PAPER • OPEN ACCESS

Transfer learning nonlinear plasma dynamic transitions in low dimensional embeddings via deep neural networks

To cite this article: Zhe Bai *et al* 2025 *Mach. Learn.: Sci. Technol.* **6** 025015

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
4 May 2024

REVISED
11 March 2025

ACCEPTED FOR PUBLICATION
8 April 2025

PUBLISHED
16 April 2025

PAPER

Transfer learning nonlinear plasma dynamic transitions in low dimensional embeddings via deep neural networks

Zhe Bai[1,*], **Xishuo Wei**[2], **William Tang**[3,4], **Leonid Oliker**[1], **Zhihong Lin**[2] **and Samuel Williams**[1]

[1] Applied Mathematics and Computational Research Division, Lawrence Berkeley National Lab, Berkeley, CA 98195, United States of America
[2] Department of Physics and Astronomy, University of California, Irvine, CA 92697, United States of America
[3] Princeton Plasma Physics Laboratory, Princeton, NJ 08543, United States of America
[4] Princeton University, Princeton, NJ 08544, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: zhebai@lbl.gov

Keywords: transfer learning,model order reduction, embeddings, plasma physics, nonlinear dynamics, bifurcation

Abstract

Deep learning algorithms provide a new paradigm to study high-dimensional dynamical behaviors, such as those in fusion plasma systems. Development of novel, data-driven model reduction methods, coupled with detection of abnormal modes with plasma physics, opens a unique opportunity to identify plasma instabilities through automated construction of parsimonious models that can be tuned to balance accuracy and cost. Our fusion transfer learning (FTL) model demonstrates success in rapidly reconstructing nonlinear kink mode structures by learning from a limited amount of nonlinear simulation data. The knowledge transfer process leverages a pre-trained neural encoder–decoder network, initially trained on linear simulations, to effectively capture nonlinear dynamics. The low-dimensional embeddings extract the coherent structures of interest, while preserving the inherent dynamics of the complex system. Experimental results highlight FTL's capacity to capture transitional behaviors and dynamical features in plasma dynamics—a task often challenging for conventional methods. The model developed in this study is generalizable and can be extended broadly through transfer learning to address various magnetohydrodynamics modes.

1. Introduction

Modeling, design, and control of fusion plasma that characterize a high-dimensional, strong nonlinear system are a central challenge for tokamak-based plasma devices. In such systems, the wide range of scales in both space and time necessitates a large amount of data collection, processing, and computation to resolve all physically relevant features for system identification and control. The intricate, multi-scale dynamics usually render classic control methods impractical for real-time feedback control of tokamak experiments. Dynamic model-based approaches provide a viable alternative, enabling the implementation of more adaptive and robust control strategies. However, their efficacy relies on the availability of mathematical models to accurately emulate the system dynamics. Consequently, computational modeling plays an essential role in understanding, estimating, and eventually developing model-based control for such complex physical processes.

Characterizing the evolution of plasma dynamics is crucial for effective plasma control. In magnetically confined plasmas, macroscopic magnetohydrodynamic (MHD) instabilities can be excited by equilibrium currents or pressure gradients. Various MHD instabilities, including fishbones, sawtooth, neoclassical tearing modes (NTM), and kink modes, can limit burning plasma performance and threaten fusion device integrity [1–4]. Although the kink mode does not directly cause confinement degradation, it can non-linearly excite the more destructive modes like sawtooth and NTM that affect the particle confinement or even cause a major plasma disruption. Detecting anomalies allows for early identification of the critical events, enabling

