overall 50% speed-up time of the code.

We made significant progress towards preparing equilibrium reconstruction for the burning plasma era. This requires Bayesian analysis to maximize limited information and the use of neural nets for fast evaluation. The details however require considerable effort because of the computational cost and the balance between fast, accurate, and robust approaches. Finding the optimal balance remains an outstanding issue.

3. Accomplishments

Tech-X has contributed in every phase of the project throughout its inception. In this report, we detail only work performed in the last two years of the project by Tech-X. Earlier work is summarized in the prior reports. We break the work down by categories.

Publications and conference presentations. Tech-X contributed to publications 1-4 listed below within the past two years. Publication #1 is based on an invited talk at APS-DPP in 2023 and required significant effort to develop and publish. Publication #2 is an achievement because it represents the use of cutting-edge machine learning techniques, and our goal was to present it in a physics journal. Presenting technical machine learning work to a physics audience was challenging. Presentations 3 and 4 are exciting avenues that hopefully can be pursued in other work. Tech-X contributions to this work was less involved but it's exciting.

Two other notable presentations are the first two listed under Conference Presentations below. The first is a summary of the project presented at the 29th IAEA Fusion Energy Conference (FEC 2023) in London. Considerable competition exists internationally for this work and presenting to an international audience was valuable. The second is an invited talk by Cihan Akcay that provides a summary of many of the machine learning issues that we have learned throughout the course of this work.

Integration of GPR work with OMFIT. Prior years' work developed GPR for use in Thomson scattering data from an empirical point of view where different GPR techniques were assessed. Here, we upgraded our prior tools to make it more robust and integrated it with the OMFIT tool base so that any user can perform GPR analysis as part of profile fitting. Work was ongoing to do a more systematic empirical study with Columbia when funding ended.

Work on GPR for magnetic signals. The most fundamental EFIT mode is the magnetic only EFIT. At the time of the writing of the original proposal, we expected it to be relatively straightforward based on prior work in this area. However, as we dug into the prior work in more detail, we realized that all of them were significantly limited. In this period of performance, we made progress on a more general approach based on the success of the neural network model. This is extensively detailed in the next section.

4. Optimization for Real Time Inference

A. Profiling and Optimizing the Python inference routine

The EFIT-Prime model was developed and tested using Tensorflow and associated ML libraries. These libraries are optimized for efficient batched evaluation of a large datasets (high throughput), however real time applications require fast (low latency) inference on individual time slices. As first step towards applying EFIT-Prime for real time applications we performed profiling and optimization of the full EFIT-Prime workflow. This includes the pre-processing of the data and evaluation of the ML models. The works in collaboration with our partners at ANL and GA, who did additional profiling and optimization of the ML models as part of the FES 2024 Theory and Simulation Performance target.

Before focusing on profiling, it's helpful to briefly review the EFIT-Prime model. The first step in the model processes the magnetic diagnostics. During the pre-processing step the magnetic diagnostics are first normalized and then projected onto a RZ grid. A smoothing filer is then applied that averages over overlapping diagnostics. Then the principal component transom is applied to the filtered data, and the first 30 principal components are retained.

After the pre-processing step, the principal components are used as inputs to the EFIT-Prime NAS model. The NAS Model is an ensemble composed of five optimized neural networks. The networks are independent and compute both ψ and j_{tor}. The NAS model then uses the combined output to compute the mean and quantify the uncertainty for both ψ and j_{tor}.

The subsequent timings in this subsection are from a 16-core AMD Ryzen 9 5950X CPU with 32GB RAM performing inference on the full model without any pruning or quantization applied. First, we consider the cost of of preparing the magnetic signals and projecting them on to the principal components. The batched computation projected the signals onto the RZ grid and applied the smoothing filter in real-time. These steps leveraged vectorization to get good performance. However, these two steps can be combined with the projection onto the principal components into an single Affine transformation. This transformation is computed using a matrix vector product followed by vector addition. The advantage is that the transformation matrix and offset vector only must be computed once. The cost of computing the matrix and offset vector takes on average 107ms. The subsequent cost of of transforming the signals on takes 38μs per time slice. This highlights the importance of pre-computing the transformation for real-time. The cost of transforming the signals is essentially negligible.

