Physics of Plasmas

RESEARCH ARTICLE | AUGUST 05 2025

# Machine-learned trends in mirror configurations in the large plasma device

Phil Travis ; Jacob Bortnik ; Troy Carter

Check for updates

*Phys. Plasmas* 32, 082106 (2025)
https://doi.org/10.1063/5.0270755
CHORUS

View Online
Export Citation

## Articles You May Be Interested In

The upgraded Large Plasma Device, a machine for studying frontier basic plasma physics

*Rev. Sci. Instrum.* (February 2016)

Ponderomotive force driven density modifications parallel to B on the LAPD

*Phys. Plasmas* (April 2022)

Wave coupling and propagation from the helicon antenna in LAPD

*AIP Conf. Proc.* (August 2023)

13 January 2026 06:45:00

AIP Publishing

AIP Publishing
Physics of Plasmas
Special Topics Open for Submissions
Learn More

# Machine-learned trends in mirror configurations in the large plasma device

Cite as: Phys. Plasmas **32**, 082106 (2025); doi: 10.1063/5.0270755
Submitted: 13 March 2025 · Accepted: 21 July 2025 ·
Published Online: 5 August 2025

View Online Export Citation CrossMark

Phil Travis,[1,a] Jacob Bortnik,[2] and Troy Carter[1,3]

## AFFILIATIONS

[1]Department of Physics and Astronomy, University of California, Los Angeles, Los Angeles, California 90095, USA
[2]Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles, Los Angeles, California 90095, USA
[3]Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, USA

[a]Author to whom correspondence should be addressed: phil@physics.ucla.edu

## ABSTRACT

This study demonstrates the efficacy of machine learning (ML)-based trend inference using data from the large plasma device (LAPD). The LAPD is a flexible basic plasma science device with a high discharge repetition rate (0.25–1 Hz) and reproducible plasmas capable of collecting high-spatial-resolution probe measurements. A diverse dataset is collected through random sampling of LAPD operational parameters, including the magnetic field strength and profile, fueling settings, and the discharge voltage. Neural network ensembles with uncertainty quantification are trained to predict time-averaged ion saturation current ($I_{sat}$—proportional to density and the square root of electron temperature) at any position within the dataset domain. Model-inferred trends, such as the effects of introducing mirrors or changing the discharge voltage, are consistent with current understanding. In addition, axial variation is optimized via comprehensive search over $I_{sat}$ predictions. Experimental validation of these optimized machine parameters demonstrates qualitative agreement, with quantitative differences attributable to Langmuir probe variation and cathode conditions. This investigation demonstrates, using ML techniques, a new way of extracting insight from experiments and novel optimization of plasmas. The code and data used in this study are made freely available.

© 2025 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). https://doi.org/10.1063/5.0270755

13 January 2026 06:45:00

## I. INTRODUCTION

Understanding and controlling plasma behavior in fusion devices is necessary for developing efficient fusion reactors for energy production. Because of the complex, high-dimensional parameter space, traditional experimental approaches are often time-consuming and require careful planning. This work explores how machine learning (ML) techniques can accelerate this understanding by studying the effect of machine parameters in a basic magnetized plasma device. Trend inference is the process of relationship discovery. While ML methods, particularly neural networks (NNs), have become increasingly prevalent in fusion research for control and stabilization, their application to systematic trend discovery remains largely unexplored.

Many studies have used ML for profile prediction on a variety of tokamaks, particularly for real-time prediction and control. For example, NNs were used to predict electron density, temperature, and other quantities in DIII-D,[1] and reservoir NNs have demonstrated the ability to quickly adapt to new scenarios or devices.[2] Temporal evolution of parameters has been successfully modeled using recurrent neural networks (RNNs)[3] for multiple devices, including the EAST[4] and KSTAR tokamaks.[5,6] These predictions enabled training of a reinforcement learning-based controller.[5,6] In addition, a decision tree-based controller was trained to maximize $\beta_N$ while avoiding tearing instabilities[7] on DIII-D. Electron temperature profiles have also been predicted using dense NNs on the J-TEXT tokamak.[8]

A parallel focus has been on instability prediction and mitigation in tokamaks, particularly of disruptions. Notable achievements in disruption prediction include RNN-based disruption prediction[9] and random forest approaches,[10] with a comprehensive review available by Vega *et al.*[11] Recent work has extended to active control, such as the mitigation of tearing instabilities in DIII-D using reinforcement learning.[12]

While ML has proven effective for prediction and control tasks, inferring trends using data-driven methods has been relatively uncommon. Notable exceptions include finding scaling laws on the JET tokamak[13] via classical ML techniques and the development of the Maris density limit,[14] which outperforms other common scalings (including the Greenwald density limit) in predictive capability.

The use of machine learning and Bayesian inference in fusion research has been recently reviewed by Pavone *et al.*[15]


© Author(s) 2025

Outside of magnetized plasmas, the laser plasma community has embraced ML techniques for various applications, enumerated in a review by Döpp *et al.*[16] Data-driven plasma science, in general, has been reviewed by Anirudh *et al.*[17] Notably, a similar quasi-random method (Sobol sequences) was used to collect a spectroscopy dataset on a plasma processing device over diverse machine settings.[18] This process is similar to what is performed in our work here, but a generative variational autoencoder was instead trained to be used as an empirical surrogate model.

This work advances data-driven methods in plasma physics by taking these methods one step further: instead of learning a model for particular task (e.g., disruption prediction or profile prediction), we infer learned trends directly from the model itself.

The goal of this study is to develop a data-driven model that can provide insight into the effect of machine parameters on plasmas produced in large plasma device (LAPD) in lieu of a theoretical model. In contrast with tokamaks and other fusion devices, the LAPD is particularly well-suited for ML data collection because of its flexibility and the high repetition rate. We demonstrate the capability to infer trends in a particular diagnostic signal, the time-averaged ion saturation current ($I_{sat}$), for any mirror (or anti-mirror) field geometry in a variety of machine configurations. Langmuir probes are commonly used to measure density, temperature, and potential in virtually all plasma devices in low-temperature (less than tens of an electron volts) regimes. The $I_{sat}$ signal, in particular, is almost always used in the LAPD for calculating local plasma density.

This study performs two firsts in magnetized plasma research: using NNs to directly infer trends and collecting data efficiently with partially randomized machine parameters. We also demonstrate optimizing LAPD plasmas, given any cost function by minimizing axial variation in $I_{sat}$. This global optimization is only possible using ML techniques. This work demonstrates the usefulness of a pure ML approach to modeling device operation and shows how this model can be exploited. We encourage existing ML projects and experiments to consider this approach if possible. Acquiring sufficiently diverse datasets may require assuming some risk because diverse data, such as discharges from randomly sampled machine settings, may not be amenable to conventional analysis techniques.

All the processed data used for training the models in this study are freely available.[19] Other devices have also made data publicly available. Namely, data for H-mode confinement scaling have been available since 2008,[20] and more recently, some MAST[21] and all LHD[22] data are now publicly available.

This paper is organized as follows: Sec. II discusses the LAPD and the data acquisition methodology. Sections III and IV detail the development of the model and uncertainty quantification. Section V presents model validation, followed by trends in discharge voltage and gas puff duration in Sec. VI. Section VII demonstrates optimization of the axial $I_{sat}$ profile, with the discussion and conclusion in Secs. VIII and IX.

## II. DATA COLLECTION AND PROCESSING

### A. The large plasma device (LAPD)

The large plasma device (LAPD)[23,24] is a basic plasma science device located at the University of California, Los Angeles. The LAPD produces up to 18 m long, 1 m diameter plasmas with densities up to $3 \times 10^{13}$ cm$^{-3}$ and temperatures up to 20 eV, though typical operation yields temperatures around 5 eV. Probes can sample virtually any point in this plasma through unique ball valves placed every 32 cm along the length of the device, enabling the collection of time-series data with high-spatial resolution. The discharge repetition rate is configurable between 0.25 and 1 Hz. Additionally, the LAPD has 13 independently controllable magnet power supplies to shape the geometry of the axial magnetic field. The discharge is formed by 35 cm diameter lanthanum hexaboride (LaB6) cathode[24] and 72 cm molybdenum anode 0.5 m away (−z direction) at the southern end (+z) of the device. A cartoon of the LAPD and relevant coordinate system can be seen in Fig. 1.

The LAPD has many experimental control parameters for various physics studies. While the device can accommodate various insertable components, this study focuses on the parameters fundamental to the operation of the main cathode. Specifically, halfway between the cathode and anode are three gas puff valves: East, West, and top. The aperture, duration, and triggering of these valves have a large impact on plasma formation. A static gas fill system also exists, but it is not used in this study. The cathode–anode voltage (and consequently, discharge power) strongly influences plasma density and temperature downstream of the source. Additionally, the magnetic field configuration substantially shapes the plasma column. One crucial variable not considered in this study is the cathode temperature, as its adjustment and equilibration require many hours, limiting dataset diversity. This combination of diagnostic coverage, high repetition rate, and extensive configurability renders the LAPD particularly suitable for machine learning studies.

### B. Data collection and processing of $I_{sat}$ signals

The ion saturation current, denoted as $I_{sat}$, is obtained by applying a sufficiently negatively bias to a Langmuir probe to ensure the exclusive collection of ions. This collected current is proportional to $Sn_e\sqrt{T_e}$, where $n_e$ and $T_e$ are the electron density and temperature, respectively, and $S$ is the effective probe collection area. To account for differences in probe tip geometry, the $I_{sat}$ values are normalized to area. Under typical conditions, an $I_{sat}$ value of 1 mA/mm$^2$ corresponds to $n_e \approx 1$–$2 \times 10^{12}$ cm$^{-3}$ for a $T_e$ from 4 to 1 eV.

Data collection was conducted in two campaigns separated by 14 months. The initial run set is designated as DR1 and the subsequent run set as DR2. These run sets are further broken down into *dataruns*, which are series of discharges (shots) with identical operational machine parameters. A total of 67 dataruns were collected over both campaigns.