proactive measures to mitigate them and maintain stable plasma conditions. Investigations of plasma instabilities can enhance understanding and prediction of core plasma dynamics and transport, as well as edge plasma-material interactions critical for operation of future fusion devices like ITER [5, 6]. The evolution of instabilities typically encompasses of linear and nonlinear phases, with the former characterized by small perturbations and the later with saturated amplitude. As perturbation amplitude increases exponentially in the linear phase, nonlinear damping effects compete against the linear instability drive, leading to the decrease of the growth rate and eventually the saturation of the perturbation amplitude, with fluctuation amplitudes reaching a steady level, quenching to near-zero, or exhibiting limit cycle oscillation patterns [7, 8]. In fusion experiments, directly observing the linear growing phase is challenging. Understanding the nonlinear evolution of fluctuations induced by instabilities is crucial for validating simulations, predicting physical dynamics, and preventing plasma disruptions. Nonlinear bifurcations in the dynamics of the plasma are frequently observed, exemplified by transitions such as from low-confinement mode (L-mode) to high-confinement mode (H-mode) induced by auxiliary heating, variations in turbulence-zonal flow interactions due to distinct damping parameters, and the formation of internal transport barriers resulting from the saturation of macroscopic MHD modes. The bifurcations signify the existence of multiple steady states of plasma turbulence and often indicate pathways to enhanced confinement. Such bifurcations appear at the nonlinear saturation of unstable waves, driven by reductions in free energy and increased damping effects.

High-fidelity simulations of systems governed by nonlinear partial differential equations (PDEs) often entail substantial computational costs, making them impractical for decision-making tasks in real-world applications. To address this issue, model order reduction methods have been developed to approximate the behaviors of such system while significantly reducing computational overhead. Reduced-order models (ROMs) play a pivotal role in characterizing, estimating, and controlling high-dimensional complex systems [9–12]. In an approach orthogonal to physics-based reduced order models based on first-principle equations, data-driven methods study low-dimensional latent space trajectories for dynamical systems using deep learning informed by simulation and/or experiment data measurements. Although the fully-resolved state space of a complex system, e.g. turbulence and tokamak plasmas, may contain millions of degrees of freedom, these nonlinear systems exhibit inherent low-dimensional, dominant patterns in latent space, and typically evolve on a low-dimensional attractor that can be leveraged for effective control [13, 14]. The essential steps in model reduction include identifying an optimal state space for the attractor and characterizing the system's dynamical behavior when evolved on this attractor. Conventional approaches are projection-based, such as proper orthogonal decomposition (POD)-Galerkin method that performs the Galerkin projection of the first principle equations onto a linear subspace of modes obtained via POD [15–17]. Dynamic mode decomposition (DMD) is another dimensionality reduction technique to decompose high-dimensional data into dominant spatiotemporal coherent structures [18, 19]. As a numerical technique to approximate the Koopman operator [20–22], DMD and its variations [23, 24] have established a strong connection to the analysis of nonlinear dynamical systems by constructing modes that oscillate at a fixed frequency and growth or decay rate in a linear evolution model. Non-intrusive model reduction based on projection methods enables approximating the low-dimensional operators using regression techniques [25, 26]. However, these methods pose challenges when applied to multi-scale dynamical systems due to the inherent linear characteristics of the model. Such limitations become particularly pronounced in the study of plasma dynamics, where transitions can dramatically change in an unpredictable manner due to the complex, nonlinear nature. Therefore, the development of new models is critical to effectively capture the spatiotemporal structures in a low-dimensional space that accurately represents the physical observations.

In this work, we introduce fusion transfer learning (FTL), a data-driven reduced order model that efficiently characterizes and reconstructs MHD modes. The encoder–decoder based network, as illustrated in figure 1, captures the essential spatial features from the simulations of the internal kink modes and is capable of rapidly detecting anomalous mode structures when presented with snapshots of dynamical plasma structures. The network, trained on non-dynamical modes, is extended to embed time-evolution kink instabilities through transfer learning. One of the principal challenges for ROMs is modeling transient system behaviors particularly when the system is activated by smoothly varying external forcing. Trajectories in the latent space provide informative changes in a low-dimensional embedding manner. We focus on identifying the bifurcation point in the low dimensional space observed when plasma instabilities transit from the linear to the nonlinear phase.

