Machine learning methods for probabilistic locked-mode predictors in tokamak plasmas


ARTICLES YOU MAY BE INTERESTED IN

Statistical analysis of non-Maxwellian electron distribution functions measured with angularly resolved Thomson scattering
Physics of Plasmas 28, 082102 (2021); https://doi.org/10.1063/5.0041504

Turbulence transport in the solar corona: Theory, modeling, and Parker Solar Probe
Physics of Plasmas 28, 080501 (2021); https://doi.org/10.1063/5.0055692

Precession drift reversal and rapid transport of trapped energetic particles due to an energetic particle driven instability in the Large Helical Device
Physics of Plasmas 28, 080701 (2021); https://doi.org/10.1063/5.0059683
Machine learning methods for probabilistic locked-mode predictors in tokamak plasmas

Cihan Akçay,1,a) John M. Finn,2 Dylan P. Brennan,3 Thomas Burr,4 and Doğan M. Kurkcuoğlu5

AFFILIATIONS
1 General Atomics, 3550 General Atomics Drive, San Diego, California 92121, USA
2 Tibbar Plasma Technologies, Incorporated, 274 DP Rd., Los Alamos, New Mexico 87544, USA
3 Princeton University, Princeton, New Jersey 08544, USA
4 Los Alamos National Laboratory, Los Alamos, New Mexico 87544, USA
5 Fermilab, Batavia, Illinois 60510-5011, USA

a) Author to whom correspondence should be addressed: akcayc@fusion.gat.com

ABSTRACT
A rotating tokamak plasma can interact resonantly with the external helical magnetic perturbations, also known as error fields. This can lead to locking and then to disruptions. We leverage machine learning (ML) methods to predict the locking events. We use a coupled third-order nonlinear ordinary differential equation model to represent the interaction of the magnetic perturbation and the plasma rotation with the error field. This model is sufficient to describe qualitatively the locking and unlocking bifurcations. We explore using ML algorithms with the simulation data and experimental data, focusing on the methods that can be used with sparse datasets. These methods lead to the possibility of the avoidance of locking in real-time operations. We describe the operational space in terms of two control parameters: the magnitude of the error field and the rotation frequency associated with the momentum source that maintains the plasma rotation. The outcomes are quantified by order parameters that completely characterize the state, whether locked or unlocked. We use unsupervised ML methods to classify locked/unlocked states and note the usefulness of a certain normalization of the order parameters. Three supervised ML classifiers are used in suite to estimate the probability of locking in the region of control parameter space with hysteresis, i.e., the set of control parameters for which both locked and unlocked states can exist. The results show that a neural network gives the best estimate of the locking probability. An analogy of the present locking model with the van der Waals equation of state is also provided.

I. INTRODUCTION
The MHD response of a rotating toroidal plasma to non-axisymmetric error fields,1,2 a form of driven magnetic reconnection or driven tearing mode (TM) activity, can be important in tokamaks because it can lead to locked MHD modes, which can cause disruptions.3 A disruption in a tokamak involves a rapid loss of plasma confinement. Specifically, during a major disruption, the rapid loss of confinement can lead to a fast release of thermal energy. Such disruptions can cause surface melting of plasma-facing components and high electromagnetic loads.

Error fields can arise due to imperfections or alignment errors in external coils, or due to disadvantageously placed current feeds. The error fields can cause locking of a rotating plasma by exerting a large Maxwell torque,5 which leads to the amplification of the driven magnetic perturbations, with the magnetic perturbation and the decreased rotation synergistically amplifying each other. In many tokamak experiments, momentum is injected, e.g., by unbalanced neutral beams, to maintain the rotation to prevent locking, but this mechanism will be unavailable in ITER.6 Disruption avoidance, by means other than injecting momentum, requires understanding and forecasting the response of a plasma to error fields in the presence of plasma rotation.

Disruption forecasting/avoidance via machine learning (ML) techniques have been common practices in tokamak operations for nearly two decades. Artificial neural networks (NNs) were trained on the available diagnostic data as early as mid-2000s to detect disruptions.6–8 More recently, a disruption predictor based on deep learning, combining recurrent and convolutional neural networks, was developed by Kates-Harbeck et al.9 References 10–12 are further examples of deep learning methods developed on various devices for disruption prediction. Fu et al.13 applied the ensemble learning methods to DIII-D data to judge the “tearability” or “disruptivity” of their plasmas.
for real-time feedback control where their tearing mode predictor is used to control the neutral beam power in order to affect the plasma rotation. Their ML feedback detects the tearability in a probabilistic fashion after the onset of locking. Because of the danger posed by disruptions, any such detection of ITER disruptions should require an extremely low false negative (where “negative” means unlocked) rate, and achieving this is a serious challenge for ML algorithms.

This paper presents a proof-of-principle of using the machine learning classifiers (MLCs) to forecast mode-locking by calculating the probability of a rotating plasma to lock to a static error field. The appeal of this approach is the possibility of generating a meaningful, quantitative result, i.e., the locking probability, using sparse data. This capability is relevant to the realistic scenarios where the amounts of data that can be obtained from either experiments or high-fidelity simulations can be limited. The data sparsity is simulated here by having a single measurement (of several physical quantities) for each point in the parameter space and then going to coarser grids in the parameter space to determine where the accuracy of the MLCs breaks down. Another advantage of the present method is its probabilistic interpretation of locking. The need for this probability originates from the solutions of the ODE model that exhibit hysteretic and bifurcations between the low-rotation (locked) and high-rotation (unlocked) branches. This implies an inherent sensitivity to noise or perturbations in the initial conditions. In such hysteretic systems, knowing the probability of locking, conditional on the control parameters of the system, can be insightful because it quantifies the sensitivity to sudden changes in the plasma conditions and, thus, is more representative of the underlying phenomenology. In fact, this probability is a measure of the robustness of an unlocked state to a disturbance like a large sawtooth or a large edge-localized mode (ELM). These phenomena, among others, are known to trigger the neoclassical tearing modes (NTMs), which are nonlinearly destabilized TMs that arise due to significant bootstrap current. NTMs can be locked by eddy currents induced in resistive walls (RWs) in conjunction with the residual error fields.

The capability described here, when trained or calibrated on realtime experimental data, has the potential to become a forecasting tool for disruptions, which can be used in active feedback control. An example of such a capability is a probabilistic locked-mode predictor (LMP) that has been applied to Joint European Torus (JET) data, using support-vector machine (SVM) classifiers. The authors of Ref. state that their probabilistic LMP is very competitive in performance compared to other ML methods and is fully suitable for integration into the more general decision support system.

The third-order ODE model featured in this work has three dependent variables, $\psi/\psi_w$, $\theta$, and $\Omega$, which represent the magnitude of the magnetic field perturbation, its phase relative to that of the error field, and the toroidal plasma rotation frequency, all at the mode rational surface. These, in the time-asymptotic state, are the order parameters of the system. The initial condition for each order parameter is sampled randomly over a prescribed range. The control parameters of the model are the magnitude $\psi/\psi_w$ of the error field at the edge and the frequency $\Omega_0$ of a momentum source sustaining the plasma rotation. This rotation can applied externally in the case of unbalanced neutral beam injection (NBI) or intrinsically—and at low level—as expected in ITER. We choose a uniform grid with a nominal resolution of $200 \times 200$ in this 2D control space. Other parameters in the model are kept fixed.

The equations advancing $\psi/\psi_w$ and $\theta$ represent the linear growth of a weakly stable (intrinsic) tearing mode which is driven by the error field $\psi/\psi_w$ in the presence of rotation $\Omega$. Spontaneous stability is assured by setting the stability parameter of the mode $\Delta_1$, one of the fixed parameters of the model, to a small but negative number. For simplicity, the tearing mode is taken to be in the viscoresistive (VR) regime. The equation for $\Omega$, represents the competition between the quasilinear Maxwell torque, slowing the plasma down and the injected or intrinsic source driving $\Omega \rightarrow \Omega_0$. We examine two cases: one that is biased toward locking and one that is relatively unbiased or neutral. These two cases represent, respectively: a tokamak with a high level of fluctuations, such as sawteeth or ELMs, which may cause locking, and another more quiescent tokamak that operates in a safer regime. The chosen range in the initial conditions controls the amount of the said bias, i.e., predetermines the tendency of the system to lock.

The ODE model has time-asymptotic solutions that are steady state, consisting of locked and unlocked states. Loosely speaking, the locked states are characterized as having large steady-state values of $\psi/\psi_w$, $\theta$, and $\Omega$, while the unlocked solutions cluster nearby. This merging is assured by normalizing the order parameters to span $[0, 1]$. Under this normalization, the phase $\theta$ is redundant. The normalization of the remaining two-order parameters onto a unit square $|\psi/\psi_w, \Omega|$ redistributes more than 99% of the solutions into two tight clusters, representing the locked and unlocked states of the plasma. The locked solutions cluster around $(\psi/\psi_w, \Omega) \approx (1, 0)$, while the unlocked solutions cluster around $(\psi/\psi_w, \Omega) \approx (0, 1)$.

The steady-state solutions of this model can be found analytically by solving a cubic equation in $\Omega$, representing the torque balance on the plasma at the rational surface (see Refs. 1 and 2). These solutions show the presence of a hysteretic region $(\psi/\psi_w, \Omega)$, where the cubic equation of torque balance has three real roots. Two of these three roots, the attractors of the system, represent the locked and unlocked states of the plasma. The hysteretic region is separated in control parameter space from two other regions, where the cubic equation has strictly one real root. In one of these two regions, the states are locked and in the other unlocked. These two regions are separated from the hysteretic region by analytically defined boundary segments that we call the bifurcation boundary segments. These two segments come together at a critical point (CP) below which smooth transitions are seen to occur between the locked and unlocked states. This is analogous to the CP encountered in liquid-to-gas phase transitions, as described by the van der Waals (VDW) equation of state (EOS).