We now consider the cost of the ensemble model inference. Again, this process is broken into two steps. The first is to read and setup on the five NAS models, and the second is running the inference on the models. The time it takes to read and setup the models takes on average about 1.05s, but this initialization is only needed once. The cost of running the inference of a single time step is a little more nuanced. On average it takes about 107ms to run the inference on all five models. The first time the five models are analyzed there is an additional initialization cost of 912ms.

These timings are computed by evaluating each of the five models in the NAS ensemble in serial. The evaluation of the models is independent and this step is trivially parallelizable. The average time to evaluate a time slice for each model is 21.4ms and ranges from 20ms to 23.5ms. Assuming efficient parallelization on similar hardware the cost to run the inference on a single time slice, which will be limited by the slowest model evaluation, is around 23.5ms.

Ideally, real time inference requires around 10ms inference. These initial performance studies indicate real time is plausible, but roughly a factor of two performance improvement is needed.

We expect some benefit from using more modern hardware. Additional paths to improved performance include exploring the const of inferring on small grids (here 129x129 grids were used), optimizing the NN inference using standard techniques (pruning, quantization, distillation, and decomposition), and exploring NN libraries optimized for low latency (real-time) applications.

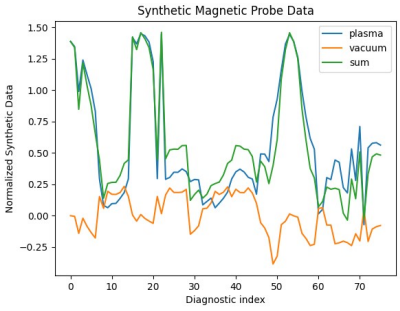
B. Development of Bootstrapping

In this section we develop the tools needed to address two questions:

- How does EFIT-Prime compare to traditional inference-based reconstructions
- How does EFIT-Prime behave if a diagnostic produces a faulty signal

The EFIT-Prime model is trained using roughly 145000 EFIT "magnetics only" equilibria generated using roughly 800 discharges from the 2019 DIIID experimental campaign. The EFIT-Prime neural networks are trained by comparing the inferred ψ and J_tor on a 2D mesh with the EFIT inferred values on the same mesh. A Prime model is considered a "good fit" based on how well it agrees with the EFIT prediction. However, from a experiment standpoint we really want to know how well the EFIT-Prime prediction agrees with the measure data. If The Prime prediction agrees with the EFIT prediction but both show large data mismatch, then that is ultimately a "bad-fit". To make this comparison we need to compute synthetic diagnostics that model the real diagnostics on the experiment. These diagnostics allow for a direct calculation of the data-mismatch (the difference between a measured diagnostic and a modeled diagnostic) and subsequent measures of the quality of fit (e.g. χ^2). The ability of calculated the synthetic diagnostics enables a direct apples-apples comparison between EFIT-Prime, experimental data, and reconstructions computed using alternative methods.

The ability to compute synthetic diagnostics also enables the possibility to bootstrap the EFIT-Prime model. The model is trained using a collection of 145 diagnostic measurements (76 magnetic probes, 44 flux loops, the measured plasma current, 6 Ohmic coil currents, and 18 poloidal field coil currents). The model is trained assuming that all diagnostics are available and provide reliable measurements. However, in experiments diagnostics often drop out or provided erroneous data. Thus it's important to characterize how sensitive the model is to faulty data, and explore methods to correct for that data. One possible method to correct for faulty data is to bootstrap the Prime model starting with a few bad data points. The idea is to first call the model once with the faulty data, and then compute synthetic data from the inferred data. Then call the model a second data replacing faulty diagnostics with the synthetic data.



C. Characterizing the magnetic PCA

EFIT-Prime uses 145 magnetic diagnostics to infer ψ and J_{tor}. These diagnostics are pre-processed before inputted into the ensemble neural network model. The pre-processing first normalizes the data. Then the data is projected onto a 2D grid and smoothing is used to averaging overlapping signals. Finally, principal component analysis is applied to the smoothed data, and the first 30 components are retained for subsequent analysis. In this section we explore the result of this analysis, and how sensitive the principal components are to in the diagnostics. For analysis we use DIII-D discharge 180087 and consider a time slice half-way through the discharge (3.58s).