The remainder of the paper is organized as follows. In section 4, we elaborate the results of mode reconstruction, detected anomalous modes, as well as learned nonlinear modes and their elicited bifurcation using the FTL model. Section 5 discusses the physical interpretation of the tipping points observed in the latent space, the limitation of this work and potential future directions. In section 3, we describe gyrokinetic

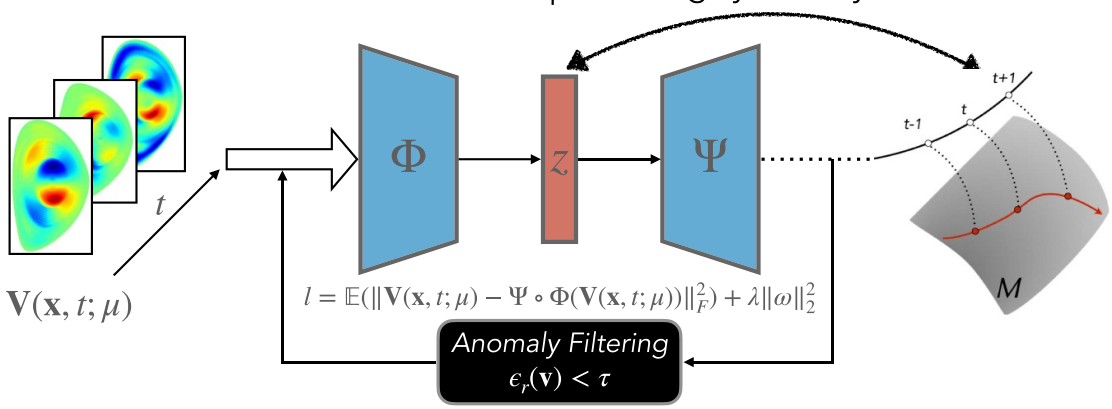

Figure 1. Schematic diagram of the FTL model using an encoder (Φ) - decoder (Ψ) network. The coherent structure is compressed to a low-dimensional vector $\boldsymbol{Z}$ while the inherent dynamics are preserved in the latent space through embedding. Regularization is employed to impose penalties on the model's complexity. An anomaly filtering subroutine is utilized for the identification of outliers, thereby improving the fidelity of the pre-trained model.

toroidal code (GTC) simulation configurations, the architecture of the FTL network, the algorithm to detect anomalous modes, and our approach to learn dynamics through pre-trained models on non-sequential data from linear kink simulations. The contributions for this work are summarized as follows:

- We propose FTL, a data-driven ROM, wherein low-dimensional embeddings enable efficient reconstruction of spatiotemporal MHD modes.
- We develop a method for anomaly detection and filtering based on an encoder–decoder network architecture, aiming to identify outliers from raw datasets and improve the fidelity of the pre-trained model.
- We demonstrate the feasibility of the proposed FTL on out-of-sample regimes in the parameter space, highlighting its capability for extrapolation and generalization to complex mode structures while achieving accelerated convergence with a parsimonious model through transfer learning.
- We investigate the plasma evolution and dynamic transitions in the latent space, providing insights into the correlation between the encoded representations of dynamic transitions and their manifestations in both real and frequency domains, thereby enhancing the interpretability of the FTL model in physics.

2. Background

2.1. Plasma instabilities

Plasma instabilities such as MHD modes, micro-instabilities, and Alfvén eigenmodes (AEs) are extensively studied using physics models and computational algorithms, including MHD codes [27–32], kinetic particle-in-cell codes [33–36], and kinetic Eulerian codes [37–39], among others. GTC [33] stands as a prominent plasma physics simulation tool, incorporating multiple simulation models such as gyrokinetic, fully-kinetic, and fluid covering multi-scale physics. The validation of multi-scale nonlinear simulations can be quantitatively achieved through their comparison with experimental data [40, 41]. Recent advances in scientific machine learning (SciML) present new possibilities to address these challenges in fusion science. Kates-Harbeck *et al* proposed the use of recurrent neural networks in FRNN [42] to forecast disruptions for large burning plasma systems ITER. A recent study on deep reinforcement learning [43] explores an autonomous, self-learning control strategy and produces a set of plasma configurations to control tokamak plasma by acquiring knowledge interacting with its environment. The experimental investigations have revealed that the disruptions are associated with the current-driven MHD modes and various additional effects [44]. However, a comprehensive physical understanding of the disruption process remains elusive, and addressing this necessitates the utilization of a simulator capable of elucidating the physics mechanism underlying plasma behaviors at specific time points based on real-time measurement signals. The Surrogate of GTC (SGTC) uses deep learning methods to predict plasma instability [45]. Trained using GTC data, SGTC can accurately predict linear stability and the 2D structure of the kink mode in a few milliseconds—five orders of magnitude faster than conventional first-principle based simulations. Nevertheless, as any real fusion plasma system is fundamentally nonlinear, complex and multi-scaled, determining control laws directly from the high-dimensional data is hardly feasible considering its constraints. Therefore, a