For the ODE system presented here, the hysteresis described in the preceding paragraph appears as a sizable region in the 2D control space of $(\psi/\psi_w, \Omega)$, featuring a strong mixing of locked and unlocked solutions, often neighboring each other. This mixing implies a
sensitivity of the ODE solutions in the hysteretic region to the randomly chosen initial conditions, whose range in fact influences the density of locked and unlocked states within the hysteretic region. This implies, for a different realization of the random initial conditions, that a point identified previously as locked can jump to an unlocked state. In our model, this tendency to "jump" is quantified in terms of the locking probability, which is a measure of the size of the domain of attraction of the locked root on the slow-rotation branch, given the domain of initial guesses. Or in the context of tokamak physics, the said probability is a measure of the robustness of an unlocked state to a disturbance like a large sawtooth or a big ELM.

For applying ML methods to the problem of locking probabilities, we follow a twofold approach. The first stage aims to classify unambiguously each solution of the ODE system as locked or unlocked in the space of the normalized order parameters, without human input. This is accomplished by subjecting the input vector comprising all of the values of the normalized order parameters to K-means clustering (KMC), which is an unsupervised classification algorithm. KMC is a geometric method that makes a "hard" assignment to each sample. The classification results have also been benchmarked with Gaussian mixture models, a probabilistic method that fits anisotropic Gaussians to the data.

The second stage involves training a series of supervised classifiers to calculate the conditional (or posterior) probability of locking \( p(L|\tilde{\psi}_w, \Omega_b) \) in the control space \((\tilde{\psi}_w, \Omega_b)\). The locking probability satisfies \( 0 \leq p(L|\tilde{\psi}_w, \Omega_b) \leq 1 \) within the hysteretic region, and an area where \( p(L|\tilde{\psi}_w, \Omega_b) \) is close to one is an area susceptible to locking.

The first step in calculating \( p(L|\tilde{\psi}_w, \Omega_b) \) entails switching from the 2D space of normalized order parameters \((\tilde{\psi}_w, \Omega_b)\) to the 2D space of the control parameters \((\psi_w, \Omega_b)\). Next, a pre-processing step is carried out where the binary class labels (unlocked and locked) emerging from KMC are assigned to each point in the control space as the targets of the supervised training. This converts the control space phase diagrams into a speckle plot of 0’s and 1’s in the hysteretic region. This binary map is then fed into the supervised classifiers which convert it into a smooth map of locking probabilities. The input vector space of the training consists of the values of the control parameters \((\psi_w, \Omega_b)\) at each grid point in the control space. Thus, we have \(200 \times 200\) samples with two features \((\psi_w, \Omega_b)\) each and 40,000 binary target values that enter the cost function of the classifiers.

The three algorithms chosen to estimate \( p(L|\tilde{\psi}_w, \Omega_b) \) are support-vector machines (SVMs), logistic regression (LR), and a fully connected feed-forward neural network (NN). The support-vector machine uses a nonlinear kernel in the form of a radial basis function (RBF), and the logistic regression uses a rational function or Padé approximants for the argument of its sigmoid function. These nonlinearities are introduced to deal with difficulties associated with shape of the hysteretic region, especially near the CP. The neural network contains four hidden layers, each consisting of hundreds of nodes. Note that SVM has already been used in this fashion and trained on inductance and mode amplitude data from jet discharges, to construct a probabilistic LMPT.

To assess the accuracy of these estimates of \( p(L|\tilde{\psi}_w, \Omega_b) \), a "ground truth" (GT) locking probability is also computed by a direct, but quite expensive, Monte Carlo method, where the ODEs at each point in control space are integrated for \( N_i = 10,000\) different random initials conditions. The locking probability at each point is then given by the ratio of the tally of the locked points to \( N_i\). Note that more realistic systems will not grant us the luxury of knowing the GT. The reader should also bear in mind the cost of the GT calculation for the ODE system, which takes approximately a week on several dozen processors, in contrast to any of the three MLCs that take approximately a minute each on a single processor. The orders-of-magnitude gain in the computation time makes it possible for one or several of these MLCs to be employed in real-time operations.

Of the three MLCs that have been used, the NN produces the most accurate probabilities. Trials for optimizing the NN suggest that the most accurate results are obtained for an architecture with four hidden layers, each containing 50–200 nodes. The other two methods, RBF SVM and rational LR do not perform as well in capturing the finer features near the CP and along the boundary segments of the hysteretic region.

This manuscript is organized as follows: The analysis and solutions of the ODE model are presented in Sec. II. The results from the MLCs for the biased case are presented in Sec. III, with the unsupervised classification results appearing in Sec. III B. The locking probabilities calculated by the supervised classifiers as well as the GT probability appear in Sec. III C. Section IV mirrors the structure of Sec. III for the neutral case. A summary and discussion of the results as well as future directions are provided in Sec. V. Appendix A presents an analogy between the equations for the locking-unlocking bifurcation diagram and those related to the equation of state for a VDW gas. Appendix B describes the unsupervised and supervised MLCs that are used in this work.

II. THE MODEL: COUPLED ODE SYSTEM OF LOCKING/UNLOCKING BIFURCATIONS

The coupled ODE system derived here represents the interaction of the magnetic perturbation with the error field in the presence of plasma rotation. This system is sufficient to describe the well-known characteristic bifurcations involving the locking and unlocking of a rotating plasma. The advantage the simple model provides is that it can produce tens of thousands of solutions in a few minutes that can then be used to train the ML algorithms. The lessons learned from such a process might be helpful in designing an ML approach for getting the best use out of a simulation or experimental campaign, where the data will necessarily be more sparse.

We begin with a linear response model for the complex-valued reconnected flux \( \tilde{\psi}_w \), as described in Refs. 20 and 21. For simplicity, we assume that the tearing layer is in the constant-\(\psi\) viscoresistive (VR) regime, without the complications of pressure gradient and curvature in the tearing layer.\(^{22}\) The time-dependent linear homogeneous ODE for \( \tilde{\psi}_w \) takes on the following form:

\[
\frac{d}{dt} \tilde{\psi}_w + i \Omega_i - \Delta_i \tilde{\psi}_w = I_{121} \tilde{\psi}_w,
\]

where \( \Omega_i \) (real-valued) is the instantaneous plasma rotation at the rational surface, \( I_{121} \) is an inductance-like factor related to the geometry and the current density profile and \( \Delta_i \) is the intrinsic stability parameter.

The angular momentum of the plasma in the tearing layer at the rational surface is subject to a Maxwell torque as well as a restoring force due to the injected or intrinsic momentum source \( \Omega_{bi} \).\(^{19}\)
\[ I \frac{d}{dt} \Omega = -\text{Im} \left( \psi^* \frac{d\psi}{dt} \right) + \mu (\Omega_0 - \Omega), \]  

where \( I \) is the moment of inertia of the tearing layer near the rational surface and the first term on the right represents the Maxwell torque, with the jump in the derivative of the reconnected flux at the rational surface (proportional to the current density there) given by \( \psi^* = \Delta_i \psi + i_3 |\psi| \). The parameter \( \mu \) represents a physical drag term, related to the viscosity of the plasma.\(^1\) We write Eqs. (3)–(5) to zero, which yields

\[ I \frac{d}{dt} \Omega = i_1 \psi^* \psi \sin \theta_1 + \mu (\Omega_0 - \Omega). \]  

Splitting Eq. (1) (multiplied by \( e^{-it} \)) into its real and imaginary parts yields the coupled ODE system of Eq. (3) with

\[ \frac{d}{dt} \psi = \Delta_1 \psi_1 + \frac{\psi}{\psi^*} \cos \theta_1, \]  

\[ \frac{d}{dt} \psi_1 = -\Omega - \frac{\psi}{\psi^*} \sin \theta_1, \]  

where we choose \( I = 1 \) and \( \psi^* \rightarrow i_1 \psi^* \) for convenience.\(^23\) We normalize the time to a nominal tearing time and fix \( \Delta_1 = -0.1 \) and \( \mu = 0.01 \). In these units, \( 1/\mu \) is the timescale for momentum transport without the electromagnetic torques. The choice of \( \Delta_1 \) indicates a weakly stable spontaneous tearing mode. The value of \( \mu \) is chosen such that the growth of the mode (in response to \( \psi^* \)) with this value of \( \Delta_1 \), proceeds on a faster timescale than \( 1/\mu \); the ordering \( \mu \ll |\Delta_1| \) is consistent with the observed modes in the experiments.\(^1,2,3,4,5\) We refer to the trio \((\psi, \psi^*, \Omega)\) in the time-asymptotic state as the order parameters, characterizing the state of the system.

The system of equations in Eqs. (3)–(5) always leads to a steady time asymptotic state. Note that when \( \psi \) is zero, the system converges to \( \psi_1 = 0 \) and \( \Omega = \Omega_0 \) for \( t \rightarrow \infty \). For large \( \Omega_0 \) and moderate \( \psi^* \), the system converges to a state with \( \Omega \approx \Omega_0 \) and \( \psi^* \) small (unlocked), while for more moderate \( \Omega_0 \) and large \( \psi^* \) the system may converge to a state with \( \Omega \approx \Omega_0 \) and \( \psi^* \) large (locked).