$I_{sat}$ measurements were averaged over 10 to 20 ms to exclude plasma ramp-up and fluctuations. Example $I_{sat}$ probe data can be seen in Fig. 2 alongside gas puff timings. Profile evolution is not studied to minimize computational requirements. $I_{sat}$ characteristics vary significantly between axial (z) position machine parameters. For $I_{sat}$ measurements on the same probe as a Langmuir sweep (DR2 port 26, z = 863 cm), the averaging process excludes the sweep period with an additional 40 $\mu$s buffer. $I_{sat}$ measurements in DR1 that saturated either the isolation amplifier or digitizer are excluded from the dataset. Only 484 shots were removed out of ≈132 000, so the impact on the aggregate dataset is minimal.

The LAPD control parameters varied in this study were the source field, mirror field, midplane field, gas puff valve voltage, gas puff duration, and discharge voltage. The magnetic field regions are labeled in Fig. 1 and effectively control the width of the plasma relative

13 January 2026 06:45:00


© Author(s) 2025

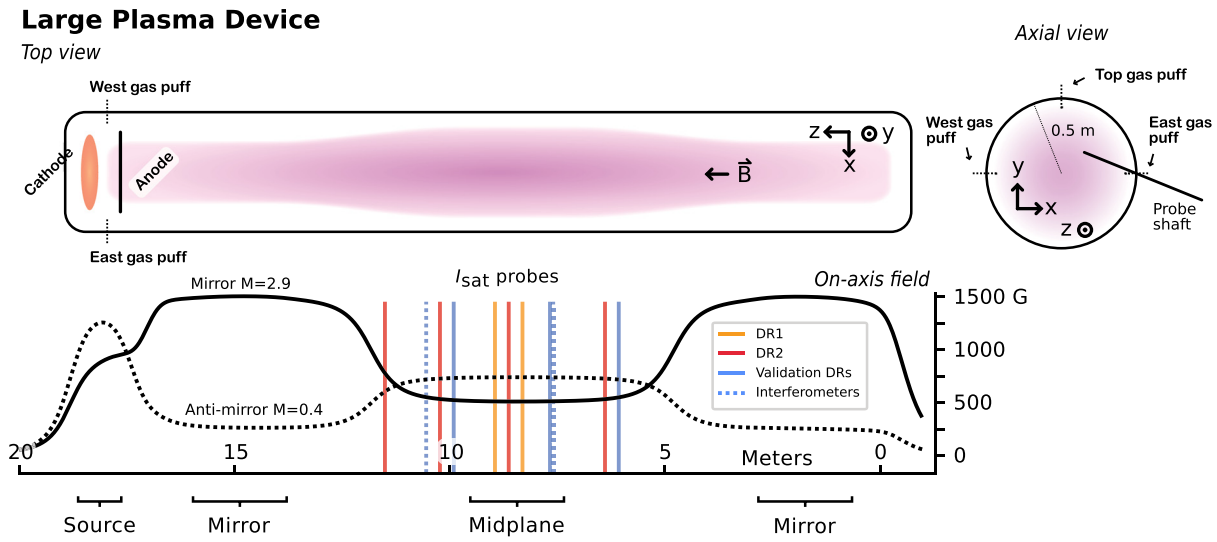


**FIG. 1.** A cartoon of the large plasma device, the coordinate system used, examples of a mirror and anti-mirror magnetic field configuration, and probe locations used in this study. The source, mirror, and midplane regions are labeled; the three fields were programmed independently.

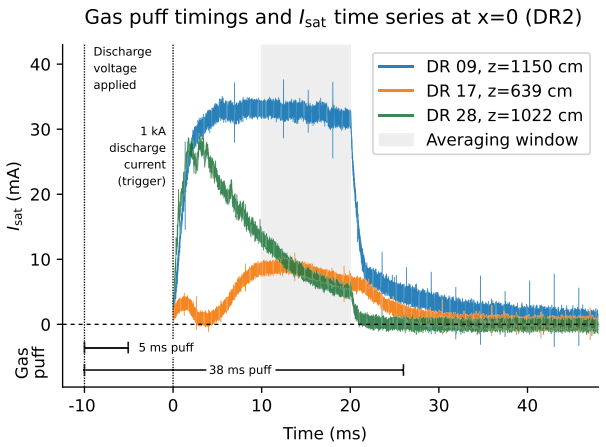


**FIG. 2.** Gas puff timings and example $I_{sat}$ time series at three different z-axis locations from three different dataruns. Note that some discharges do not achieve steady state in $I_{sat}$.

to the cathode in their respective regions. The gas puff voltage governs gas flow rate into the chamber, though this relationship is not yet quantified, and the gas puff duration defines the piezo valve activation period. The discharge voltage is applied across the cathode and anode 10 ms after gas puff initiation. While discharge voltage correlates with discharge current (and thus power), the current depends on the machine configuration and downstream conditions and cannot be predetermined.

These machine parameters—with the exception of gas puff duration—were randomly sampled via Latin-hypercube sampling (LHS) for 44 of the dataruns. Data were then collected with these settings. Gas puff duration was reduced for the last seven runs to 20, 10, or 5 ms (see Fig. 2 for timings relative to $I_{sat}$ signals). The breakdown of each setting in the dataset is given in Appendix A, Table IV. The top gas puff valve was used for only the first nine dataruns of DR2 because of equipment issues.

$I_{sat}$ measurements were acquired along $y = 0$ lines (51 dataruns total) or x–y grids (16 dataruns total) with spatial resolutions varying between 1.5 and 2 cm. The fixed axial locations of the probes were 895 cm and 831 in DR1 and 1150, 1022, 863, and 639 cm for DR2 (Fig. 1). Six shots were recorded at each position except for the first four dataruns in DR1 with five shots each. While $I_{sat}$ exhibits a small degree of shot-to-shot variation, the present model only learns the expected value, leaving distributional modeling to future generative approaches.

### C. Dataset distribution and bias

Neural network inputs comprise 12 variables: source field, mirror field, midplane field, gas puff voltage, discharge voltage, gas puff duration, probe coordinates (x, y, z), probe rotation, run set identifier, and top gas puff flag. These variables can be interpreted as six control parameters, four probe coordinates, and two flags. These inputs are mean-centered and normalized to the peak-to-peak value with no outliers in the dataset. The baseline models trained in Sec. III B did not contain the run set identifier or top gas puff flag.

Measurements over an x–y plane, constituting ≈64% of all shots, are predominantly acquired overnight for maximal machine utilization. These longer dataruns lead to particular machine configurations being overrepresented in the dataset.

The dataset predominantly contains gas puff durations of 38 ms. Only 6 runs in the training set have gas puff durations less than 38 ms: three have 5 ms and three have 10 ms, each having mirror ratios 1, 3, and 6 but otherwise identical configurations in an attempt to see mirror-related interchange instabilities in higher-temperature, lower-collisionality regimes. The 20 ms gas puff duration case is in the test set. This sampling bias toward the 38 ms gas puff duration suggests that poor model performance is to be expected in shorter gas puff

13 January 2026 06:45:00


© Author(s) 2025

regimes. The top gas puff valve was operational for only the first nine runs of DR2.

The $I_{sat}$ distribution is dissimilar between DR1 and DR2: DR1 appears to have a more uniform distribution. Combining the two datasets results in many $I_{sat}$ examples less than 2 mA/mm$^2$ and a sharp decrease in number of examples above 10 mA/mm$^2$. Thus, we expect the model to perform better for smaller $I_{sat}$ values than larger ones. Data bias is further discussed in Appendix A.

## III. MODEL DEVELOPMENT, TRAINING

### A. Training details

For initial experiments in training the model, a mean-squared error (MSE) loss is used:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{m}\sum_{i=1}^{m}\left(f(x_i) - y_i\right)^2, \tag{1}$$

where $x_i$ represents the input vector for the ith example, $y_i$ the target measurement, m the batch size, and f the NN. During training, overfitting was assessed via the validation set MSE with a traditional 80-20 train-validation random split. Unless stated otherwise, a dense neural network, four hidden layers deep and 256 units wide ($\approx$200 000 parameters), was trained with AdamW using a learning rate of $3 \times 10^{-4}$. Leaky ReLU activations (the nonlinearities in the NN) and adaptive gradient clipping[25] (cutting gradients norms above the 90th percentile of recent norms) were used to mitigate vanishing gradients and mitigate exploding gradients, respectively. The models were evaluated after training concluded at 500 epochs.

### B. Baselines for mean-squared error

A model was first trained with zeroed-out inputs as a baseline and to validate the data pipeline. This model effectively has only a single, learnable bias parameter at the input. This process yields a validation loss (simply MSE in this case) of 0.036. A linear model, though incapable of reasonably fitting the data, was trained as a performance baseline and to validate the data pipeline. This baseline linear model reaches a training and validation loss (MSE) of around 0.014. These initial models used tanh activations, though the impact of using a different activation function is minimal in these cases. A summary of these baselines is seen at the top of Table I.

### C. Effects of training set and model sizes

To study the effects of reduced diversity, the number of unique dataruns in the training set was systematically reduced while evaluating on a fixed test set. The test set loss monotonically increased with this decrease in datarun count. Part of this decrease may be caused by a simple reduction in training set size. In addition, models were individually trained and evaluated on DR1 only or DR2 only. When evaluated on the left-out run set, the test set losses were high, near or above the zero-input baseline of $3.6 \times 10^{-2}$. This result suggests that both run sets contain significant information missing in the other, and training on both provides beneficial information on the structure of the $I_{sat}$ measurement despite different probe calibrations and cathode state.

A larger model, consisting of a 12-deep 2048-wide dense network, was trained on the full training dataset, evaluated at 30 epochs. This larger model yielded a test MSE of $2.8 \times 10^{-3}$, indicating that these NNs are behaving as expected. Longer training or larger models may yield better test set results but will likely not come close to the training and validation losses, which are on the order of $10^{-5}$. Combined models with differing initializations (an ensemble) were trained to measure the MSE variance over model parameters, which was about 16%. When the $I_{sat}$ predictions were averaged, the test set MSE was $2.9 \pm 1.1 \times 10^{-3}$, achieving the best performance for that model size. These test set losses are also seen in Table I.