low-dimensional representation characterizing bifurcation analysis is needed to provide insights into how such complex systems behave, in order to understand and anticipate critical transitions.

Many physical models use simplified equations to obtain insight and intuition in a low-dimensional physical space. Examples include the Kadomtsev model, which describes the sawtooth cycle [46]; the predator–prey model that describes zonal flow evolution [8]; and the modified Rutherford equation [47] to describe the NTM instability. The reduced MHD equations for kink simulation used in GTC and other MHD codes are also a type of reduced physical model. These models derive simplified equations from the complete equations by applying assumptions that are valid within a certain parameter range. In contrast, data-driven methods provide a general model that analyzes complex simulation or experimental data, directly extracting latent features from raw data. Compared to other reduced models, the ML-based models offers significant advantages: they are much faster, with an inference time typically around 1ms; their dimension reduction is more aggressive than that of usual reduced physical models; and its methodology can be readily applied to other physical modes such as NTM and AEs without deriving a new set of reduced equations.

2.2. Gyrokinetic toroidal simulation and data generation

The GTC model employed to simulate the 5000 DIII-D experimental shots [45] is the single-fluid model in low-frequency, long-wavelength limit [48]. GTC uses the perturbation method, in which the physical quantities are separated to equilibrium and perturbed parts. The equilibrium quantities, including temperature, density, magnetic field and the flux surface shape are taken from the reconstruction of DIII-D experiments, and the perturbed quantities are solved from GTC equations. The equations of single-fluid model include the continuity equation to solve charge density δn, Poisson's equation to solve $\delta\phi$, the Ampere's law to solve $\delta u_{\parallel}$, the Faraday's law with the assumption $E_{\parallel}=0$ to solve $\delta A_{\parallel}$, and the perpendicular force balance equation to solve $\delta B_{\parallel}$. The first equation is the continuity equation for gyrocenter charge density,

$$\begin{aligned}&\frac{\partial\delta n}{\partial t}+\mathbf{B}_0\cdot\nabla\left(\frac{n_0\delta u_{\parallel}}{B_0}\right)-n_0\mathbf{v}_*\cdot\frac{\nabla B_0}{B_0}+\delta\mathbf{B}_{\perp}\cdot\nabla\left(\frac{n_0u_{\parallel 0}}{B_0}\right)\\&-\frac{\nabla\times\mathbf{B_0}}{eB_0^2}\cdot\left(\nabla\delta P_{\parallel}+\frac{\left(\delta P_{\perp}-\delta P_{\parallel}\right)\nabla B_0}{B_0}\right)\\&+\nabla\cdot\left(\frac{\delta P_{\parallel}\mathbf{b}_0\nabla\times\mathbf{b}_0\cdot\mathbf{b}_0}{eB_0}\right)-\frac{\mathbf{b}_0\times\nabla\delta B_{\parallel}}{e}\cdot\nabla\left(\frac{P_0}{B_0^2}\right)\\&-\frac{\nabla\times\mathbf{b}_0\cdot\nabla\delta B_{\parallel}}{eB_0^2}P_0=0,\end{aligned}\tag{1}$$