The data required for training the MLCs in Secs. III and IV are obtained by integrating Eqs. (3)–(5). However, some insights can be gleaned by analytically examining the time-asymptotic solutions of the system, which are steady-state and found by setting the left side of Eqs. (3)–(5) to zero, which yields

\[ \bar{\psi} = \frac{\psi^*}{\sqrt{\Delta_1^2 + \Omega_0^2}}, \]  

\[ \tan \theta_1 = \frac{\Omega_0}{\Delta_1}, \]  

\[ \mu (\Omega_0 - \Omega_0) = -\Omega_0 \psi^* \frac{\psi^*}{\Delta_1^2 + \Omega_0^2}, \]

Equation (8) shows the possibility of either one (real) root for \( \Omega_0 \) or three, and Eq. (7) shows \( -\pi/2 < \theta_1 < 0 \) and \( 0 < \Omega_0 < \Omega_0 \). For small \( \Omega_0 \), specifically \( \Omega_0 \ll |\Delta_1| \), we have \( \psi \rightarrow \psi^* \left/ |\Delta_1| \right. \) and \( \theta_1 \rightarrow 0 \), i.e., the maximum response to the error field. For large \( \Omega_0 \), specifically \( |\Delta_1| \ll \Omega_0 < \Omega_0 \), we have \( \psi \rightarrow \psi^* \Omega_0 \Omega_0 \approx \Omega_0 \), and \( \tan \theta_1 \approx -\Omega_0 / |\Delta_1| < -1 \), leading to \( \theta_1 \rightarrow -\pi/2 \).

Equation (8), representing the well-known torque balance in steady-state,\(^1\) can be recast as a cubic equation for \( \Omega_0 \):

\[ \Omega_0^3 - \Omega_0 \Omega_0^2 + \left( \frac{\Delta_1^2}{\mu} + \frac{\psi^*}{\mu} \right) \Omega_0 - \Omega_0 \Delta_1 = 0. \]

The locked/unlocked phases of a rotating plasma interacting with an error field have traditionally been described in terms of the solutions of this equation, where, loosely, \( \Omega_0 / \Omega_0 \ll 1 \) means locked and \( \Omega_0 / \Omega_0 \approx 1 \) means unlocked. This cubic equation has real coefficients, and the conditions to have three real roots or one real root are well-known. (These equations can be alternately expressed in terms of a cubic equation for \( \psi^* \), with the same condition for real roots; and a third equation for \( \theta_1 \) can also be solved.)

As we shall discuss further, in the case of a single real root, the distinction between locked and unlocked states is subject to interpretation. In the case with three real roots for \( \Omega_0 \), the largest root corresponds to the unlocked state and the smallest root to the locked state. These two roots are the attractors while the middle root is not, and is thus physically unobservable, i.e., "forbidden," as the initial conditions near the middle root converge to one of the two other roots. This region in control parameter space \((\psi^*, \Omega_0)\) where there are two stable equilibrium states is the region of hysteresis. At the boundary of this hysteretic region, an abrupt jump (a bifurcation) can occur between the locked and unlocked phases. This is analogous to the jump that an order parameter undergoes during a first-order phase transition. In fact, the system described by Eq. (9) is analogous to the van der Waals equation of state for a non-ideal gas. This equation is also cubic in the number density of the gas \( n \), which is the order parameter of the system. We elaborate on this analogy in Appendix A.

We integrate Eqs. (3)–(5) over \( t = [0, 10^4] \) for a uniform \( 200 \times 200 \) grid of the control parameters \((\psi^*, \Omega_0)\), spanning \( \psi^* = [0, 0.2] \) and \( \Omega_0 = [0, 0.5] \). To deal with the issue of the probability of locking, especially in the hysteretic regime, a single initial condition \((\psi_0, \theta_0, \Omega_0(0))\) is chosen randomly for each pair of control parameters \((\psi^*, \Omega_0)\). The size of the space of initial conditions affects the system’s tendency to lock, with some initial conditions biased toward locking and others less biased, as explained in Secs. III and IV.

III. THE LOCKING-BIASED CASE

A. The solutions of the ODEs

To bias the solution inside the hysteretic regime toward locking the range for \( \psi_0(0) \) is augmented to \([0, 20]\), while that for \( \Omega_0(0) \) is reduced to \([0, 1] \). The initial condition for the phase \( \theta_0(0) \) is always sampled over \([0, 2\pi]\), regardless of the locking bias.

The ODE solutions solved on the \( 200 \times 200 \) grid of control parameters are shown in Fig. 1(a), where two of the three-order parameters \((\psi, \Omega_0)\) are plotted as a scatterplot, while \( \theta_0 \) is ignored for convenience at this stage. These two-order parameters form an “L”-shaped pattern in the solution space, where the vertical “arm”
corresponds roughly to unlocked solutions and the horizontal arm to locked solutions, with the density of points not noticeably different anywhere in the L-shape.

As argued in the last section, the locked states have $\tilde{\psi}_t \approx \tilde{\psi}_w/|\Delta|$, $\Omega_t \approx 0$, and $\theta_t \approx 0$, whereas the unlocked states have $\tilde{\psi}_t \approx 0$, $\Omega_t \approx \Omega_m$, and $\theta_t \approx -\pi/2$. Motivated by these observations, we define the normalized order parameters $\tilde{\psi}_w \equiv |\Delta|\tilde{\psi}_w/\tilde{\psi}_w$ and $\Omega_w \equiv \Omega_t/\Omega_m$, and discuss the advantage of using them rather than the raw order parameters. Because we have $-\pi/2 < \theta_t < 0$, it is not necessary to scale $\theta_t$ further. In terms of these normalized parameters, the locked states have $\tilde{\psi}_u \approx 1$, $\Omega_u \approx 0$ and $\theta_u \approx 0$, whereas the unlocked states have $\tilde{\psi}_u \approx 0$, $\Omega_u \approx 1$, and $\theta_u \approx -\pi/2$. From Eqs. (6) and (7), we have $\tilde{\psi}_u = \cos \theta_t$, a direct relationship independent of the control parameters $(\tilde{\psi}_w, \Omega_u)$. For this reason, when normalized order parameters are used, the variable $\theta_t$ becomes redundant. In contrast, the relation between $\tilde{\psi}_u$ and $\Omega_u$ depends on the control parameter $\Omega_m$. It is important to note that the redundancy of $\theta_t$ also reduces the dimensions of the order parameter space from 3D to 2D. In terms of the remaining normalized order parameters, it becomes far less ambiguous what constitutes a locked or an unlocked state of the plasma. On the unit square in $(\tilde{\psi}_u, \Omega_u)$ shown in Fig. 1(b), the unlocked states are clustered near the top left corner, and the locked states near the bottom right corner; more than 99% of the solutions fall into these two clusters.

There is a small population of non-clustering points that appears in the center of Fig. 1(b). However, the concentration of clustering points near the two corners is several orders of magnitude greater than that of these non-clustering points, as evidenced by the darkening of the color (indicating the density of points) in Fig. 1(b). As we will see, this clustering due to the choice of normalization aids in classification greatly. Also, with this normalization choice, with parameters on the unit square, it is natural to use a Euclidean metric tensor for distances, used in the unsupervised classification of Sec. III B.

The crucial role played by normalization of the order parameters is further illustrated in Figs. 2(a) and 2(b). Figure 2(a) shows the 1D cumulative distribution functions (CDFs) of the time-asymptotic raw order parameters $\tilde{\psi}_t$ (dashed green) and $\Omega_t$ (solid blue), while Fig. 2(b) shows the 1D CDF of the time-asymptotic normalized order parameters $\tilde{\psi}_w$ (dashed green) and $\Omega_w$ (solid blue). The figures also include the probability density function (derivative of the CDF) shown as histograms in the insets, with the blue bars labeling $\Omega_w$ and $\Omega_m$ and the green bars labeling $\tilde{\psi}_t$ and $\tilde{\psi}_u$. The CDFs of the raw order parameters in Fig. 2(a) show a tendency toward clustering around zero for both $\Omega_t$ and $\tilde{\psi}_t$, but vary noticeably for intermediate values of these parameters. The jumps in $\Omega_w$ and $\tilde{\psi}_w$ correspond roughly to the horizontal and vertical wings of the scattered solutions shown in Fig. 1(a), respectively.