**TABLE I.** Summary of test set losses for different training data and ensembles.

| Model | MSE $\times 10^{-3}$ |
|---|---|
| Zeroed-input | 36 (validation) |
| Linear model | 14 (validation) |
| 9 dataruns | 7.0 |
| 19 dataruns | 6.9 |
| 29 dataruns | 4.2 |
| 39 dataruns | 4.1 |
| 49 dataruns | 3.4 |
| DR1 only | 6.4 |
| DR2 only | 5.4 |
| Full set, large model | 2.8 |
| Full set average | $3.6 \pm 0.56$ |
| Full set ensemble | $2.9 \pm 1.1$ |
| "Run set" flag ensemble | $1.9 \pm 0.64$ |
| "Top gas puff" flag | 1.8 |

### D. Improving performance with machine state flags

Data from DR1 and DR2 were collected 14 months apart leading to differing machine states. In DR1, only one turbo pump was operating leading to much higher neutral pressures than in the DR2 run set. A new parameter (mean-centered and scaled) was added to the inputs to distinguish between these two run sets. All the predictions in this study use the DR2 run set flag (a value of 1.0) because turning off the turbopumps is not a commonly desired mode of operating the LAPD. The inclusion of this parameter also provides the model the ability to distinguish between the probe calibration differences between DR1 and DR2. An ensemble prediction with this run set flag brings the test set MSE down to $1.9 \times 10^{-3}$.

A flag indicating when the top gas puff valve was enabled in DR2 was also added to all training data, allowing the model to further distinguish between different fueling cases. The addition of this flag incrementally improved test set MSE to $1.8 \times 10^{-3}$. The effect on MSE on the inclusion of these new parameters is compared to the performance of other models in Table I.

### E. Learning rate scheduling

Modifying the learning rate over time (scheduling) is known to improve model learning. The following schedules were compared: constant learning rate ($\gamma = 3 \times 10^{-4}$), $\gamma \propto \text{epoch}^{-1}$, $\gamma \propto \exp(-\text{epoch})$, and $\gamma \propto \text{epoch}^{-1/2}$. The epoch is the training step divided by the number of batches in one epoch, so "epoch" in this case takes on a floating-point value. $\gamma \propto \text{epoch}^{-1}$ appears to give the best test set loss by a test

13 January 2026 06:45:00


© Author(s) 2025

MSE difference of $1 \times 10^{-4}$, and any schedule beats a constant learning rate by a difference of $2 - 4 \times 10^{-4}$.

## IV. UNCERTAINTY QUANTIFICATION

### A. $\beta$-NLL loss

Instead of predicting a single point, the model can predict a mean $\mu$ and variance $\sigma^2$ using the negative-log likelihood (NLL) loss[26,27] by assuming a Gaussian likelihood. An adaptive scaling factor $\text{StopGrad}(\sigma_i^{2\beta})$ is introduced, which can be interpreted as an interpolation between an MSE loss and Gaussian NLL loss, yielding the $\beta$-NLL loss:

$$\mathcal{L}_{\beta-\text{NLL}} = \frac{1}{2}\left(\log \sigma_i^2(\mathbf{x}_n) + \frac{(\mu_i(\mathbf{x}_n) - y_n)^2}{\sigma_i^2(\mathbf{x}_n)}\right) \text{StopGrad}\left(\sigma_i^{2\beta}\right), \tag{2}$$

for example $n$ and model $i$, with an implicit expectation over training examples. $\beta = 0$ yields the original Gaussian NLL loss function and $\beta = 1$ yields the MSE loss function. This factor improves MSE performance by scaling via an effective learning rate for each example (which necessitates the StopGrad operation)[28] and improves both aleatoric and epistemic uncertainty quantification.[29] $\beta = 0.5$ was used by default in this study. This $\beta$-NLL loss function also improved training stability.

This NLL-like loss assumes the prediction—the likelihood of $y$ given input $\mathbf{x}$: $p(y|\mathbf{x})$—following a Gaussian distribution. Treating each prediction as an independent random variable [considering each model in the ensemble is sampled from some weight distribution $\theta \sim p(\theta|\mathbf{x}, y)$] and finding the mean of the random variables yields a normal distribution with mean $\mu_*(\mathbf{x}) = \langle \mu_i(\mathbf{x}) \rangle$ and variance $\sigma_*^2 = \langle \sigma_i^2(\mathbf{x}) + \mu_i^2(\mathbf{x}) \rangle - \mu_*^2(\mathbf{x})$, where $\langle\rangle$ indicates an average over the ensemble.

The ensemble predictive uncertainty can be broken down into the aleatoric and epistemic components:[29] the aleatoric uncertainty is $\langle \sigma_i^2(\mathbf{x}) \rangle$ and the epistemic uncertainty is $\langle \mu_i^2(\mathbf{x}) \rangle - \mu_*^2(\mathbf{x}) = \text{Var}[\mu_i(\mathbf{x})]$. The intuition behind these uncertainties is that the random fluctuations in the recorded data are captured in the variance of a single network, $\sigma_i^2$. If the choice of model parameters was significant, we would expect the predicted mean for a single model, $\mu_i$, to fluctuate as captured by $\text{Var}[\mu_i(\mathbf{x})]$.

### B. Cross-validation MSE

For cross-validation, multiple train-test set pairs were created. Test set 0 comprises deliberately chosen dataruns to encompass a diverse set of machine settings and probe movements. The other six datasets were compiled with randomly chosen dataruns (without replacement) while keeping the number of dataruns from DR1 and DR2 equal. Seven model ensembles (5 NNs per ensemble—35 NNs total) were trained to evaluate the effect of test set choice on model MSE. The median ensemble test set MSE for these seven sets was $2.13 \times 10^{-3}$ with a mean of $3.6 \times 10^{-3}$. The handpicked dataset had an ensemble test set MSE of $1.85 \times 10^{-3}$, indicating that the choice of dataruns was adequately representative. This median MSE will be used to estimate model prediction error in addition to uncertainty quantification. This cross-validation also provides an error estimate if the models were to be trained on *all* dataruns. Ensembles always outperformed the average error of single-model predictions.

All validation set MSEs fall between 1 and $6 \times 10^{-5}$, with the average training MSE falling within that range as well. These MSEs indicate that the model is able to fit the training data to a high degree of accuracy regardless of which dataruns are held out. The loss and MSE curves over training epochs can be seen in Fig. 11 in Appendix B.

### C. Model calibration via weight decay

The predicted uncertainty may not provide an accurate range of $I_{\text{sat}}$ values when compared to the measured value. Calibrating the model means changing the predicted uncertainty range so that the measured values fall within that range according to some distribution, such as a Gaussian in this case. One of the ways assessing this calibration is by the z-score of predictions, where $z_n = (x_n - \mu_n)/\sigma_n(x_n)$ for example $x_n$, predicted mean $\mu_n$, and standard deviation $\sigma_n$. Perfect model calibration would lead to identical z-score distribution $\mathcal{N}(\mu = 0, \sigma = 1)$ for the training test sets. When evaluated on the training set, the distribution should be a Gaussian with a standard deviation of 1.

Increased weight decay can lead to better model calibration.[30] Weight decay penalizes large parameter values by adding the L2 norm of model weights to the loss. Model ensembles were trained with weight decay coefficients between 0 and 50 to determine the best calibrated model determined by the distribution of z-scores of the training and test sets. The results of this weight decay scan are seen in Fig. 3. Increasing the weight decay increases the test MSE and decreases its z-score standard deviation. This large standard deviation is caused by outliers. Excluding z-scores magnitudes above 10, or 4.4% of the test set, yields a standard deviation of 2.53. This long tail indicates that the distribution of predictions on the test set is not Gaussian. Nonetheless, the trend remains that increasing weight decay leads to smaller test set z-score standard deviations. However, the test set MSE increases after a weight decay of 1. This increase in test MSE implies that the model is making less accurate predictions but is better calibrated. Highly biased models are better calibrated but come at great expense of mean

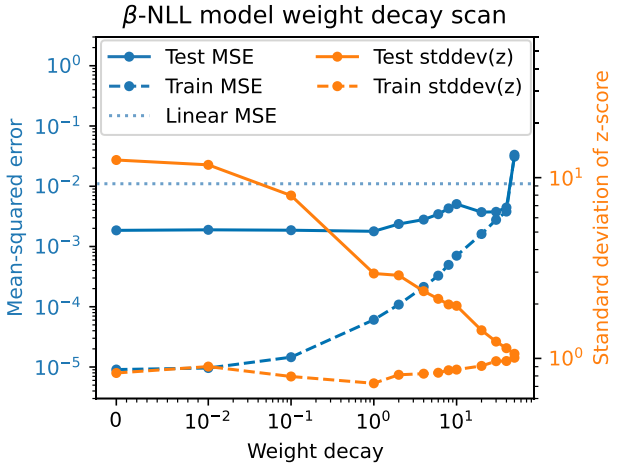


**FIG. 3.** Model performance and calibration for different weight decays. Highly biased models are better calibrated but come at great expense of mean prediction error. At the weight decay value of 50, the model has worse error than a linear model. Note the linear scale below $10^{-2}$.

13 January 2026 06:45:00


© Author(s) 2025

prediction error. At the weight decay value of 50, the model has worse error than a linear model. Despite the attempts using weight decay, the model never becomes well-calibrated: the predicted uncertainty is always too low by a factor of 2–5.

Despite the better calibration, the uncertainty predicted by a model with a large weight decay is decidedly worse: the uncertainty is similar across an entire profile, and when projected beyond the training data, the total uncertainty remains largely constant as seen in Fig. 4. The zero weight decay model exhibits relatively increasing uncertainty beyond the bounds of the training data. Although not well-calibrated, this uncertainty can provide a hint of where the model lacks confidence relative to other predictions, even though the uncertainty is much less than it should be.