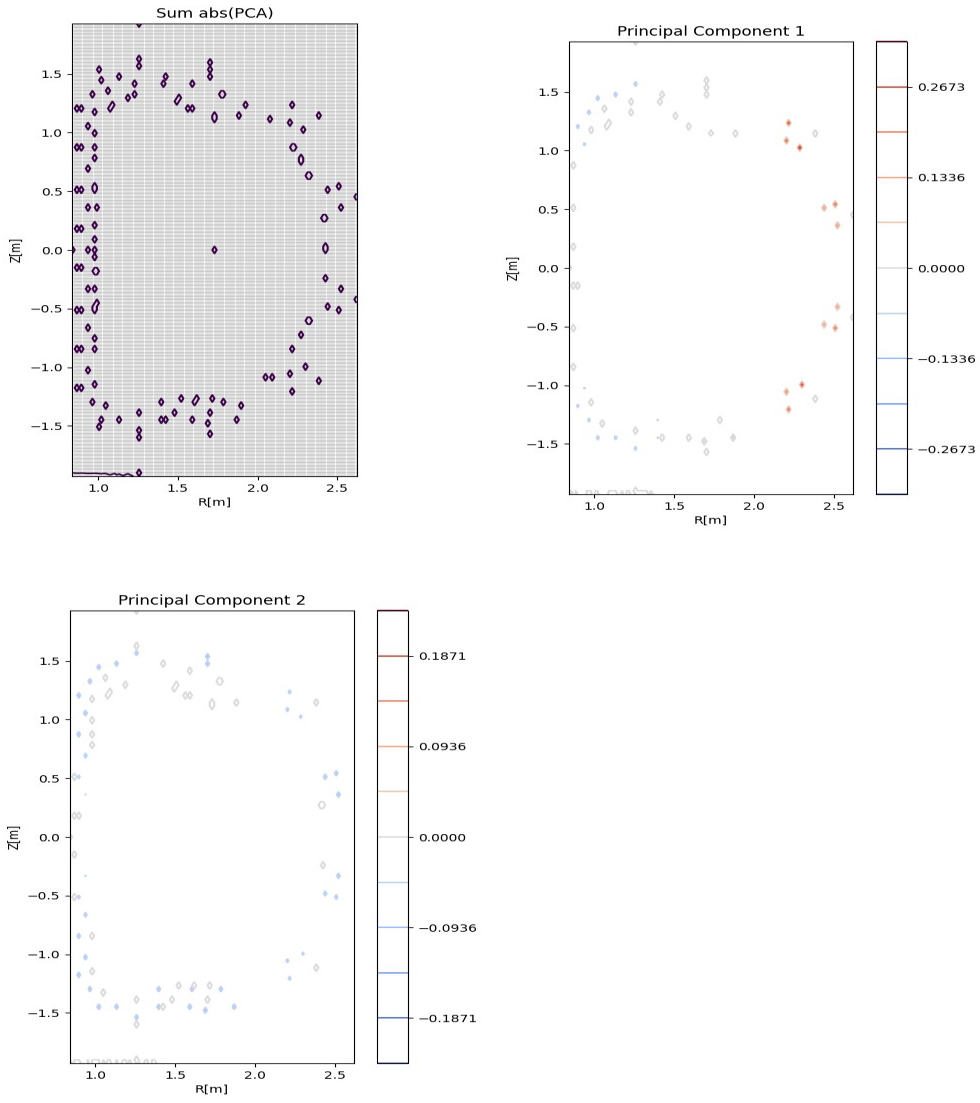


Figure 1a shows the span of the 30 components on the standard EFIT-Prime "pca-mesh." This is a 129x129 grid that is enlarged relative to the EFIT grid. The regions within the dots represent the RZ location of diagnostics (the total plasmas current is assumed to be located at the mesh center). The Regions that span multiple grid locations indicate places where diagnostics overlap, and their measured singles are averaged but the smoothing. Diagnostics that are located at the same R-Z

location but different toroidal planes are also averaged. One of the motivations for using a the 2D mesh was to enable the PCA to average adjacent diagnostics. The averaging should in principle make the model more robust to faulty data. However, due the high-resolution mesh used here, there is little overlap and little averaging is performed. Future work should focus on reducing the resolution of the pca-mesh to test if the averaging impacts the sensitivity to faulty data.