The cumulative distribution functions are re-plotted as a function $\tilde{\psi}_w$ and $\Omega_w$ in Fig. 2(b) to demonstrate the dramatic effect of our choice of normalization of the solutions. The clustering behavior, described above for Fig. 1(b), re-emerges here as well. The CDF traces now evince two areas of fast increase: one around $\Omega_w \approx 0.05(\tilde{\psi}_w \approx 0.9)$ and another around $\Omega_w \approx 0.8(\tilde{\psi}_w \approx 0.1)$. There is very little increase in the CDF for intermediate values, showing that the number of intermediate solutions between the locked and unlocked states is very few. Similarly, the histograms shown in the upper inset for $\Omega_w$ and lower inset for $\tilde{\psi}_w$ feature two spikes located at $\Omega_w \approx 0$ ($\tilde{\psi}_w \approx 1$) and $\Omega_w \approx 1$ ($\tilde{\psi}_w \approx 0$), respectively. Thus, the normalization makes it clear that the majority of the population clusters into two groups: one rotating very slowly with the magnetic perturbation $\tilde{\psi}_t \approx \tilde{\psi}_w/|\Delta|$ (locked) and the other rotating fast at nearly $\Omega_w$ with $\tilde{\psi}_t \approx 0$ (unlocked). The minimum of the histograms, corresponding to the flattest part of the CDF, suggest the possibility of a threshold for locking in terms of a single-order parameter. These minima indicate locking for $\Omega_w \sim 0.5$ or $\tilde{\psi}_u \sim 0.5$. Both criteria result in approximately same rate of locking: 56% of the total population is locked for this particular range of control parameters. We will compare the locking criteria based on the CDFs to those from the classification results in Sec. III B.
Phase diagrams, i.e., the time-asymptotic solutions as a function of the control parameters \( (\tilde{\psi}_w, \Omega) \) on a 200 \( \times \) 200 uniform grid, are shown in Fig. 3. Here, (a) and (b) show color plots of the time asymptotic normalized rotation frequency \( \tilde{X}_n \) and mode amplitude \( \tilde{\psi}_w \), respectively. As discussed, \( \tilde{\psi}_t \) is redundant and therefore omitted. The phase diagrams show three distinct regions: two that are solely occupied with solid red or blue pixels and a third region, the hysteretic region, with an intermixing of the blue and red pixels. The regions of solid color correspond to the cubic equation having one real root. The solid blue represents the cases with \( \Omega \approx 1, \tilde{\psi}_t/\tilde{\psi}_w \ll 1 \) in Figs. 3(a) and 3(b), i.e., unlocked solutions. These are the solutions that are hardly influenced by the error field. The solid red represents the cases with \( \Omega \approx 1, \tilde{\psi}_t/\tilde{\psi}_w \approx 1 \) in Figs. 3(a) and 3(b), i.e., the locked solutions. These are the solutions that are strongly influenced by the error field. The third region, squeezed between the solid blue and red region, where the pixels with a spectrum of colors ranging from red to blue...
are intermixed, corresponds to the hysteretic region. This is where the analytic solution of the torque balance, Eq. (9), yields three real roots. In this region, there can be two nearby points in the control space, randomly initialized with two different values of the order parameters, with one going to a locked state and the other going to an unlocked state, and hence the “speckling” pattern.

The speckles apparent to the eye in Fig. 3 depend on the particular value of \( (\psi_w, \Omega_0) \) as well as on the range of the initial conditions of the randomly chosen order parameters and the actual realization of these random variables. On the other hand, the density of speckles is insensitive to the different realization of the random initial conditions so long as the range of the initial conditions is kept the same from one execution of the ODE system to the other. Thus, the location of the individual speckles inside the hysteretic region in the control space corresponds to a microscopic quantity, sensitive to fluctuations, while the density of the speckles therein corresponds to a more robust macroscopic quantity. The MLCs used in this work estimate this density.

Above the upper black curves appearing in Fig. 3, the cubic has only unlocked states; below the lower black curve, the cubic has only locked states. These curves bound the hysteretic region. These two bifurcation boundary segments merge at the critical point (CP) of the system, situated at \( (\psi_w, \Omega_0) = (0.0283, 0.520) \). Near the CP, the two curves meet at a tangency, \( \delta \Omega_0 \propto (\psi_w)^{1/2} \). This point will be of importance when we deal with supervised classification to calculate the probability of locking in this region. Below and to the left of the CP, one state of the plasma can smoothly transition into another without undergoing bifurcations, much like continuously transitioning between the gaseous and liquid phases of a VDW gas (Appendix A). Note the difference in the color pattern below the CP for the two plots. The widely scattered points between the two clusters in Fig. 1(b) correspond exactly to these smoothly transitioning points. These smoothly transitioning points can be visually eliminated by re-plotting only the solutions that are generated from the control parameters that lie above the CP, in this case specifically for \( (\psi_w, \Omega_0) \geq (0.05, 1.0) \), which is marked by the gold stars in Fig. 3. This filter produces the two well-separated clusters shown in Fig. 4. Another related point is that the states in the vicinity of the CP usually have small values of \( \psi_w \) and, therefore, the occurrence of disruptions due to locking around the CP is not such a serious concern.

The dashed horizontal white lines in Figs. 3(a) and 3(b) represent a 1D cut at \( \Omega_0 = 3.0 \) along which we further illustrate the hysteresis inherent in our ODE system. See Fig. 5, which shows the bifurcation in the solutions \( \Omega_n \) and \( \psi_w \) as a function of \( \psi_w \) for \( \Omega_0 = 3.0 \). The solid black \( (\Omega_n) \) and red \( (\psi_w) \) curves are the analytic curves that are extracted by solving the time-asymptotic solutions given in Eqs. (6)–(8). The upper branch of the red curve and the lower branch of the black curve correspond, respectively, to locked states; the lower branch of the red curve and the upper branch of the black curve correspond to the unlocked states. The “forbidden” solutions in the hysteretic region are represented by the dashed portions of the traces.

In Secs. III A, III B, and IV B that follow, we employ a two-step strategy, with each stage leveraging machine-learning classifiers (MLCs) in a particular way. Stage 1 entails an unsupervised classification in terms of the normalized order parameters of the ODE as locked or unlocked. This step uses K-means clustering (KMC), described in Appendix B. This gives us an unambiguous locking criterion in the normalized order parameter space, \( (\psi'_w, \Omega_n) \). As discussed above, the normalization of the solution space onto a unit square facilitates the task of classification. In more complex systems, e.g., in numerical simulations, it may not be possible to do such a scaling to aid the classification process; in such cases, it will be important to use unsupervised classifiers to determine the class of each solution. In stage 2, we use the supervised MLCs to calculate the conditional probability of locking within the hysteretic region in the control parameter space. For this process, we use the control parameters \( (\psi'_w, \Omega_0) \) as the input vectors instead of the normalized order parameters, and the sample classes labeled according to KMC as the target values, to train the MLCs.
This ML-based probability is then compared to a Monte Carlo-type calculation of probability, described in Sec. III C, which we call the ground truth probability.

B. Classifying the ODE solutions as locked or unlocked

The main classification result of this section is based on using the normalized order parameters, $\psi_n$ and $\Omega_n$, as the input for the unsupervised classification. We refer to this procedure as a 2D classification. The classification is also repeated in 1D, using either $\Omega_n$ or $\psi_n$ as the single input. The results show the 2D classification to be only marginally better than the 1D classification based on either $\Omega_n$ or $\psi_n$ alone. This is another indication of the fact that the normalizations [compare Figs. 1(a) and 1(b)] facilitate this task of unsupervised learning greatly, by redistributing most of the samples into the two aforementioned clusters.

We use K-means clustering $^{29,30}$ (KMC) to classify each solution of the ODEs as locked or unlocked in an unsupervised fashion, i.e., without human input. The details of this algorithm are described in Appendix B. Briefly, KMC geometrically partitions $N$ sample points into $K$ clusters by computing the centroid or cluster center of mass and then assigning each sample to the cluster with the nearest centroid. (This classification in terms of distance shows the importance of picking an appropriate metric in the order parameter space.) KMC is a technique that is used in image segmentation and compression.

The results of the classification by K-means clustering trained on the values of $(\psi_n, \Omega_n)$ are displayed in Fig. 6. The blue dots represent unlocked cases and the red crosses their locked counterparts. The centroids, marked by two large yellow diamonds, are located at $(\psi_n, \Omega_n) = (0.98, 0.01)$ and $(0.05, 0.94)$ for the locked and unlocked states, respectively. Since the two cluster centers are located in the upper left and lower right corners, KMC classification in this case produces almost a perfectly diagonal decision boundary (dashed line), separating the locked and unlocked populations. This means any ODE solution with $\psi_n \geq \Omega_n$ corresponds to a locked phase of the plasma. This criterion indicates 55% of the population to be locked, in good agreement with the results based on the minimum derivative of the CDFs shown in Fig. 2(b).

For the classifications based on a single input consisting of the values of $\psi_n$ or $\Omega_n$ as the training input, KMC splits the population into two classes by a vertical (respectively, horizontal) line near the middle of the range for each order parameter: the classification based on $\psi_n$ indicates locking for $\psi_n \geq 0.51$ and classification based on $\Omega_n$ indicates locking for $\Omega_n \leq 0.48$. These locking thresholds are in good agreement with the results based on the minimum derivative of the cumulative distribution functions shown in Fig. 2(b). The two individual criteria in this case lead nearly to the same number of locked states, 56%, as the decision boundary based on both $(\psi_n, \Omega_n)$. The reason for such a close agreement between the locking tallies in spite of three seemingly different criteria for locking is the fact that the widely scattered data in Fig. 6 are very sparse.

The KMC results discussed here have been found to agree well with another unsupervised classifier called Gaussian-mixture models (GMM), which fit K-many anisotropic Gaussians to a dataset, where each Gaussian represents one cluster. Unlike KMC, which is a geometric method that makes hard assignments, GMM are a probabilistic method that makes soft assignments to the data points, and can account for the cluster density. In fact, the cluster density is one of the parameters that the GMM algorithm optimizes. However, since the data in use here have nearly the same density for both clusters, the classification results with GMM are very similar to those with KMC. Thus, we proceed with the KMC classification results for the remainder of the manuscript.

C. Calculating locking probabilities via supervised classification

In stage 2, we use the supervised classifiers described in Appendix B.2 to calculate the probability of locking $p_L \equiv p_L(L|\psi_n, \Omega_n)$, conditional on the control parameters of our model in the hysteretic region of the control space. This involves switching from the 2D space of normalized order parameters $(\psi_n, \Omega_n)$ of the previous section to the 2D space of the control parameters $(\psi_n, \Omega_n)$, appearing as the $200 \times 200$ uniform grid in Figs. 3(a) and 3(b). We then assign the values from $t \in \{0, 1\}$ (unlocked, locked, respectively), to each control space point $(\psi_n, \Omega_n)$ according to the classification of Sec. III B. This conversion is done by imposing the KMC classification on all of the $200 \times 200$ points on this grid whereby each point with $\psi_n > \Omega_n$ is classified as locked (red) and each point with $\psi_n \leq \Omega_n$ is classified as unlocked (blue). Figure 7 shows the result of this procedure: a speckling pattern within the hysteretic region emerges. It is this binary map that is fed into the supervised classifiers in this section, which convert it into a smooth map of locking probabilities $p_L$.