## V. EVALUATING MODEL PERFORMANCE

Model performance is evaluated in three ways by comparing against intuition from geometry, an absolute measurement, and extrapolated machine conditions.

### A. Checking geometrical intuition

Assuming magnetic flux conservation, we know that modifying the mirror geometry can control the effective width of the plasma. One way to check that the model is learning appropriate trends is to check with this intuition. If the magnetic field at the source is not equal to the field at the probe, the probe will see the plasma expanded (or contracted) by roughly a factor of $\sqrt{B_{\text{probe}}/B_{source}}$. The cathode is about 35 cm in diameter, so a magnetic field ratio of 3 would give a plasma approximately 60 cm in diameter. All the probes used in this study are in or very close to the zero-curvature midplane region of a mirror.

To check this intuition, the model is given the following inputs: $B_{\text{source}} = 500$ G, $B_{\text{mirror}} = 1500$ G, $B_{\text{midplane}} = 500$ G, discharge voltage = 110 V, gas puff voltage = 70 V, gas puff duration = 38 ms, run set flag = DR2, and top gas puff = off. The discharge voltage and gas puffing parameters were arbitrarily chosen. The x coordinate is scanned from 0 to 30 cm, and the z coordinate from 640 to 1140 cm. This discharge is then modified by separately changing $B_{\text{source}}$ to 1500 G and $B_{\text{midplane}}$ to 750 G (M = 1.5). The x profiles at the midplane (z = 790 cm) of the reference M = 3 prediction, source field change, and midplane field change, all scaled to cathode radius, can be seen in Fig. 5. Changing the source field to 1500 G increases the $I_{\text{sat}}$ toward the edge of the plasma, as expected. When the midplane field is increased, the $I_{\text{sat}}$ values decrease at the edge and increase at the core (x = 0 cm), implying a thinner plasma column, and are consistent with previously measured behavior. When only the mirror field is modified (not shown), the strongest effect on $I_{\text{sat}}$ is on or near x = 0 cm, and the plasma column width does not appear to appreciably change.

### B. Directly comparing prediction to measurement

$I_{\text{sat}}$ measurements were taken with the following LAPD machine settings: $B_{\text{source}} = 1250$ G, $B_{\text{mirror}} = 500$ G, $B_{\text{midplane}} = 1500$ G, discharge voltage = 90 V, gas puff voltage = 90 V, gas puff duration = 38 ms, run set flag = DR2, and top gas puff = off. These settings were from a previous discharge optimization attempt. The probes utilized were the permanently mounted 45° probe drives. These probes were known to have identical effective areas relative to each other from the previous experiment and from analyzing the discharge ramp-up.

Because of data acquisition issues, only a single useful shot was collected at a nominal −45° angle (relative to the x-axis) 10 cm past the center (x = 0 cm, y = 0 cm) of the plasma on three probes at z-positions of 990, 767, and 607 cm (ports 22, 29, and 34, respectively). The probe drives were slightly uncentered, leading to the real coordinates of the probes to be around x = 9.75 cm and y = −8.4 cm. Note that the model can predict anywhere in LAPD bounded by the training data, so off-axis measurements are not an issue. The resulting predictions using these coordinates and machine conditions can be seen in Fig. 6. The model reproduces the axial trend well but slightly underestimates $I_{\text{sat}}$ on an absolute basis. However, given the lack of absolute $I_{\text{sat}}$ calibration and variable machine state, the agreement of the absolute $I_{\text{sat}}$ values may be coincidental. Nevertheless, the trend exhibited by this validation study matches the predicted trend and increases our trust in model predictions.

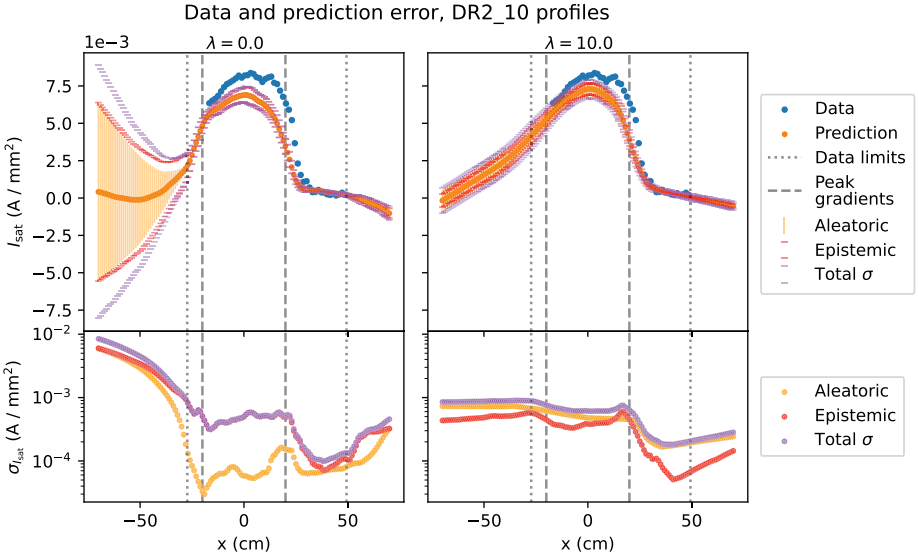


**FIG. 4.** Model extrapolation performance (top plots) with uncertainty (bottom plots) for a model ensemble trained on a $\beta$-NLL loss function. DR2 run 10 was chosen as an illustrative example. The *relative* uncertainty appears to be more useful when zero weight decay ($\lambda = 0$, left) is used: the uncertainty increases when the model is predicting outside its training data along the x-axis.

13 January 2026 06:45:00


© Author(s) 2025

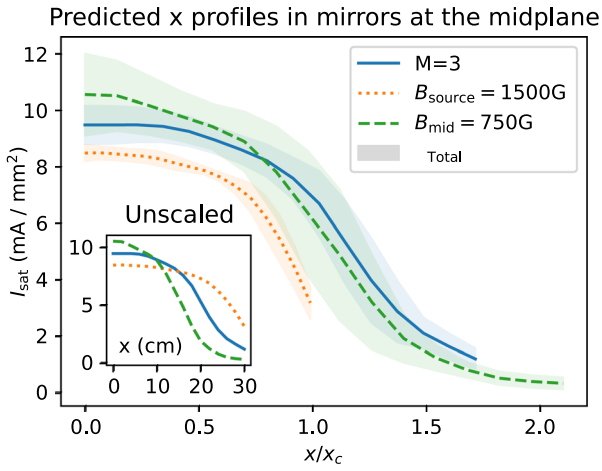


**FIG. 5.** Plot of various mirror configurations scaled to the cathode radius $x_c = 17.5$ cm at the midplane (z = 790 cm). When scaled according to the expected magnetic expansion, the profiles generally agree. The smaller the plasma diameter (and thus smaller volume), the higher the peak in $I_{sat}$ at the core, as expected.

An additional validation datarun was performed. For this run, the discharge voltage was increased to 160 V, and the source field changed to 822 G. The training data contain discharge voltages up to 150 V, so this case tests the extrapolation capabilities of the model. The comparison of model predictions and the measured data can be seen in Fig. 6. As stated earlier, the absolute uncertainty provided of the model is not calibrated. However, note that the level of uncertainty provided by the model and the large spread in model predictions are much greater than seen in the interpolation regime (Fig. 6) and eclipse the cross-validated test set root mean-squared error (RMSE). We conclude that this model has good interpolation capabilities, but extrapolation—as with any model—remains difficult.

## VI. INFERRING TRENDS

A systematic study of the impact of discharge voltages on $I_{sat}$ profiles has not been conducted using conventional techniques. Collecting both z- and x-axis profiles over a wide range of discharge voltages would take a considerable amount of time, mostly from the requirement to dismount and reattach the probes and probe drives along the length of the LAPD. This study has now been performed using the learned model, circumventing these time-consuming challenges. Model input parameters were chosen to be common, reasonable values: 1 kG flat field, 80 V gas puff, 38 ms gas puff duration, run set = DR2, and top gas puff off. The 38 ms puff is used in these predictions because it is the most common gas puff duration in the training set—the model is biased in favor of this gas puff setting. The results of changing the discharge voltage only can be seen in Fig. 7. Notably, the $I_{sat}$ increases across both axes. Steeper axial gradients are seen with lower discharge voltages, but peaked x-profiles are seen at higher discharge voltages. The area closer to the source region (+z direction) appears to have a steep drop, but flatter profiles down the length of the machine.

Unfortunately, the discharge current was not included as an output in the training set. Otherwise, the effect of changes in discharge power, rather than simply voltage, could be computed. The discharge

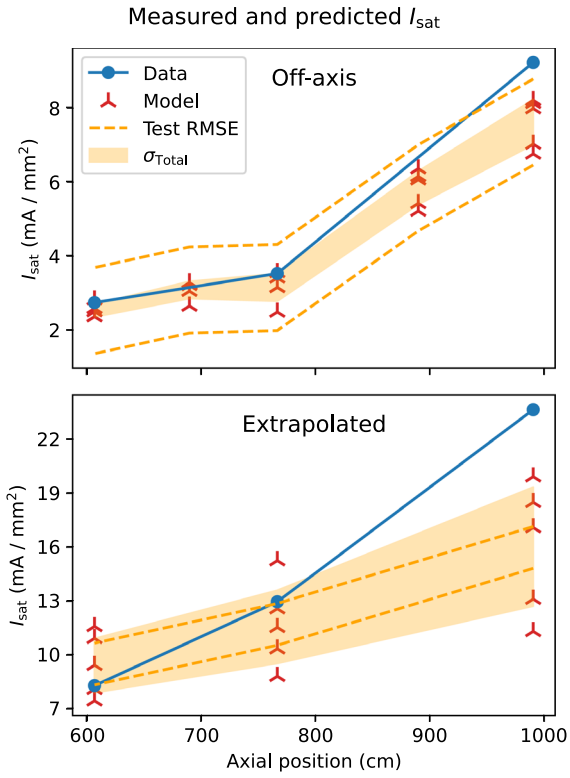


**FIG. 6.** Top: data collected at off-axis positions around x = 9.75 cm and y = −8.4 cm are compared with predictions from the machine learning model at the same points in addition to two interpolating predictions. The model predicts the trend well but underestimates $I_{sat}$ in general. The shaded orange region is the total model uncertainty ($\sigma = \sqrt{\text{Var}}$). Bottom: Measured vs predicted $I_{sat}$ values for an odd machine configuration with $B_{source} = 822$ G and discharge voltage = 160 V. The training data only cover discharge voltages up to 150 V. The machine was also in an odd discharge state, so it is no surprise that the predicted uncertainty bounds are very large (even greater than the test set RMSE value) and that accuracy suffers.

current—thus discharge power—is set by cathode condition, cathode heater settings, and the downstream machine configuration and, thus, cannot be set to a desired value easily before the discharge. Discharge voltage, however, can remain fixed.