The 2D representation of the measurements allows for a physical analysis of the different pca components. Figures 1b and 1c show the 1st and 2nd components. The first components shows a polarity in the radial direction. Measurements on the outboard side are positive, and those on the inboard side are negative. This suggests that this component is capturing the radial shift of the plasma. The 2nd components shows a symmetric response. The associated diagnostics all have the same sign. It's interpretation is less clear, but it could capture changes in the plasma inductance.

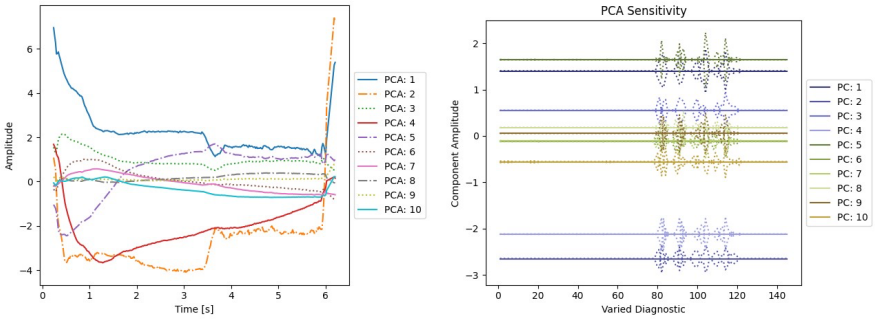


Figure 4 shows the evolution of the first 10 principal components over the course of discharge 180087. This figure illustrates several features. First the components are order by amplitude. The amplitude of the first for components is larger than the amplitude of the later components. Second, large changes are visible in the first few components. These changes represent different phases of the discharge (start-up, l-mode, h-mode, shut-down). In contrast the higher order components show less variation and evolve smoothly throughout the discharge.

Figure 5 shows the sensitivities to the principal components amplitudes to changes in an individual diagnostic. This is computed for by interdependently varying each signal by 50%. Each signal is both increased and decreased by 50%. In Figure 5 the diagnostics are ordered as follows. Signals 1-76 are magnetic probes, signals 77-120 are flux loops, 121 is the total plasma current, 122-127 are the ohmic coil currents, and 128-145 are the poloidal field coil currents. The figure indicates that the principal components are most sensitivities to a changes in the poloidal field coils. The 50% changes in the individual flux loop signals result similar changes in the principal component amplitudes. However, the component amplitudes are much less sensitive to changes in the other diagnostics. The 50% changes in the magnetic probe signals on results in a change of a few percent or less in the component amplitudes. Based on this analysis, we suspect that the EFIT-Prime model should be robust to changes individual faulty magnetic probe and coil current diagnostics. However, faulty flux loop measurements provide a great concern.

D. Future directions

We ran out of funding before this work was completed, so there are obvious next steps. First is to compare the synthetic data generated from the EFIT-Prime prediction with experimental data. This can be used to compare EFIT-Prime with other reconstruction methods (both traditional and ML based). Second, we want to explore the sensitivity of the full EFIT-Prime model to diagnostic errors. The preceding analysis only consider the sensitivity of the input principal components.

Finally, we want to test the viability of bootstrapping the EFITPrime model. Note that for magnetics only reconstruction one could generate reasonable surrogate data using simple models of the plasma current and coil currents. However, our goal is real time kinetic reconstitution, where such simple surrogates are not as easily to computed.