The justification for using a single speckling diagram is related to the aforementioned insensitivity of the macroscopic quantities, such as the density and number of locked states, within the hysteretic region to the different realizations of the randomly chosen initial conditions. This insensitivity has been confirmed by solving the ODEs on the same control grid five times with five different realizations of the initial conditions and observing the variance in the number of locked states within the hysteretic region to be $\sim 1\%$. Thus, it suffices to generate only a single speckling map, as shown in Fig. 7.

To gauge the accuracy the conditional probabilities calculated by the MLCs, we first calculate the ground truth (GT) conditional probability $p_{GT}(L|\psi_n, \Omega_n)$ or $p^{GT}_{L}$ for short, by solving the ODEs...
repeatedly at each point in the control space for \( N_i = 10,000 \) randomly selected initial conditions. In contrast, a single initial condition at each point in the control space is used to generate the ODE solutions presented in Sec. III A, which make up the data for the MLCs used below. The ground truth probability for locking is calculated by a Monte Carlo-type approach: by accepting each solution satisfying the locking criterion, \( \psi_n < \Omega_n \) based on the diagonal decision boundary shown in Fig. 6, and rejecting the solutions that fail this criterion. We then take the ratio of the number of accepted (locked) outcomes over the total number of initial conditions used for each of our 200 \( \times \) 200 points in the control space to arrive at \( p_{L}^{(GT)} \). The resulting GT probability is shown in Fig. 8(a). The color represents the conditional probability \( p_{L} \), with \( p_{L} = 0 \) corresponding to a unlocked state (blue) and \( p_{L} = 1 \) corresponding to a locked state (red). The figure also shows the analytic bifurcation boundary segments as thick black lines that demarcate the hysteretic region. As evidenced by the dashed black curve marking \( \psi_n > \Omega_n \), the contours of low probability are fairly close to the left boundary segment. In other words, the hysteretic region is taken up mostly by a region with a higher probability of locking, consistent with the locking bias introduced into the initial conditions. In a realistic scenario like an experimental or MHD simulation campaign, we will not have the luxury of calculating the ground truth probabilities.

The accumulation of the low probabilities toward the left boundary segment is strongly suggested in Fig. 8(b), which shows a histogram of the values of \( p_{L}^{(GT)} \)-a probability density of probabilities-restricted to points in the hysteretic region. This distribution is peaked, with a mean of roughly 0.55 (slightly shifted toward locking) and 0.4 \( \leq p_{L}^{(GT)} \leq 0.7 \) in a large area of the hysteretic region, consistent with the large white region in Fig. 8(a). The peaked histogram also shows an unexpected and peculiar structure in the form of noticeable gaps in the distribution for \( p_{L} \leq 0.25 \) and \( p_{L} \geq 0.9 \). The distribution of GT probabilities changes significantly for the neutral (less biased) case, discussed in Sec. IV, where a displacement of the large flat region to lower probabilities is observed.

To calculate the ML-based probability, we use the two control parameters \( (\psi_n, \Omega_n) \) as the input vector for the training. The target values of the training are the binary classification outcomes shown as the red and blue dots in Fig. 7. Note we do not hold out a fraction of the data for testing as it is commonly done in ML, because we are able to compare with the GT probabilities.

Three supervised classifiers—described in Appendix B 2—are used to estimate \( p_{L} \), and their results are displayed in Fig. 9: panel (a) shows a support vector machine (SVM) with a kernel that uses Gaussian radial basis functions (RBF); panel (b) shows logistic regression (LR) with a rational basis function; and panel (c) shows the neural network (NN) with four hidden layers, each consisting of 200, 100, 100, and 200 nodes, respectively. A rectified linear unit, “ReLU,” activation function is used to propagate the information forward through
the NN. The same color coding as that in Fig. 8(a) is used, with blue colors indicating a low probability of locking and red colors indicating a high probability of locking. The dashed black curve marks the \( p_L = 0.5 \) contour in each panel, and the solid black curves demarcate the hysteresis region. An accurate probability should reproduce \( p_L \to 0 \) at the upper (left) boundary segment and \( p_L \to 1 \) at the lower (right) boundary segment. The tangency at the critical point \( \beta = 0 \) \( \delta \Omega = (\delta \psi_w)^{3/2} \), leading to a thin region over which the probability varies rapidly, presents a serious challenge to the ML algorithms. The error in the ML probabilities is shown in Fig. 10 in terms of a residual, which is the difference between the GT probability and those calculated by each one of the three MLCs. The residuals are only plotted in regions where \( 0.01 \leq p_L \leq 0.99 \), to focus on the hysteretic region and its immediate surroundings. The root mean square error (RMSE) with respect to the GT probability for the same region is also reported in each panel of Fig. 10.

A support-vector machine with a radial basis function kernel (RBF SVM) [Fig. 9(a)] captures the rough structure, including the accumulation of low probability contours toward the left boundary segment and the sharp transition between the locked and unlocked regions below the CP. The latter property is a consequence of imposing a binary classification on the smoothly transitioning solutions. This is also where the largest residuals appear for all three MLC-calculated probabilities [Figs. 10(a)–10(c)]. The general shape of the SVM probability contours do not conform to the shape of the hysteretic region: the SVM probability contours do not conform to the shape of the hysteretic region marked by the solid black curves. This is also evident in the residual plot, Fig. 10(a), which indicates a RMSE \( = 0.10 \) for this classifier. An over-fitting problem in the form of wiggles in the left of the left hysteresis boundary segment and in the upper right corner of Fig. 9(a) appears as well. The reader can find other examples of SVM-generated probability plots, applied to mode amplitude and plasma inductance data from JET, in Figs. 7, 8, and 10 of Ref. 17. Note their phase space does not feature a CP where the probability contours of \( p_L = 0 \) and \( p_L = 1 \) come together.

The results from the rational logistic regression (RLR) are shown in Fig. 9(b). As described in Appendix B.2, the RLR uses for the argument of the sigmoid a rational function that is a quadratic function of the two inputs \( \langle \psi_w, \Omega_b \rangle \) in the numerator and a linear function \( \langle \psi_w, \Omega_b \rangle \) in the denominator. Similar to the results of the RBF SVM, the \( p_L \) contours as calculated by RLR also spill grossly outside the hysteresis boundary segments marked by the solid black curves. This is also evident in the residual plot, Fig. 10(a), which indicates a RMSE \( = 0.10 \) for this classifier. An over-fitting problem in the form of wiggles in the left of the left hysteresis boundary segment and in the upper right corner of Fig. 9(a) appears as well. The reader can find other examples of SVM-generated probability plots, applied to mode amplitude and plasma inductance data from JET, in Figs. 7, 8, and 10 of Ref. 17. Note their phase space does not feature a CP where the probability contours of \( p_L = 0 \) and \( p_L = 1 \) come together.

The neural network (NN) shown in Fig. 9(c) captures the accumulation of the low \( p_L \) toward the left boundary segment due to the locking bias, like the previous two classifiers. In terms of other metrics however, the NN clearly outperforms RBF SVM and RLR in accurately computing \( p_L \). The neural network probability \( p_L \) contours mostly capture the critical property that \( p_L \to 0/p_L \to 1 \) at upper/lower

![Image](https://via.placeholder.com/150)
bifurcation boundary segments; much better than what can be achieved with the other two classifiers. This is also evident in the residual plot of Fig. 10(c) that evinces an area that mostly conforms to the shape of the hysteretic region, and conforms well in the vicinity of the CP tangency. The NN results also have significantly smaller residuals than those shown in Figs. 10(a) and 10(b), producing the smallest RMSE (0.055), as indicated by Fig. 10(c). However, what stands out the most about the NN’s performance is the powerful smoothing effect produced by the use of an architecture with several hidden layers, each containing hundreds of nodes. For each pixel, it seems that the NN accurately brings in contributions from the neighboring pixels to result in an accurate probability, without doing too much smoothing near the boundary segments and especially the CP. This smoothing effect breaks down for fewer than approximately fifty nodes in most of the layers. Further details of the NN architecture are discussed in the next paragraph.

The sensitivity of the NN results to the network architecture has been investigated in terms of the number of hidden layers, the number of nodes contained within each hidden layer, and finally the type of nonlinear activation function used for each node. The locking probability exhibits notable sensitivity to these parameters, and the best accuracies are observed for architectures that use at least two hidden layers with at least 50 nodes in each layer. It should be noted that the choice of optimal network architecture has no closed form solution and has traditionally been found with trial-and-error. However, there are now automated machine learning techniques like neural architecture search (NAS) that can be used to optimize the network architecture.

The cost function of the network is minimized with some form of stochastic gradient descent (SGD) where the gradients of the hyperparameters are evaluated using backpropagation. This is discussed in Appendix B 2.

To illustrate further the differences between the three MLCs, we take horizontal and vertical slices through the 2D probability plots of Fig. 9 to compare the profiles of the MLC probabilities to those extracted from the GT probability shown in Fig. 8(a). The results are shown in Fig. 11: panel (a) shows $p_L$ profiles as a function of $\psi_w$ at $\Omega_0 = 1.5$ (black traces on the left) and 3.8 (red traces on the right); panel (b) shows $p_L$ profiles as a function of $\Omega_0$ at $\psi_w = 0.06$ (black traces on the left) and 0.15 (red traces on the right). One feature that immediately stands out here is the aforementioned gap in the GT probability, which appears here as very sharp increases and even apparent jumps in the solid curves near $p_L = 0$ and 1. Neither SVM (dashed-dotted) nor RLR (dotted) is able to capture this peculiar property, while the NN partially reproduces it, albeit a smooth version of it. The neural network also captures the flatness of $p_L$ in the hysteretic region that is evident in Fig. 9(c).