where n is the density, B is the magnetic field, $u_{\parallel}$ is the parallel flow velocity, and P is the pressure. The subscript 0 denotes the equilibrium quantity, and the equilibrium is prescribed in a single GTC simulation. The δ prefix means the perturbed quantity, the evolution of which is calculated through the simulation. Here, e is the elementary charge with $e=-q_e$, and $\mathbf{b}_0=\mathbf{B}_0/B_0$ is the unit vector in the direction of magnetic field line. The subscript $\parallel$ and $\perp$ refer to the component parallel and perpendicular to the magnetic field of a vector. Note that $\delta n=\delta n_e+q_i\delta n_i/q_e$ stands for the difference of ion and electron density, and $\delta u_{\parallel}=\delta u_{\parallel e}+q_i\delta u_{\parallel i}/q_e$ denotes the difference of ion and electron flow. We have $\mathbf{v}_*=\mathbf{b}_0\times\nabla\left(\delta P_{\parallel}+\delta P_{\perp}\right)/\left(n_0m_e\Omega_e\right)$, where m_e is the electron mass, and $\Omega_e=eB_0/m_e$ is the electron cyclotron frequency. The perturbed electron parallel flow $\delta u_{\parallel}$ can be solved from Ampere's law,

$$\delta u_{\parallel}=\frac{1}{\mu_0en_0}\nabla_{\perp}^2\delta A_{\parallel},\tag{2}$$

where μ_0 is the permeability of vacuum. $\delta A_{\parallel}$ is the perturbed vector potential. In the single fluid model, $E_{\parallel}=0$ is assumed. Then $\delta A_{\parallel}$ can be solved from

$$\frac{\partial A_{\parallel}}{\partial t}=\mathbf{b}_0\cdot\nabla\phi,\tag{3}$$

and the electrostatic potential ϕ can be solved from gyrokinetic Poisson's equation (the quasi-neutrality condition)

$$\frac{c^2}{v_A^2}\nabla_{\perp}^2\phi=\frac{e\delta n}{\epsilon_0},\tag{4}$$

where c is the speed of light, v_A is the Alfvén velocity, and ϵ_0 is the dielectric constant of vacuum. The parallel magnetic perturbation δB is given by the perpendicular force balance,

$$\frac{\delta B_\parallel}{B_0} = -\frac{\beta_e}{2}\frac{\delta P_\perp}{P_0} = -\frac{\beta_e}{2}\frac{\partial P_0}{\partial \psi_0}\frac{\delta \psi}{P_0}. \tag{5}$$

The perturbed pressure in the fluid limit can be calculated by

$$\delta P_\perp = \frac{\partial P_0}{\partial \psi_0}\delta\psi - 2\frac{\delta B_\parallel}{B_0}P_0, \tag{6}$$

$$\delta P_\parallel = \frac{\partial P_0}{\partial \psi_0}\delta\psi - \frac{\delta B_\parallel}{B_0}P_0. \tag{7}$$

In these equations, ψ_0 and $\delta\psi$ is the equilibrium and perturbed magnetic flux, and the evolution of $\delta\psi$ is solved from

$$\frac{\partial \delta\psi}{\partial t} = -\frac{\partial \phi}{\partial \alpha_0}, \tag{8}$$

where α_0 is from the Clebsch representation of $\mathbf{B}$ field, $\mathbf{B}_0 = \nabla\psi_0 \times \nabla\alpha_0$.

For all the 5000 simulations in DIII-D tokamak geometry, we use the $100 \times 250 \times 24$ mesh grids in the radial, poloidal and parallel direction. The time step is set to $\Delta t = 0.01 R_0/C_s$, where R_0 is the distance between magnetic axis to the geometric center of the tokamak, $C_s = \sqrt{T_e/m_i}$ is the acoustic velocity. The physical time of Δt depends specific parameters. For typical DIII-D parameters, $R_0 \approx 1.6$ m, $T_e \approx 10$ keV, so $\Delta t \approx 1.6 \times 10^{-8}$ s. The total number of simulation steps is set to 3000. Due to the free energy contained in the pressure gradient and the parallel current, the internal kink mode can be driven unstable. The volume averaged mode amplitude is measured to calculate the linear mode growth rate, and a number of snapshots during the simulation are taken to study the mode structure. In the study of kink instability, only one toroidal mode number $n = 1$ is kept in the simulations, among which 1972 cases of the simulated kink modes are identified as unstable. Several empirical methods are used to peel off the cases with large numerical noise, and we have a total number of 1605 used to train the ML model. The cases with stable kink modes cannot show physical kink mode structure, therefore not investigated in this paper.