Of particular interest for some LAPD users is achieving the most uniform axial profile possible. We explore this problem in the context of mirrors. The gas puff duration is known to be a large actuator for controlling density and temperature and so is explored as a way of shaping the axial profile. We predict discharges with a flat 1 kG field with the probe in the center. The discharge voltage was set at 110 (a reasonable, middle value) with the run set flag = DR2 and top gas puffing = off. The inferred effect of gas puff duration on the axial gradient and axial gradient scale length can be seen in Fig. 8. Care was taken to handle the aleatoric (independent) uncertainty separately from the axially dependent epistemic uncertainty. As seen in the figure, the mean axial gradient decreases when the gas puff duration is shortened. The gradient scale length also increases, so the mean gradient is not decreasing simply because the bulk $I_{sat}$ is decreasing. This result suggests that the gas puff duration may be a useful actuator to consider when planning future experiments.

13 January 2026 06:45:00


© Author(s) 2025

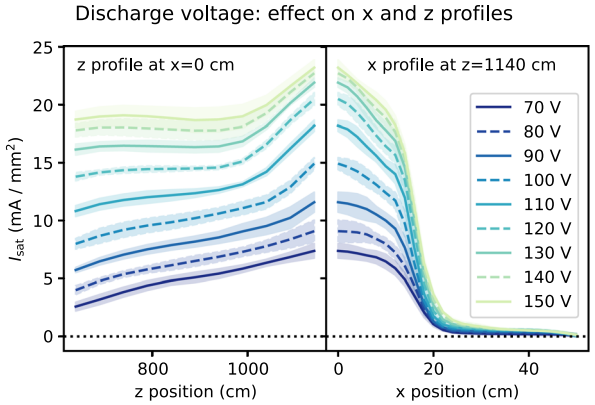


**FIG. 7.** The z profile at x = 0 cm and x profile at z = 1140 cm for different discharge voltages. The $I_{sat}$ decreases with increasing voltage, and the error (filled regions) stays roughly the same, but, in general, increases slightly toward the cathode and at higher discharge voltages.

These applicability of these results are somewhat muted because the gas puff duration was not chosen randomly in the training discharges. Given this lack of data diversity, the accuracy and applicability of this study must be interpreted cautiously. When a model is trained on *all* data available (using the cross-validated test set MSE as a guide for error), which includes the 20 ms gas puff case, the predicted gradient scale length decreases uniformly across the duration scan by 1 m. The fact that the trend remains intact when an additional, randomized intermediate gas puff case is added gives confidence in the predictions of the model despite the lack of data diversity.

## VII. OPTIMIZING PROFILES

One particular issue seen in LAPD plasmas is sharp axial density and temperature gradients. Some experiments require flat gradients, such as Alfvén wave propagation studies. We explore optimizing the axial $I_{sat}$ variation as an approximation to this problem. In addition,

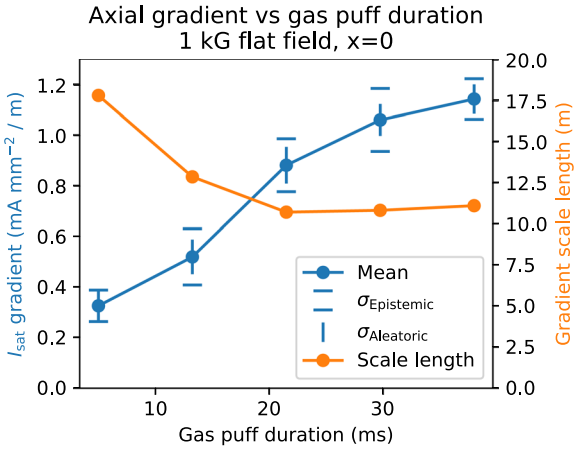


**FIG. 8.** ML prediction: mean axial gradients decrease with decreased gas puff duration. Five durations are plotted between 5 and 38 ms (which are the bounds of the training data), evenly spaced. The gradient scale length also increases, indicating that the gradient change was not just from a decrease in the bulk $I_{sat}$.

in this case, the optimization problem is used as a tool to evaluate the quality of the learned model. This is a very demanding task because the trends inferred by the model along all inputs must simultaneously be accurate. Constraints on this optimization further increase the difficulty of the problem. Success in optimization provides strong evidence that the model has inferred relevant trends in predicting $I_{sat}$. We quantify the uniformity of the axial profile by taking the standard deviation of $I_{sat}$ of 11 points along the z-axis ($x, y = 0$). The required LAPD state for attaining the most uniform axial profile can be found by finding the minimum of this standard deviation with respect to the LAPD control parameters and flags:

$$\text{Inputs} = \underset{\text{Inputs}\neq z}{\arg\min}\, \underset{z}{\text{sd}}(I_{sat}|_{x=0}). \tag{3}$$

The largest axial variation can likewise be found by finding the maximum. The model inputs used for this optimization can be found in Table II.

For this optimization, we use an ensemble of five $\beta$-NLL-loss models with weight decay $\lambda = 0$. The $\lambda = 0$ model is used because it appears to give the most useful uncertainty estimate (seen in Fig. 4). The optimal machine actuator states are found by feeding a grid of inputs into the neural network. This variance estimate is not well-calibrated: the error of the predictions on the test set falls far outside the predicted uncertainty. However, this uncertainty can be used in a relative way: when the model is predicting far outside its training range, the predicted variance is much larger. The ranges of inputs into this model are seen in Table II. These inputs yield 127 234 800 different machine states (times five models), which takes 151 s to process on an RTX 3090 ($\approx 4.2 \times 10^6$ forward passes per second) when implemented in a naive way. The number of forward passes can be reduced by a factor of 51 if the x value is set to 0 cm. Note that gradient-based methods can be used for search because the network is differentiable everywhere but this network and parameter space is sufficiently small that a comprehensive search is computationally tractable.

Like any optimization method, the results may be pathologically optimal. In this scenario, the unconstrained minimal axial variation is found when the $I_{sat}$ is only around 1 mA/mm$^2$, which is quite small and corresponds to 1–2 $\times 10^{12}$ cm$^{-3}$ depending on Te. The inputs corresponding to this optimum are given in the second column of

**TABLE II.** Machine inputs and actuators for model inference.

| Input or actuator | Range | Step | Count |
|---|---|---|---|
| Source field | 500–2000 G | 250 G | 7 |
| Mirror field | 250–1500 G | 250 G | 6 |
| Midplane field | 250–1500 G | 250 G | 6 |
| Gas puff voltage | 70–90 V | 5 V | 5 |
| Discharge voltage | 70–150 V | 10 V | 9 |
| Gas puff duration | 5–38 ms | 8.25 ms | 5 |
| Probe x positions | −50 to 50 cm | 2 cm | 51 |
| Probe y positions | 0 cm | ⋯ | ⋯ |
| Probe z positions | 640–1140 cm | 50 cm | 11 |
| Probe angle | 0 rad | ⋯ | ⋯ |
| Run set flag | Off and on | 1 | 2 |
| Top gas puff flag | Off and on | 1 | 2 |

13 January 2026 06:45:00


© Author(s) 2025

**TABLE III.** Machine inputs and actuators for optimized axial profiles.

| Input or actuator $I_{sat}$ constraint (mA/mm$^2$) | Weakest $I_{sat}$ = any | Weakest $I_{sat} > 7.5$ | Strongest $I_{sat} > 7.5$ | Intermediate $I_{sat} > 7.5$ |
|---|---|---|---|---|
| Source field (G) | 750 | 1000 | 500 | 2000 |
| Mirror field (G) | 1000 | 750 | 500 | 1250 |
| Midplane field (G) | 250 | 250 | 1500 | 750 |
| Gas puff voltage (V) | 70 | 75 | 90 | 90 |
| Discharge voltage (V) | 130 | 150 | 150 | 120 |
| Gas puff duration (ms) | 5 | 5 | 38 | 38 |
| Run set flag | On | On | On | On |
| Top gas puff flag | On | Off | Off | Off |

Table III. This density range is below what is required or useful for many studies in the LAPD.

Since many physics studies require higher densities, we constrain the mean axial $I_{sat}$ value to be greater than 7.5 mA/mm$^2$ (roughly 0.5–2 $\times$ 10$^{13}$ cm$^{-3}$). The "run set flag" is set to "on" for cases to be validated (bolded in Table III) because we wish to keep the turbopumps to represent typical LAPD operating conditions. In addition, the "top gas puff flag" was set to "off" to minimize the complexity of operating the fueling system on follow-up dataruns and experiments. Turning the top gas puff valve on is predicted to decrease the average $I_{sat}$ by $-2$ mA/mm$^2$ for strongly varying profiles, but otherwise, the shapes are very similar. The inputs corresponding to the maximum and minimum axial variation under these constraints can be seen in columns 3 and 4 of Table III. For contrast, we also consider what machine settings would lead to the greatest axial variation. The results of both of these optimizations can be seen in Fig. 9. The optimizations yield profiles that have the largest $I_{sat}$ values closest to the cathode, which is expected.

The prediction for an intermediate axial variation case is also seen in Fig. 9. The intermediate case was chosen as somewhere around halfway between the strongest and weakest case with a round index number (15 000, in this case). The parameters for intermediate case are also enumerated in Table III.