Another important issue addressed here regards the amount of data required for accurate training, as it relates to the sparsity of data in more realistic situations. A high-fidelity simulation campaign with an MHD framework like NIMROD may take months, even years to complete. In the case of a large-scale experiment, data can take years to collect. Thus, it is imperative to determine the minimum sample size required to train the MLCs for obtaining accurate predictions. To achieve this, we subject smaller datasets to the same procedure as described above to determine the threshold in sample size at which unacceptable loss of accuracy occurs. We have determined this

---

**FIG. 10.** Residuals between the conditional probability of locking ($p_L$) obtained from the ML classifiers of Fig. 9 and those obtained from the ground truth Monte Carlo calculation $p_L^{GT}$: (a) a support-vector machine (SVM) with a radial basis function (RBF) kernel, (b) logistic regression (LR) with a rational argument, and (c) a fully connected feed-forward neural network (NN) with four hidden layers, each consisting of 200, 100, 100, and 200 nodes, respectively. The residuals are only shown in regions where $0.01 \leq p_L \leq 0.99$, to focus on the hysteretic region and its immediate surroundings. The root-mean-square error (RMSE) of $p_L$ with respect to $p_L^{GT}$ over the same region is also shown.
threshold to be approximately $40 \times 40$ samples for the NN. Below this threshold, the RMSE exceeds 0.1 for the NN. The probabilities calculated by SVM and RLR are not as sensitive to sample size as the NN, possibly because of the much smaller number of free parameters (weights) that these algorithms optimize. This estimate is likely more applicable to simulations than experiments, as experimental data tend to be clustered and often complex in structure, as can be evidenced by Figs. 7 and 8 of Ref. 17.

**IV. THE NEUTRAL CASE**

**A. The solutions of the ODEs**

In this section, we repeat the procedures of Sec. III by applying them to a second case that is free of the locking bias. Many of the illustrations and much of the analysis of Sec. III are condensed to avoid repetition. We begin once again by integrating Eqs. (3)–(5) over the same range in time for the same $200 \times 200$ grid of control parameters, spanning $\psi_W = [0, 0.2]$ and $\Omega_n = [0, 5]$. However, the locking bias previously introduced into the initial conditions is reduced in this case by narrowing the range in $\psi_w(0)$ to $[0, 2]$ and increasing the range in $\Omega_n(0)$ to $[0.5]$. The range in $\theta_1(0)$ is kept the same as before $[0, 2\pi]$.

The phase diagram, i.e., time-asymptotic solutions of the ODEs as a function of $(\psi_w, \Omega_n)$, is shown in Fig. 12, with the color indicating the value of $\Omega_w$ as in Fig. 3. Once again, the color blue represents the high frequencies ($\Omega_w \approx 1.0$) and the color red the low frequencies ($\Omega_w \approx 0.0$). A feature that immediately stands out compared to Fig. 3 of Sec. III A is the very low density of the locked states (red/orange pixels) residing in the hysteretic regime. In fact, only a small fraction (≈9%) of solutions inside the hysteretic region lock. This is a consequence of the reduced locking bias applied to the initial conditions of the ODEs.

The cumulative distribution functions (CDFs) of the normalized order parameters are shown in Fig. 13. The most notable difference between these distribution functions and those shown in Fig. 2 is that the $\psi_w$ and $\Omega_n$ curves appear to have traded places, indicating that the majority of the solutions are now unlocked. The unlocked cluster also seems more diffuse, based on the distribution of $\psi_w$ curves, than the other two methods, the NN fairly captures the sharp increases in the GT profiles (solid trace).

The histograms (insets) also reflect these findings. The minima of the CDFs suggest locking for $\Omega_n \leq 0.3$ or $\psi_w \geq 0.5$, with the first criterion notably reduced compared to the threshold obtained from the CDF for the locking-biased case. Both criteria indicate approximately 42% of the total population to be locked, a significant reduction from 55% to 56% observed for the locking-biased case.

**B. Classification and locking probabilities**

We start the process as before by classifying the solutions of the ODEs, using the unsupervised classifier KMC. The figure for the KMC results is omitted here, as it shows approximately the same diagonal threshold.

**FIG. 11.** Profiles of $p_L^{(GT)}$ and $p_L$ obtained by the three MLCs, described in captions of Figs. 9 and 10. Panel (a) shows horizontal slices, i.e., the locking probability $p_L$ as a function of $\psi_w$ at $\Omega_n = 1.5$ (black traces on the left) and 3.8 (red traces on the right). Panel (b) shows vertical slices, i.e., $p_L$ as a function of $\Omega_n$ at $\psi_w = 0.08$ (black traces on the left) and 0.15 (red traces on the right). Unlike the other two methods, the NN fairly captures the sharp increases in the GT profiles (solid trace).

**FIG. 12.** The phase diagram: time-asymptotic solutions of the ODE system in the control space $(\psi_w, \Omega_n)$ for the “neutral” case, i.e., one that is less biased toward locking. The color represents the normalized rotation frequency $\Omega_w = \Omega_n/\Omega_b$. The two black curves bound the hysteretic region that merges at the critical point (CP) of the system.
The conditional probability of locking calculated by the neural network (NN) is shown in Fig. 16(a) for the less biased case under study here. The probability from the other two classifiers is omitted as Sec. III C has already established that NN yields the most accurate results. The neural network architecture has been altered from the previous configuration to produce once again the most accurate results with the smallest RMSE. To be specific, the network still consists of four hidden layers, but these four layers now contain 100, 100, 50, and 25 nodes, respectively. A ReLu activation function is used again to propagate the information forward through the NN. Only the configuration to produce the most accurate results. The residual between the GT probability and the NN probability for the neutral case is shown in Fig. 16(b). The same intermediate step that converts the phase diagrams, displaying the color plots of the normalized order parameters in the control space, to a strict binary map is carried out here as well. The resulting speckling diagram, which is the counterpart of Fig. 7, appears in Fig. 14. The ground truth conditional probability of locking $p_L^{(GT)}$ for the neutral case is shown in Fig. 15(a). This probability is calculated by the same Monte Carlo procedure as that described in Sec. III C, except now the GT calculation uses the same range in the initial conditions as that used to produce Fig. 14. For clarity, only the probability in the hysteretic region is shown; $p_L^{(GT)} = 0$ holds above the top boundary segment and $p_L^{(GT)} = 1$ holds below the bottom segment, and are not plotted. Observe the much smaller values of $p_L^{(GT)}$ appearing in the color bar on the right, which is also indicated by the histogram of $p_L^{(GT)}$ values in the hysteretic region, shown in Fig. 15(b). A gap in $p_L^{(GT)}$ appears here as well, similar to the one encountered in the case with the locking bias [Fig. 8(b)]. However, this gap now only appears on the side of the high probabilities. In fact, no point in the hysteretic region features a locking probability greater than ~40%. A profile plot like in Fig. 11 (not shown) shows an extremely rapid increase in probabilities as the bottom boundary segment is approached.

The conditional probability of locking calculated by the neural network (NN) is shown in Fig. 16(a) for the less biased case under the basic aspects of the locking dynamics that are useful for the MLCs. The advantage of this model is the rapid generation of data—a few minutes on a few dozen processors—required for the MLCs, in addition to the well-demonstrated success of similar modeling in capturing the realistic scenarios in which the amount of data that can be obtained from either experiments or high-fidelity simulations can be limited.

The data required for the MLCs are generated by solving a simplified third-order ODE system that describes the locking dynamics. The advantage of this model is the rapid generation of data—a few minutes on a few dozen processors—required for the MLCs, in addition to the well-demonstrated success of similar modeling in capturing the realistic scenarios in which the amount of data that can be obtained from either experiments or high-fidelity simulations can be limited.

This paper presents a proof-of-principle of using MLCs to forecast mode-locking by calculating the probability of a rotating tokamak plasma to lock to a static error field. It is demonstrated here that this locking probability can be applied successfully to sparse data and provide meaningful and quantitative insight. This capability is relevant to the realistic scenarios in which the amount of data that can be obtained from either experiments or high-fidelity simulations can be limited.
The locking probability is conditional on the two control parameters of the model, which are the error field magnitude $\psi_w$ and the frequency $\Omega_0$ associated with the injected or intrinsic momentum source. The ODEs are integrated to the nonlinear steady-state for each point on a 200x200 uniform grid of these two control parameters. The dependent variables are the amplitude and phase of the tearing mode, $\psi_t$ and $\theta_t$ as well as $\Omega_t$ the plasma rotation frequency at the tearing layer. The nonlinearly saturated values of these variables are the order parameters of the system. A single set of initial conditions for three-order parameters are chosen randomly, over a prescribed range for each point in the $$(\psi_w, \Omega_0)$$ grid. The relative dimensions of the set of initial conditions from which a sample is taken randomly determine the tendency of the system to lock. Two scenarios are considered here: one with a strong locking bias and another with a reduced locking bias (neutral).