5. Publications

A. Journal Articles since Beginning of Project

1. Kruger, S.E., Leddy, J., Howell, E.C., Madireddy, S., Akcay, C., Bechtel Amara, T., McClenaghan, J., Lao, L.L., Orozco, D., Smith, S.P. and Sun, X., 2024. Thinking Bayesian for plasma physicists. *Physics of Plasmas*, *31*(5).
2. Madireddy, S., Akçay, C., Kruger, S.E., Amara, T.B., Sun, X., McClenaghan, J., Koo, J., Samaddar, A., Liu, Y., Balaprakash, P. and Lao, L.L., 2024. EFIT-Prime: Probabilistic and physics-constrained reduced-order neural network model for equilibrium reconstruction in DIII-D. *Physics of Plasmas*, *31*(9).
3. McClenaghan, J., Akçay, C., Amara, T.B., Sun, X., Madireddy, S., Lao, L.L., Kruger, S.E. and Meneghini, O.M., 2024. Augmenting machine learning of Grad–Shafranov equilibrium reconstruction with Green's functions. *Physics of Plasmas*, *31*(8).
4. Sun, X., Akcay, C., Amara, T.B., Kruger, S.E., Lao, L.L., Liu, Y., Madireddy, S. and McClenaghan, J., 2024. Impact of various DIII-D diagnostics on the accuracy of neural network surrogates for kinetic EFIT reconstructions. *Nuclear Fusion*, *64*(8), p.086065.
5. L.L. Lao, S. Kruger, C. Akcay, P. Balaprakash, T.A. Bechtel, E. Howell, J. Koo, J. Leddy, M. Leinhauser, Y.Q. Liu, S. Madireddy, J. Meclenaghan, D. Orozco, A. Pankin, D. Schissel, S. Smith, X. Sun, and S. Williams, "Application of machine learning and artificial intelligence to extend EFIT equilibrium reconstruction", *Plasma Phys. Control. Fusion*, Volume 64, 074001, 2022.
6. Leddy, J., Madireddy, S., Howell, E. and Kruger, S., 2022. Single Gaussian process method for arbitrary tokamak regimes with a statistical analysis. *Plasma Physics and Controlled Fusion*, *64*(10), p.104005.
7. Liu, Y., Akcay, C., Lao, L.L. and Sun, X., 2022. Surrogate models for plasma displacement and current in 3D perturbed magnetohydrodynamic equilibria in tokamaks. *Nuclear Fusion*, *62*(12), p.126067.

B. Conference Paper/Presentation (in last two years):

S. Kruger, L.L. Lao, C. Akcay, O. Antepara, T.A. Bechtel, E. Howell J. Leddy, Y.Q. Liu, S. Madireddy, J. McClenaghan, D. Orozco, A. Pankin, D. Schissel, S.P. Smith, X. Sun, S. Williams, Improving the Accuracy and Speed of Equilibrium Reconstructions of Tokamak Plasmas Using Machine Learning, presented at 29th IAEA Fusion Energy Conference (FEC 2023) 16-21 October 2023, London, United Kingdom.

Akcay, C., Madireddy, S., Sun, X., Bechtel, T., McClenaghan, J., Samaddar, A., Kruger, S., Lao, L., Liu, Y. and Team, E.A., 2023. EFIT-AI neural network surrogates for magnetic, MSE, and kinetic equilibrium reconstruction. In *APS Division of Plasma Physics Meeting Abstracts* (Vol. 2023, pp. TI01-006).

Leddy, J., Howell, E., Kruger, S., Madireddy, S., Akcay, C., Bechtel, T., Lao, L., McClenaghan, J., Orozco, D., Smith, S. and Pankin, A., 2023. Gaussian Process Regression for Equilibrium

Reconstruction in DIII-D and ITER Plasmas. In *APS Division of Plasma Physics Meeting Abstracts* (Vol. 2023, pp. JO09-011).

Bechtel, T., Orozco, D., Kruger, S., Pankin, A., McClenaghan, J., Lao, L., Akcay, C., Sun, X., Smith, S. and Meneghini, O., 2023. Production of a FAIR Tokamak Equilibria Database for Analysis and Machine Learning. In *APS Division of Plasma Physics Meeting Abstracts* (Vol. 2023, pp. JO09-012).

Amara, T., Akcay, C., Sun, X., McClenaghan, J., Madireddy, S., Samaddar, A., Kruger, S., Pankin, A., Liu, Y., Howell, E. and Antepara, O., 2024. Improved Surrogate Models for DIII-D Equilibrium Reconstruction and Tools for Uncertainty Analysis Delivered by EFIT-AI for the Theory and Simulation Performance Target. *Bulletin of the American Physical Society.*