The predicted configurations with the run set flag on and top gas puff flag off (bolded in Table III) were then applied on the LAPD. The data collected, compared with the predictions, can be seen in Fig. 9.

For the optimized axial profiles, the absolute value of the $I_{sat}$ predictions compared to measurement does not agree. All of the predicted profiles have overlapping predictions (within the predicted error) at the region furthest from the cathode, but the measured values do not show that behavior. Although the mean $I_{sat}$ value was constrained to be greater than 7.5 mA/mm$^2$, the measured mean was 2 mA/mm$^2$ for the weakest case.

The important result is that the optimized LAPD settings, when implemented on the machine, do yield profiles with strong, intermediate, and weak axial variation. Although the minimum-$I_{sat}$ constraint was violated for the case of weakest axial variation case, this optimization would, nonetheless, be very useful for creating axial profiles of the desired shape.

There are three contributing factors to the mismatch of the ML-predicted values and the real measured values. First, the condition of the machine, such as the cathode emissivity or temperature or the downstream neutral pressure, is unquantified and cannot be compensated for in data preprocessing or in the model itself. Second, the calibration of the Langmuir probes could differ substantially between runs. The probes in the training data run sets (DR1 and DR2) were well-calibrated to each other within the run set but were not absolutely calibrated. The probes used for verifying the optimization were not calibrated. A rough calibration was performed by linearly extrapolating interferometer measurements and using triple probes (dotted lines on the right panel in Fig. 9). A configuration identical to DR2 run 10 was

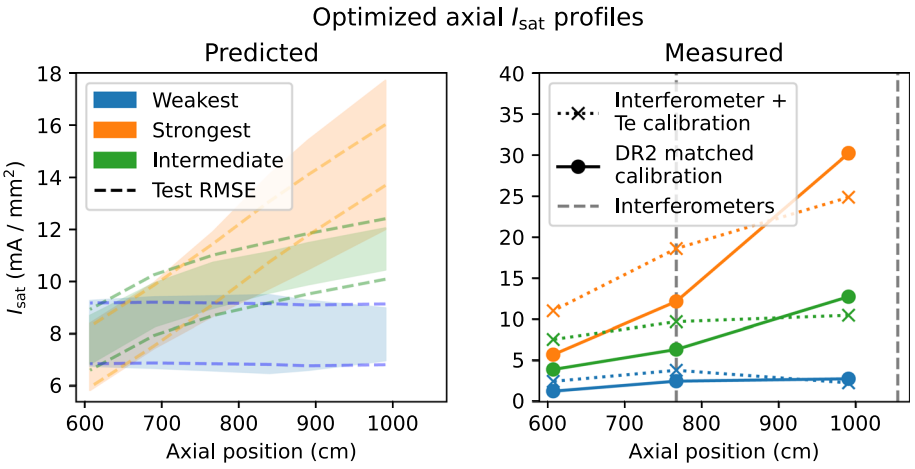


**FIG. 9.** Axial profiles, predicted and measured, for the optimized weakest (blue), intermediate (green), and strongest (orange) cases. (a) The shaded region covers the mean prediction ± one standard deviation, and the dashed lines are ± the median cross-validation RMSE values. (b) The measured $I_{sat}$ values are calibrated to DR2 run 10 (solid lines) or using triple probe Te measurements on the probe and linearly extrapolating the interferometer measurements (dotted lines). The absolute values disagree between the predicted and measured values, but axial trends are consistent with the optimization.

13 January 2026 06:45:00


© Author(s) 2025

also measured to simultaneously correct cathode condition and probe calibration (solid lines on the right panel in Fig. 9). Langmuir probe calibration is discussed further in Appendix C. Third, the original dataset may not have sufficient diversity to make accurate predictions on such a constrained optimization problem.

If this optimization were performed using the dataset instead of the model, the constrained search would encompass just 10 084 shots out of the 131 550 shots total in the training dataset or around 7.7%. Including the on-axis constraint reduces the number of shots down to 303 (270 in the training set) or 0.23% of all shots in the dataset. We conclude that this optimization of an arbitrary objective function, as done here, would be intractable using traditional, non-machine learning techniques because orders of magnitude more dataruns would need to be collected.

Optimization requires correctly learning the trends of all inputs and how they interact. In addition, as seen from the shot statistics, the model was trained on very few shots in the constrained input and output space. These two factors—the need for the model to learn all trends and the constrained search space—combine to make an incredibly difficult task that functions as a benchmark for the model. These factors were considered that it is not surprising that the model incorrectly predicts the absolute value. The uncertainty predicted by the model, though not well-calibrated, was, nonetheless, very large compared to the median test set RMSE. The model did predict the trends correctly; however, the optimized, measured profiles were strong, intermediate, or weak.

We did not check to see if the predicted optima were actually true optima: an approximation of the local derivative using a finite-difference technique would require much more run time on the LAPD than was available.

## VIII. DISCUSSION

### A. Key achievements

To the authors' knowledge, this work is one of the first instances that machine learning has been used to infer specific trends and optimize profiles in magnetized plasmas. Three examples of trend inference were shown in this paper: influence of magnetic geometry on plasma width, changes in the axial and radial profiles with changing discharge voltage, and the relationship of gas puff duration with axial gradient scale length. In addition, the axial profile was optimized by minimizing (or maximizing) the axial standard deviation. There is no other way of simultaneously uncovering many trends or finding optima without using an ML model trained on a diverse dataset. Traditionally, such studies would require extensive scans over grids to map the parameter space, but here, it was accomplished with a relatively small amount of data.

The trends inferred in this work, such as changing discharge voltages, gas puff durations, or mirror fields, would traditionally require a grid scan (varying one parameter at a time) in LAPD settings space. Here, instead, we are able to extract any trend covered by the training set with only a minimal amount of machine configurations sampled. Both data collection runs lacked absolute $I_{sat}$ calibration and had potential differences in cathode condition. Despite these issues, the model learned trends that were exploited via optimization.

In addition, this work demonstrates uncertainty quantification broken down into epistemic and aleatoric components by using ensembles and a negative-log likelihood loss. This uncertainty estimate is useful in gauging relative certainty between different predictions of the model, which increases confidence in the predictions of the model. In general, the total uncertainty predicted by the model increases when predictions are made outside the bounds of the training data.

Fundamentally, this model can predict $I_{sat}$ with uncertainty at any point in space covered by the training data. No other model currently exists that can perform this prediction. Traditionally, this capability would be possible only with a detailed theoretical study.

### B. Current limitations

This study would be dramatically improved by collecting more, diverse data. Only 44 of the 67 dataruns in this dataset had randomly sampled LAPD machine settings, which is very small compared to over 60 000 possible combinations. In addition, there are many other settings or parameters that were not changed in this study, such as gas puff timings, gas puff valve asymmetries, wall/limiter biasing, cathode heater settings, operation of the north cathode, and so on. The bounds of the inputs were also conservative; all settings in this study could be pushed higher or lower with a small amount of risk to LAPD operations. In addition, the placement of the probes could be further varied and placed outside the mirror cell, which would provide a more complete picture of LAPD plasmas, particularly axial effects.

Probe calibrations differed between the two training run sets (DR1 and DR2)—and a flag was introduced to distinguish between them—but despite this deficiency, combining the two run sets was shown to be advantageous for model performance. The condition of the cathode (e.g., electron emissivity and uniformity) also has a large impact on the measured $I_{sat}$. The improved model performance with the flag suggests that inconsistencies between dataruns could be compensated for using latent variables if a generative modeling approach is to be taken. At the very least, this model provides a way to benchmark these differences in machine state.

Concerning the model, hyperparameter tuning could be performed. In this study, a few extra percent in MSE performance is not meaningful, considering the limited dataset. Instead, we focused on the trends and insights that can be extracted from this model rather than simple predictive accuracy. There may also be regimes in hyperparameter space, where the uncertainty is better calibrated (perhaps using early stopping). Uncertainty estimation is important, even if the absolute uncertainty is not well-calibrated because it can provide a useful relative estimate as shown in this paper.

Trends such as the dependence of axial gradient on the gas puff duration (Fig. 8) or the effect discharge voltage on x–z profiles (Fig. 7), although intuitive, remain unverified. Verification of these trends would increase confidence of model predictions when setting LAPD parameters in future experiments.

### C. Future directions

The neural network architecture used here can readily scale to additional inputs and outputs; including time-series signals is the obvious next step. Integration of multiple diagnostics—perhaps starting with individual models before combining them—could enable inference of plasma parameters throughout the device volume. For example, combining triple probe electron temperature measurements with existing $I_{sat}$ data would allow density predictions anywhere in the plasma. This capability could enable *in situ* diagnostic cross-calibration (e.g., the Thomson scattering density measurement) and prediction of

13 January 2026 06:45:00


© Author(s) 2025

higher-order distribution moments like particle flux. The model could be further enhanced by incorporating physics constraints such as boundary conditions (e.g., zero $I_{sat}$ at the machine wall) or symmetries.

The problem presented here—learning time-averaged $I_{sat}$ trends—is fairly simple and required a relatively simple model. As demonstrated in this work, ML provides a way to explore regions of parameter space quickly and efficiently. Most physics studies on plasma devices (and fusion devices) are dedicated to a single particular problem, use grid scans, and are not useful to other experiments or campaigns. This work shows a way of using data and trends uncovered from other experimental studies. This work also demonstrates that random exploration can be a useful tool: the increased diversity of the aggregated data will generally benefit an ML model whether the experimenter discovers something new or not.

## IX. CONCLUSION

We demonstrate the first randomized experiments in a magnetized plasma experiment to train a neural network model. This learned model was then used to infer trends when changing field configuration, discharge voltage, or gas puff duration. This model was also used to optimize axial variation of $I_{sat}$ as measured by the standard deviation, which was validated in later experiments despite poor absolute error.

We strongly advocate that all ML-based analyses in plasma and fusion research should be validated and used to gain insight by inferring trends, as demonstrated here. This validation step is crucial for ensuring that ML models capture physically meaningful relationships, and the insights provided may provide direction for future research. We hope this is the first step toward automating plasma science.