The time-asymptotic steady-state solutions of the ODEs agree with the analytic steady-state solutions, one of which emerges from the well-known torque balance equation, which is a cubic equation in $\Omega_t$. The steady-state solutions show the presence of a hysteretic region in the control parameters $$(\psi_w, \Omega_0)$$, where the cubic equation of torque balance has three real roots. The upper and lower roots represent attractors, and initial conditions near the middle root converge to one of the two attractor solutions. In other areas of the control parameter space, only one real root exists, corresponding to one steady-state solution that is either locked or unlocked. The cubic equation of torque balance is analogous to the van der Waals (VDW) equation of state that describes (classical) phase transitions between liquid and gaseous phases of matter.
The ODE solutions, i.e., the order parameters, are first normalized to span \([0, 1]\). Of these normalized parameters, \(\psi_m, \Theta_l, \text{and } \Omega_m\) the phase \(\Theta_l\) is redundant in the sense that it depends on \(\psi_m\), independent of control parameters. The remaining two normalized order parameters show a strong clustering pattern, with the unlocked solutions accumulating around \((\psi_m, \Omega_m) \approx (0.0, 1.0)\) and locked solutions accumulating around \((\psi_m, \Omega_m) \approx (1.0, 0.0)\).

The hysteresis in the ODE system manifests itself in the control space \((\psi_m, \Omega_m)\) as a sizable region in which there is a strong mixing of locked and unlocked solutions. This mixing implies a dependence on the randomly selected initial conditions, leading to an easily perceived density of locked states within this hysteretic region of the control parameter space, as seen in Figs. 7 and 14. This mixed population turns homogeneous near the bifurcation boundary segments: becoming unlocked near the segment of the boundary between the hysteretic and the unlocked regions, and becoming locked near the segment between the hysteretic and the locked regions. The boundary segments of the hysteretic region merge at the critical point (CP) of the system below which smooth transitions are seen to occur between the two phases. This is analogous to the behavior near the CP in the VDW equation of state.

Because of the co-existence of the locked and unlocked states of the plasma in the hysteretic region, it is useful to describe the hysteretic behavior in terms of a probability of locking, conditional on the control parameters of the system. This probability is a measure of the relative size of the basin of attraction of the locked state. In the context of tokamaks physics, this probability is a measure of the robustness of an unlocked state to a disturbance like a large sawtooth or a large ELM.

For calculating the locking probability via ML methods, we follow a twofold approach. We first subject the normalized order parameters \((\psi_m, \Omega_m)\) of the model to an unsupervised classification scheme that labels each solution as locked or unlocked, free of human input. The K-means clustering (KMC) algorithm, which is a geometric method, is chosen for this task because of the approximately even density of solutions in the two clusters. The classification with KMC indicates locking for \(\Omega_m < \psi_m\). Alternatively, a single-order parameter can also be used as the input for KMC, which results in a locking criterion in terms of either \(\Omega_m\) or \(\psi_m\). These results have also been benchmarked with Gaussian mixture models (GMM), a probabilistic method that fits anisotropic Gaussians to the clusters. The different locking criteria from either KMC or GMM agree to better than 99% because of the very strong concentration of solutions around the two clusters. The normalization of the order parameters makes this clustering behavior evident, and this normalization leads to order parameters for which a geometric region merge at the critical point (CP) of the system below which smooth transitions are seen to occur between the two phases. This is analogous to the behavior near the CP in the VDW equation of state.

In contrast to the GT calculation, which takes approximately a week of parallel computation on several dozen processors, the ML methods take about a minute to train. (Compared with the Monte Carlo computations described above, these methods use a single randomly selected set of initial conditions at each point in the control parameter space.) These savings in computation time open up the possibility of applying these methods in real-time operations, once they are adequately trained on experimental data.

Of the three MLCs that have been used, the NN produces probabilities that best agree with the GT probabilities and results in the smallest residual and root mean square error (RMSE). The other two methods lead to the results that do not match the ground truth probabilities as well, especially in capturing the finer features near the CP and along the boundary segments of the hysteretic region. They also yield a larger RMSE. The above network architecture is modified for the neutral case to produce the smallest RMSE.

Future work will include modifying the ODEs to model the nonlinear saturation of the tearing mode associated with the flattening of the plasma current, and by incorporating resistive wall (RW) effects. We have investigated the first effect partially: the saturation term changes the distribution of the solutions in the 2D space of the normalized order parameters \((\psi_m, \Omega_m)\) and thus leads to a new criterion for locking. The incorporation of the RW makes the system fifth order, as the RW is coupled both to the tearing layer and to an outer perfectly conducting boundary. In this case, it is the difference between the phase of the tearing mode (TM) and the phase of the magnetic perturbation at the RW that determines the evolution of the TM and plasma rotation at the rational surface. Another fifth-order system modeling the effect of the \((2, 1)\) tearing mode on the control system involving tokamak I and C-coils placed on either side of the (resistive) wall is presented in Ref. 36. Modes of different helical symmetries can also be included to model the poloidal coupling between tearing modes of different helicities. It is exactly in situations like these where there are no analytic time-asymptotic solutions that the power of an unsupervised classifier truly emerges, for unambiguously distinguishing between the locked and unlocked solutions.

The ODE model of the locking physics with the above modifications can be validated on the experimental data. The idea here is to tune the model’s parameters to the internal/external magnetic field and toroidal rotation measurements of DIII-D or another tokamak to determine if the NN can first predict the observed locking. These possible steps can possibly be carried out within the framework...
of the disruption event characterization and forecasting (DECAF)
code,17 which already uses the standard torque balance/mode-locking
model for disruption prediction, and has access to many tokamak
databases.

Other possible extensions include comparing with the probabilis-
tic locked-mode predictor of Murari et al.,12 using an NN instead of
SVM, on the JET data of plasma inductance and mode amplitude;
training supervised classifiers on the DIII-D measurements of cross
c field and field-aligned current density gradients inside and outside the
minimum current density near the q = 2 surface. These gradients are
used to determine the stability of the 2/1 TM in DIII-D discharges that
simulate the ITER baseline scenario.10 The trained classifiers can then
be used to determine the stability of the 2/1, either deterministically or
probabilistically, as a function of the current density gradients for
future discharges.

ACKNOWLEDGMENTS

We thank Edward Strait, Laszlo Bardozci, Jayson Barr from
General Atomics, and Sandeep Madireddy from ANL for invaluable
discussions and their guidance. We also thank the anonymous
referees for their suggestions and helpful feedback. This work was
supported by DOE Office of Science Collaborative Grant Nos. DE-
SC0019016 and DE-SC0014005, respectively.

APPENDIX A: ANALOGY WITH THE VAN DER
WALLS EQUATION OF STATE

It is interesting to note that the van der Walls (VDW) equation
of state (EOS) is also cubic as a function of the density $n$ of a non-
ideal gas,

$$a b n^3 - a n^2 + (T + b p) n - p = 0,$$  \hspace{1cm} (A1)

where $T$ and $p$ are the temperature and pressure of the material; $a$
and $b$ are the VDW constants related to the inter-molecular poten-
tial and the molecular size, which have different values for each gas;
and $R$ is the universal gas constant.

Defining the normalized frequency $x = \Omega / \Omega_c$ for Eq. (9) with
the following simplifications $\mu = \Delta \lambda = 1$ as well as setting $a = b$
$= R = 1$ in Eq. (A1), shows that these two equations are analogous,

$$x^3 - x^2 + \left( \frac{\psi^2}{\Omega_c^2} + \frac{1}{\Omega_c^2} \right) x - \frac{1}{\Omega_c^2} = 0,$$  \hspace{1cm} (A2)

$$n^3 - n^2 + (R T + p) n - p = 0.$$  \hspace{1cm} (A3)

These equations establish an equivalence between the VDW control
parameters ($p, T$) and ours ($\Omega_c, \psi$), with the following analogs:
$n \leftrightarrow x$ and, in units with $R = 1$,

$$p \leftrightarrow 1 / \Omega_c^2, \hspace{1cm} T \leftrightarrow \psi / \Omega_c,$$  \hspace{1cm} (A4)

$$\Omega_c \leftrightarrow 1 / \sqrt{p}, \hspace{1cm} \psi \leftrightarrow \sqrt{T} / p.$$  \hspace{1cm} (A5)

The locked and unlocked states in this paper correspond, respectively,
to the gaseous and liquid phases in the VDW equation of state.

One difference between the locking–unlocking bifurcation
problem and the phase transformation problem exists in the hyste-
retic regime. In the former, there is a probability of locking
$p(L|\psi, \Omega_c)$, which varies from zero at the boundary segment with
the unlocked region (the upper black curve in Fig. 3) to unity at the
segment with the locked region (the lower black curve there), meas-
uring the respective number of initial conditions leading to a
locked state. In the phase transformation case, the gaseous and liq-
uid phases can co-exist, with the fraction of each roughly corre-
spending to $p(L|\psi, \Omega_c)$. The analogy would possibly be more
complete if the ODE for the locking problem included additive
noise that can cause a transition from locked to unlocked and back
for control parameters in the hysteretic region, similar to the effect
of fluctuations on phase transformations. With this noise, another
measure of the probability of locking might be the amount of time
spent, in the presence of the noise, near the locked state.

The VDW critical point, with large values of the control
parameters ($p, T$), corresponds to small values of ($\psi, \Omega_c$) [cf.
Eqs. (A4) and (A5)]. The possibility of traversing a path in the con-
rol space ($p, T$) above the critical point, with the state smoothly
transforming from the gaseous state to the liquid state, is the exact
analog to tracing a path in Fig. 1 below the critical point.

APPENDIX B: DESCRIPTION OF THE MACHINE
LEARNING CLASSIFICATION (MLCs)

In this appendix, we outline the ML methods we employ in our
two-step scheme. The first step is to do clustering, in order to label states
in the normalized order parameter space as locked or unlocked. In the
second step, we use this classification to estimate the probability of lock-
ing as a function of the two control parameters, especially in the hyster-
etic region, where both locked and unlocked states can co-exist.