Apart from the fluid simulations of the mentioned dataset, we also ran the linear and nonlinear gyrokinetic simulation of DIII-D discharge #141216 at $t = 1750$ ms for this paper. This shot was selected after meticulous calibration of measurements performed by the experimentalist, and the GTC simulation for this shot has been well benchmarked with several other codes [49]. We use the ion Vlasov equation to solve the perturbed ion distribution function δf_i and the electron continuity equation to solve the electron perturbed density δn_e, while other field equations are similar to the single fluid model. More details are specified in [48]. The geometry is discretized with the same grid number of $100 \times 250 \times 24$ in radial, poloidal, and toroidal directions; the time step is set to $\Delta t = 0.005 R_0/C_s = 1.483 \times 10^{-8}$ s. We run 20 000 steps for the nonlinear simulation, in which we can clearly see the nonlinear saturation and evolution. We keep both the $n = 1$ and $m = n = 0$ modes in the nonlinear simulation, and the toroidal variation of mode structure is determined by that at $\zeta = 0$ with a phase shift $\exp^{i\zeta}$. For the linear simulations, the zonal flow component is zero. Therefore, the radial-poloidal mode structure at $\zeta = 0$ provides sufficient information to represent the mode structure across the entire domain. In the linear simulation, all nonlinear terms are forced to be 0, and we run 60 000 steps until we see the clear periodic behaviors. The high-fidelity nonlinear GTC simulation for this shot provides the dataset utilized by the FTL model to learn the nonlinear plasma dynamics. For the downstream analysis, we standarize the poloidal grid resolution across all simulations by resampling to a unifrom 101×180 grid in radial-poloidal $\psi - \theta$ coordinate with interpolation applied to the field data. In this paper, we focus on analyzing the dynamics of electrostatic potential mode structures.

3. Methods

Consider a full-order model of a nonlinear dynamical system characterized by a system of ordinary differential equations ,

$$\dot{\mathbf{V}} = \boldsymbol{f}(\mathbf{V}(\mathbf{x}, t; \mu)), \tag{9}$$

$$\mathbf{V}(\mathbf{x}, 0; \mu) = \mathbf{V}^0(\mathbf{x}, \mu), \tag{10}$$

where t denotes time in $\mathbb{R}_+$, $\mathbf{x}$ denotes the coordinate, $\mathbf{V}$ denotes the state variable, $\mu \in \mathcal{D}$ denotes the system parameters, and $\mathbf{x}^0(\mu)$ is the parameterized initial condition. The mapping $\boldsymbol{f}: \mathbb{R}^N \times \mathbb{R}_+ \times \mathcal{D} \to \mathbb{R}^N$ denotes the dynamics of the system.

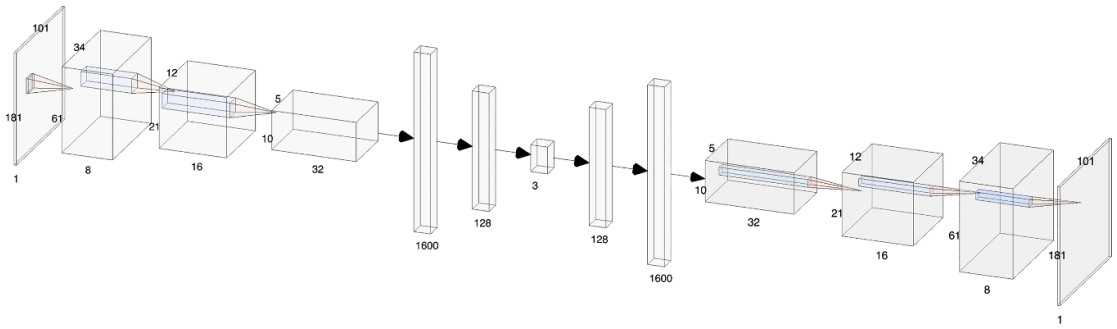

Figure 2. The architecture of the FTL network used in this work that consists of convolutional, fractionally-strided deconvolutional blocks and fully-connected layers built in an autoencoder. The architecture hyperparameters are tuned to optimize the validation performance of the kink mode structures.