## ACKNOWLEDGMENTS

The authors would like to thank Professor Walter Gekelman, Professor Christoph Niemann, Dr. Shreekrishna Tripathi, Dr. Steve Vincena, Dr. Pat Pribyl, Tom Look, Yhoshua Wug, Dr. Lukas Rovige, and Dr. Jia Han for insightful discussions and experimental support. This work was performed at the Basic Plasma Science Facility, which is a DOE Office of Science, FES collaborative user facility, and is funded by DOE (DE-FC02-07ER54918) and the National Science Foundation (NSF-PHY 1036140).

## AUTHOR DECLARATIONS

### Conflict of Interest

The authors have no conflicts to disclose.

### Author Contributions

Phil Travis: Conceptualization (equal); Data curation (lead); Formal analysis (lead); Investigation (lead); Methodology (lead); Project administration (lead); Resources (lead); Software (lead); Validation (lead); Visualization (lead); Writing – original draft (lead); Writing – review & editing (equal). Jacob Bortnik: Conceptualization (equal); Supervision (equal); Writing – review & editing (equal). Troy A. Carter: Conceptualization (equal); Funding acquisition (lead); Supervision (equal); Writing – review & editing (equal).

## DATA AVAILABILITY

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.15007853, Ref. 19.

## APPENDIX A: DATA BIAS

Despite the best efforts to randomize the machine configuration, imbalance in the dataset will be present because of the relatively small amount of samples for the given actuator space. The

13 January 2026 06:45:00

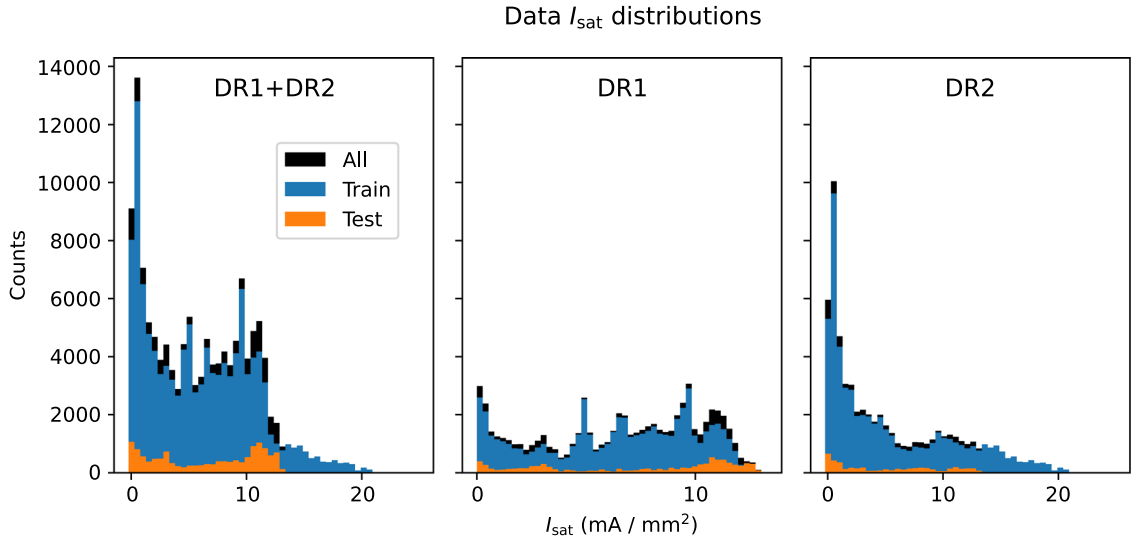


**FIG. 10.** Distribution of $I_{sat}$ signals. DR1 appears to have a more uniform distribution than DR2 does. Combining the two datasets results in many $I_{sat}$ examples near 0 mA/mm$^2$ and a sharp decrease in number of examples above 10 mA/mm$^2$. From these histograms, we expect or model to be biased toward fitting lower $I_{sat}$ values better and to perform badly in cases with very high $I_{sat}$ values.


© Author(s) 2025

**TABLE IV.** Data breakdown by class and dataset (percent).

| B source (G) | | | | B mirror (G) | | | | B midplane (G) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Train | Test | All | | Train | Test | All | | Train | Test | All |
| 500 | 4.77 | 0 | 4.29 | 250 | 4.30 | 8.41 | 4.72 | 250 | 8.25 | 21.01 | 9.55 |
| 750 | 3.34 | 12.61 | 4.29 | 500 | 30.49 | 8.41 | 28.23 | 500 | 43.80 | 8.41 | 40.19 |
| 1000 | 43.13 | 78.99 | 46.78 | 750 | 6.68 | 16.81 | 7.72 | 750 | 6.62 | 52.19 | 11.27 |
| 1250 | 12.59 | 0 | 11.30 | 1000 | 28.85 | 57.97 | 31.82 | 1000 | 26.36 | 5.78 | 24.26 |
| 1500 | 19.23 | 0 | 17.27 | 1250 | 3.34 | 4.20 | 3.43 | 1250 | 9.24 | 0 | 8.30 |
| 1750 | 1.91 | 0 | 1.71 | 1500 | 26.34 | 4.20 | 24.08 | 1500 | 5.73 | 12.61 | 6.43 |
| 2000 | 15.03 | 8.41 | 14.35 | | | | | | | | |
| **Gas puff voltage (V)** | | | | **Discharge voltage (V)** | | | | **Axial probe position (cm)** | | | |
| 70 | 12.11 | 16.81 | 12.59 | 70 | 12.22 | 8.41 | 11.83 | 639 | 12.48 | 8.41 | 12.06 |
| 75 | 6.68 | 0 | 6.00 | 80 | 5.25 | 0 | 4.72 | 828 | 17.07 | 36.28 | 19.03 |
| 80 | 11.46 | 8.41 | 11.15 | 90 | 2.86 | 8.41 | 3.43 | 859 | 12.48 | 8.41 | 12.06 |
| 82 | 41.49 | 57.97 | 43.17 | 100 | 3.34 | 8.41 | 3.86 | 895 | 33.01 | 30.10 | 32.71 |
| 85 | 14.13 | 0 | 12.69 | 110 | 8.77 | 0 | 7.87 | 1017 | 12.48 | 8.41 | 12.06 |
| 90 | 14.13 | 16.81 | 14.40 | 112 | 20.62 | 0 | 18.52 | 1145 | 12.48 | 8.41 | 12.06 |
| | | | | 120 | 3.82 | 8.41 | 4.29 | | | | |
| | | | | 130 | 0.95 | 0 | 0.86 | | | | |
| | | | | 140 | 2.86 | 8.41 | 3.43 | | | | |
| | | | | 150 | 39.30 | 57.97 | 41.20 | | | | |
| **Gas puff duration (ms)** | | | | **Vertical probe position (cm)** | | | | | | | |
| 38 | 94.27 | 91.59 | 94.00 | $\approx 0$ | 36.26 | 46.08 | 37.26 | | | | |
| $< 38$ | 5.73 | 8.41 | 6.00 | $\neq 0$ | 63.74 | 53.92 | 62.74 | | | | |

distribution of $I_{sat}$ signals can be seen in Fig. 10. The $I_{sat}$ distribution is clearly different for DR1 and DR2, with DR1 having a much flatter distribution. These distributions imply that if the model is constrained to sample from DR2 via the run set flag, then the model is expected to predict a lower $I_{sat}$ value in general. When predicting from the model in general, performance will likely be worse for $I_{sat}$ values $\gtrsim 11\,\mathrm{mA/mm^2}$.

The distribution of the selected machine settings for all the dataruns is enumerated in Table IV. Despite the randomization of the settings of 44 dataruns, the distribution is often uneven. This unevenness is exacerbated in the test set because that is a selection of 6 out of 67 dataruns. The remaining 23 nonrandom dataruns also contribute to the imbalance. For example, a source field of 1 kG and discharge voltage of 112 show up disproportionately in the dataset because data were collected at those settings, while other equipment was being adjusted or calibrated.

## APPENDIX B: DATA AND TRAINING PIPELINE VALIDATION

Multiple models were trained with varying depths and widths to verify that training loss decreases with increased model capacity. Doubling the layer width from 512 to 1024 moderately decreases the training loss; doubling the depth of the network from 4 to 8 layers has a larger impact. Increasing the width further to 2048 and depth to 12 layers has a dramatic impact on training loss, so this model and dataset are behaving nominally. The loss function for one of the NNs in the ensemble is seen in Fig. 11. The MSE decreases monotonically for the training and validation set but does not for the test set. The loss function can no longer be interpreted as a log-likelihood because of the effective per-example learning rate set by the $\beta$ term in the loss [Eq. (2)]. Note that early stopping (at around 8 epochs) would improve the test set loss, but the MSE would still be several factors higher than after 500 epochs. Early stopping was not explored in this study.