1. Unsupervised classifiers

We use K-means clustering29,30 to classify each solution of the
ODEs as locked or unlocked, without human input. The aim of
KMC is to partition $N$ sample points into $K$ clusters in which each
sample belongs to the cluster with the nearest mean cluster center
or centroid. In our application, we use this classification scheme
with $K = 2$ to classify all of the points in the normalized order
parameter space as locked or unlocked states. This technique is
used in image segmentation and compression. One begins by ran-
domly choosing $K$ points as the initial estimates of the centroids
of $K$ clusters. One then assigns each sample to the cluster with the
closest centroid, and then re-computes $K$ new centroids by using
the newly re-labeled samples in each cluster. The two phases of re-
assigning samples to the clusters and re-computing the cluster cen-
troids continue until the centroid locations no longer change or,
equivalently, there is no further change in assignments. The algo-

rithm minimizes the sum of the squares of the distances of each
data point to the closest centroid. It also has the capability of pre-
dicting the class of any new or additional samples by associating the
new sample with the cluster with the nearest centroid. Since this
method depends on distances between points, the choice of the
metric is important and, as we have discussed, choosing the
Euclidean metric on the space of normalized parameters $\psi$ and
$\Omega_c$ is a natural choice. The labels or categories of this classification
form the target values required in the training of the supervised
classifiers that are used to calculate the locking probabilities as a
function of the two control parameters.
We have also applied Gaussian-mixture models (GMM) to our data. GMM are another unsupervised classification method which fits $K$-many anisotropic Gaussians to the data, where each Gaussian represents one cluster. GMM make "soft" assignments to each data point as opposed to the hard assignments made by KMC.

2. Supervised classifiers

We make use of three classes of supervised classifiers to compute the locking probability as a function of the two control parameters. These are support-vector machines (SVM), logistic regression (LR), and a fully connected feed-forward neural network (NN), also known as a multi-layer perceptron (MLP). These methods—as do all supervised classifiers—require the training set to contain the targets as well as the inputs. The targets for the training are the binary classification of the solutions into the locked and unlocked states, supplied in this case by KMC, as described above and implemented in Sec. III B. The paragraphs below contain the descriptions of each of the three supervised classifiers used in this work. A comprehensive review of these classifiers can be found in Ref. 29.

To ease the reader into the following MLC descriptions, we first define some terms and notation. In the presence of multiple features like the height and weight of a person, the input vector samples take the form of vectors lined up to form a 2D design matrix $X \equiv X_{\mu n}$. Here, $\mu = 1, \ldots, N$ tracks the samples of a dataset with $N$ samples and $\mu = 1, \ldots, D$ tracks the components of the input or feature space. For our application, the inputs are the coordinates $(\theta_w, \Omega_w)$ of the $200 \times 200$ grid in the control space, yielding $D = 2$ and $N = 40 000$. Thus, $X = (x_1, x_2) = (x_n)$ where the bold symbol is to remind the reader that each feature $x_n$ is a vector containing all of the $N$ samples. For simplicity, the subscript $n$ is suppressed in most of these definitions. Finally, the notation $x = (x_1, x_2) = (x_n)$ is reserved for a single sample and similarly, $x_n$ for the $n^{th}$ sample.

Support vector machine (SVM) separates two classes by maximizing the margin, defined as the smallest distance between the linear decision boundary separating the two classes and any of the samples belonging to either class. Maximizing this margin leads to a particular choice of decision boundary, whose location is determined by a subset of the closest data points, known as support vectors. One can transform a nonlinear boundary in a lower dimensional space into a linear boundary in a higher dimensional space by performing the kernel trick or kernel substitution. A common choice for a kernel in SVM is radial basis functions (RBF), which produce an infinite dimensional feature space.

The kernel trick is equivalent to augmenting $X$ to a higher dimensional nonlinear feature space $\phi(X)$, e.g., $X = (x_1, x_2) \rightarrow \phi(X) = (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2)$. A linear classifier in this feature space leads to a nonlinear classification in the original space. Note that SVM has to account for overlapping class distribution in this case because of the strong intermixing of the locked and unlocked states in the hysteretic region, as indicated by Figs. 7 and 14. This is accomplished by penalizing the points on the wrong side of the boundary with a penalty that increases with the distance from the boundary. This is known as a soft-margin classifier.

A support-vector machine does not provide probabilistic outputs. However, probabilities out of an SVM classification can be obtained by fitting the SVM output to a logistic sigmoid function described in the following paragraph. This is known as the Platt’s method.47,41

Logistic Regression (LR) provides a probabilistic classifier by best fitting a logistic sigmoid function $\sigma(x) = 1/(1+e^{-x})$ to a binary classification problem. LR is a maximum likelihood estimator of the parameters of the logistic sigmoid, and in its most primitive form, provides an estimate of the probability of belonging in either one of two classes: locked and unlocked states of the plasma in this case.

For the present application and focusing on a single sample $x$ with features $(x_1, x_2)$ and bias $\beta$, $x_0 = 1$, the sigmoid function $\sigma(x \cdot \beta)$ corresponds to the probability of locking $p(L|x, \beta)$ conditional on the weights $w = (w_0, w_1, w_2)$. Here, the inner product is performed over the features: $\mathbf{w} \cdot x = w_0 x_0 + w_1 x_1 + w_2 x_2$. The conditional probability of being unlocked is $p(U|x, \beta) = 1 - \sigma(x \cdot \beta)$. For the $n^{th}$ sample of the dataset $(\mathbf{x}_n, t_n)$, we define $t_n = 1$ for locked states and $t_n = 0$ for unlocked states, so $p(t_n = 1|x_n) = \sigma(x_n \cdot \beta_n) \equiv \sigma_n$ and $p(t_n = 0|x_n) = 1 - \sigma_n$. Then the likelihood for locking on a single trial can be written as $p(t_n|x_n) = \sigma_n^{t_n} (1 - \sigma_n)^{1-t_n}$. Assuming independence of all the measurements for all values $t_n$, the likelihood to be maximized is

$$p(t|\mathbf{w}) = \prod_{n=1}^{N} \sigma_n^{t_n} (1 - \sigma_n)^{1-t_n}. \tag{B1}$$

(Interpreting this in a Bayesian sense, the posterior is—except for a constant of proportionality—equal to the likelihood, assuming that the prior is taken to be uniform.) The negative log-likelihood is the cross entropy:

$$E(\mathbf{w}) = -\log[p(t|\mathbf{w})] = -\sum_{n=1}^{N} [t_n \log \sigma_n + (1-t_n) \log(1-\sigma_n)]. \tag{B2}$$

The weights found by the process of minimizing $E(\mathbf{w})$ for the data points $(\mathbf{x}_n, t_n)$ give an estimate for the probability of belonging to the locked class.

For a 2D input space, LR produces probability contours that are parallel straight lines, which do not conform to the shape of the hysteretic region in Fig. 3. To capture the properties of the hysteretic region, the linear bases of LR are augmented to form a nonlinear feature space, as discussed above for SVM. This generalization, based on a linear combination of the feature vectors, is known as a generalized linear model (GLM), in that the argument of the sigmoid is linear in the parameters, i.e., the weights $w$. What is carried out in Sec. III C is a further generalization of this; we take as the argument of the sigmoid a rational function of the form $Q(x)/L(x)$ that represents a ratio of a quadratic to a linear polynomial. This rational logistic regression (RLR) is no longer a GLM as the argument of the sigmoid is not linear in $w$. In fact, it results in a nonlinear optimization process—of eight weights and biases in this case (six for the quadratic polynomial and three for the linear, but the bias in the denominator can be scaled to unity)—that is extremely sensitive to the initial values of the weights, and for which ScikitLearn’s LR algorithm is ill suited. Thus, we carry out the RLR
learning occurs in the neural network by changing connection weights \( w_1, \ldots, w_N \) and biases \( w_0 \) for each layer after each piece of data are processed, to minimize the amount of error in the output compared to the target \( t_k \) values (expected result). The error (cost) function is defined as

\[
E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} [y_n(\mathbf{x}, \mathbf{w}) - t_n]^2 ,
\]

for a \( N \)-many outputs.

The methods used to minimize \( E(\mathbf{w}) \) are variants of the gradient descent (GD) method; each iteration moves the weights \( \mathbf{w} \) in the direction opposite to the gradient of \( E(\mathbf{w}) \). The corrections to each weight is obtained via backpropagation where the information is propagated backward through the network, from the outputs to the inputs, for each iteration. Other enhancements, such as stochastic (online) gradient descent (SGD), are often used, because of the inefficiency of the standard GD, where the parameter updates are performed on the basis of a single sample or subset of samples that is picked randomly at each iteration. SGD is superior to the standard GD because it converges much faster. However, the solution that SGD finds is not always the global minimum, but rather a local minimum that is often acceptably close to the global minimum. The probability of finding a "bad" (high value) local minimum is non-zero for small-size networks and decreases quickly with the network size. Hyperparameter tuning can be automated with the Bayesian optimization.

We use Python’s Scikit-learn libraries to apply the said ML algorithms.

**DATA AVAILABILITY**

The data that support the findings of this study are available from the corresponding author upon reasonable request.

**REFERENCES**

3. Mode-locking is not always the primary cause of disruptions in a tokamak. EAST disruptions, for example, are primarily caused by vertical displacement events, while mode-locking due to NTM’s was the largest single cause of disruptions in JET with the carbon wall.
This calculation requires integrating the ODE system a total of $4 \times 10^8$ times.