## APPENDIX C: EFFECT OF $I_{sat}$ CALIBRATION

The Langmuir probes did not seem to be behaving correctly when the optimization validation data were taken. The probes showed an *increasing* $I_{sat}$ profile when moving further from the cathode in the lowest gas puff condition, which is in direct disagreement with previous measurements and intuition. An example of this discrepancy can be seen in Fig. 12, where a run from the original testing set (specifically DR2 run 10) is duplicate. The probes for

13 January 2026 06:45:00


© Author(s) 2025

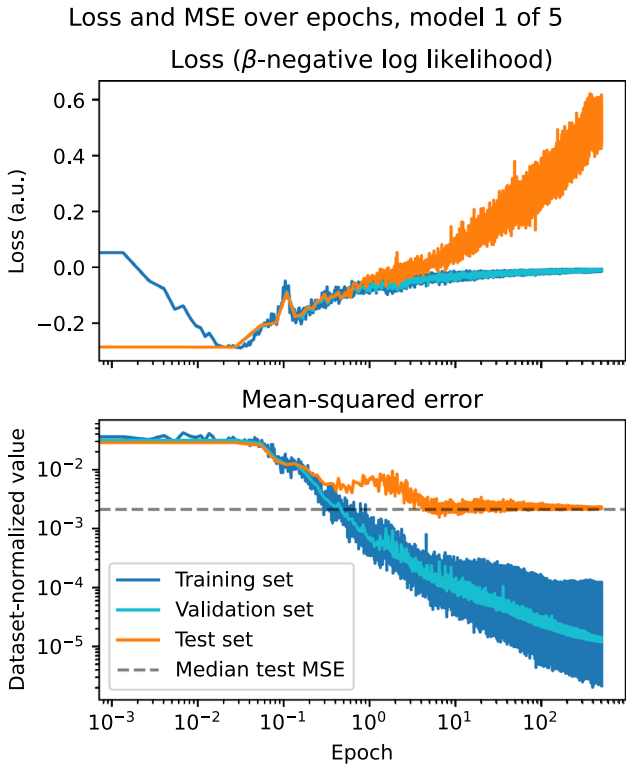


**FIG. 11.** The loss and MSE for the training, validation, and test sets over the entire training duration of 500 epochs. The inclusion of the $\beta$ term in the loss function—interpreted as a per-example learning rate—makes the loss function no longer interpretable in simple terms. The mean-squared error benefits from longer training for all sets.

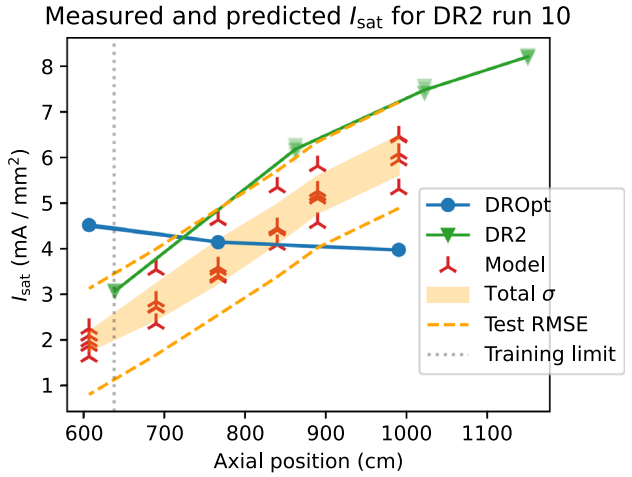


**FIG. 12.** Comparison of original DR2 profiles with the profiles from the optimization dataset (DROpt) for the same machine configuration. The $I_{sat}$ values in the DROpt dataset are not calibrated in this plot, indicating significant variation in probe calibration in this DROpt dataset.

the validation run can be either corrected for by assuming the 5 ms gas puff run has a flat axial profile or normalizing the probes to the DR2 run 10 axial profile. Calibrating the probes using the DR2 run 10 reference was the best way to go because it corrects for both probe discrepancies and changes in the condition (or emissivity) of the main cathode.

## REFERENCES


[1]J. Abbate, R. Conlin, and E. Kolemen, Nucl. Fusion **61**, 046027 (2021).
[2]A. Jalalvand, J. Abbate, R. Conlin, G. Verdoolaege, and E. Kolemen, IEEE Trans. Neural Networks Learn. Syst. **33**, 2630 (2022).
[3]I. Char, Y. Chung, J. Abbate, E. Kolemen, and J. Schneider, "Full shot predictions for the DIII-D Tokamak via deep recurrent networks," arXiv:2404.12416 (2024).
[4]C. Wan, Z. Yu, A. Pau, X. Liu, and J. Li, Nucl. Fusion **62**, 126060 (2022).
[5]J. Seo, Y.-S. Na, B. Kim, C. Lee, M. Park, S. Park, and Y. Lee, Nucl. Fusion **61**, 106010 (2021).
[6]J. Seo, Y.-S. Na, B. Kim, C. Lee, M. Park, S. Park, and Y. Lee, Nucl. Fusion **62**, 086049 (2022).
[7]Y. Fu, D. Eldon, K. Erickson, K. Kleijwegt, L. Lupin-Jimenez, M. D. Boyer, N. Eidietis, N. Barbour, O. Izacard, and E. Kolemen, Phys. Plasmas **27**, 022501 (2020).
[8]J. Dong, J. Li, Y. Ding, X. Zhang, N. Wang, D. Li, W. Yan, C. Shen, Y. He, X. Ren, and D. Xia, Plasma Sci. Technol. **23**(8), 085101 (2021).
[9]J. Kates-Harbeck, A. Svyatkovskiy, and W. Tang, Nature **568**, 526 (2019).
[10]C. Rea, K. Montes, K. Erickson, R. Granetz, and R. Tinguely, Nucl. Fusion **59**, 096016 (2019).
[11]J. Vega, A. Murari, S. Dormido-Canto, G. A. Rattá, M. Gelfusa, and JET Contributors, Nat. Phys. **18**, 741 (2022).
[12]J. Seo, S. Kim, A. Jalalvand, R. Conlin, A. Rothstein, J. Abbate, K. Erickson, J. Wai, R. Shousha, and E. Kolemen, Nature **626**, 746 (2024).
[13]A. Murari, E. Peluso, M. Lungaroni, R. Rossi, M. Gelfusa, and JET Contributors, Appl. Sci. **10**, 6683 (2020).
[14]A. Maris, C. Rea, A. Pau, W. Hu, B. Xiao, R. Granetz, E. Marmar, EUROfusion Tokamak Exploitation team, Alcator C-Mod team, ASDEX Upgrade team, DIII-D team, EAST team, and TCV team, "Correlation of the L-mode density limit with edge collisionality," arXiv:2406.18442 (2024).
[15]A. Pavone, A. Merlo, S. Kwak, and J. Svensson, Plasma Phys. Controlled Fusion **65**, 053001 (2023).
[16]A. Döpp, C. Eberle, S. Howard, F. Irshad, J. Lin, and M. Streeter, High Power Laser Sci. Eng. **11**, e55 (2023).
[17]R. Anirudh, R. Archibald, M. S. Asif, M. M. Becker, S. Benkadda, P.-T. Bremer, R. H. S. Budé, C. S. Chang, L. Chen, R. M. Churchill, J. Citrin, J. A. Gaffney, A. Gainaru, W. Gekelman, T. Gibbs, S. Hamaguchi, C. Hill, K. Humbird, S. Jalas, S. Kawaguchi, G.-H. Kim, M. Kirchen, S. Klasky, J. L. Kline, K. Krushelnick, B. Kustowski, G. Lapenta, W. Li, T. Ma, N. J. Mason, A. Mesbah, C. Michoski, T. Munson, I. Murakami, H. N. Najm, K. E. J. Olofsson, S. Park, J. L. Peterson, M. Probst, D. Pugmire, B. Sammuli, K. Sawlani, A. Scheinker, D. P. Schissel, R. J. Shalloo, J. Shinagawa, J. Seong, B. K. Spears, J. Tennyson, J. Thiagarajan, C. M. Ticoş, J. Trieschmann, J. V. Dijk, B. V. Essen, P. Ventzek, H. Wang, J. T. L. Wang, Z. Wang, K. Wende, X. Xu, H. Yamada, T. Yokoyama, and X. Zhang, IEEE Trans. Plasma Sci. **51**, 1750 (2023).
[18]G. A. Daly, J. E. Fieldsend, G. Hassall, and G. R. Tabor, Mach. Learn: Sci. Technol. **4**, 035035 (2023).
[19]P. Travis (2025). "physicistphil/lapd-isat-predict: 2025-3-11," Zenodo.
[20]C. Roach, M. Walters, R. Budny, F. Imbeaux, T. Fredian, M. Greenwald, J. Stillerman, D. Alexander, J. Carlsson, J. Cary, F. Ryter, J. Stober, P. Gohil, C. Greenfield, M. Murakami, G. Bracco, B. Esposito, M. Romanelli, V. Parail, P. Stubberfield, I. Voitsekhovitch, C. Brickley, A. Field, Y. Sakamoto, T. Fujita, T. Fukuda, N. Hayashi, G. Hogeweij, A. Chudnovskiy, N. Kinerva, C. Kessel, T. Aniel, G. Hoang, J. Ongena, E. Doyle, W. Houlberg, A. Polevoi, ITPA Confinement Database and Modelling Topical Group, and ITPA Transport Physics Topical Group, Nucl. Fusion **48**, 125001 (2008).
[21]S. Jackson, S. Khan, N. Cummings, J. Hodson, S. De Witt, S. Pamela, R. Akers, and J. Thiyagalingam, SoftwareX **27**, 101869 (2024).


13 January 2026 06:45:00


© Author(s) 2025

[22]See 10.57451/lhd.analyzed-data for "LHD experiment data repository."
[23]W. Gekelman, P. Pribyl, Z. Lucky, M. Drandell, D. Leneman, J. Maggs, S. Vincena, B. Van Compernolle, S. K. P. Tripathi, G. Morales, T. A. Carter, Y. Wang, and T. DeHaas, Rev. Sci. Instrum. **87**, 025105 (2016).
[24]Y. Qian, W. Gekelman, P. Pribyl, T. Sketchley, S. Tripathi, Z. Lucky, M. Drandell, S. Vincena, T. Look, P. Travis, T. Carter, G. Wan, M. Cattelan, G. Sabiston, A. Ottaviano, and R. Wirz, Rev. Sci. Instrum. **94**, 085104 (2023).
[25]P. Seetharaman, G. Wichern, B. Pardo, and J. L. Roux, "AutoClip: Adaptive gradient clipping for source separation networks," arXiv:2007.14469 (2020).
[26]D. Nix and A. Weigend, in *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)* (IEEE, Orlando, FL, 1994), pp. 55–60.
[27]B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," arXiv:1612.01474 (2017).
[28]M. Seitzer, A. Tavakoli, D. Antic, and G. Martius, "On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks," arXiv:2203.09168 (2022).
[29]M. Valdenegro-Toro and D. Saromo, arXiv:2204.09308 (2022).
[30]C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," arXiv:1706.04599 (2017).

13 January 2026 06:45:00


© Author(s) 2025