3.1. encoder–decoder network

In the proposed model, we deploy convolutional, fractionally-strided deconvolutional blocks and fully-connected layers built in an autoencoder architecture. The convolutional kernels employed in the network are adept at capturing spatial correlations while extracting hierarchical patterns from training data. The encoder layers constrain coherent spatial patterns in the following bottleneck layers. The operator $\Phi : \mathbb{R}^N \rightarrow \mathbb{R}^d$ maps the state $\mathbf{V}$ to a low-dimensional embeddings $\mathbf{Z}$ through convolutional kernels and a feed-forward neural network. The constructed latent space encodes the spatial structures while preserving the dynamical characteristics of the system when trained on a series of snapshots. The dimension of the latent space d is a hyperparameter that can be evaluated based on the complexity of the system and trade-off between compression ratio and reconstruction quality. The decoder Ψ network reconstruct the original input from the encodings through fully-connected layers and deconvolutional kernels, $\Psi : \mathbb{R}^d \rightarrow \mathbb{R}^N$. Batch normalization layers are used as a regularizer to improve the training stability and prevent overfitting. We choose the rectified linear unit activation function mapping the nonlinearities among the hidden layers, and the hyperbolic tangent function to constrain the output variables to the range of $[-1, 1]$. Figure 2 illustrates the architecture of the FTL model employed for the kink mode study. The network consists of three convolutional layers that sequentially reduce the spatial dimensions while increasing the number of channels: $1 \rightarrow 8 \rightarrow 16 \rightarrow 32$, utilizing a kernel size of 3×3 with batch normalization. The output from the last convolutional layer is flattened and mapped by a fully connected layer to an intermediate latent dimension of 128, followed by a hyperbolic tangent activation, which subsequently projects the latent representation into the final encoded space dimension. The architecture hyperparameters, including layer dimensions, kernel size, padding, strided convolutions and bottleneck size are tuned to optimize the validation performance. The decoder component mirrors the structure of the encoder, systematically upscale spatial features while reducing the number of channels through the transposed convolutional layers, ensuring an accurate reconstruction of the snapshot data.

We compute the parameters of the network by minimizing the Frobenius norm of the loss function over all the batches during the stochastic gradient descent of (re)training,

$$l = \mathbb{E}\left(\|\mathbf{V}(\mathbf{x}, t; \mu) - \Psi \circ \Phi(\mathbf{V}(\mathbf{x}, t; \mu))\|_{\mathrm{F}}{}^2\right) + \lambda\|\omega\|_2^2. \tag{11}$$

As the model learns to reconstruct the input data while minimizing the reconstruction error, we set up a threshold τ based on the percentile of the distribution of the relative residual,

$$\epsilon_r(\mathbf{v}) = \frac{\|\mathbf{v}(\mathbf{x}; \mu) - \Psi \circ \Phi(\mathbf{v}(\mathbf{x}; \mu))\|_{\mathrm{F}}}{\|\mathbf{v}(\mathbf{x}; \mu)\|_{\mathrm{F}}}, \tag{12}$$

and separate normal samples from anomalies. Algorithm 1 provides a summary of the FTL offline training, including the anomaly filtering procedure. Network hyperparameters, including the learning rate η, batch size n_{batch}, maximum number of epochs n_{epoch} are tuned to optimize the performance on the validation sets. The regularization term $\lambda\|\omega\|_2^2$ penalizes large weights in the model to prevent overfitting for improving generalization performance. Adjusting the parameter λ controls the strengths of the regularization applied to the model. Early stopping is also recommended as a measure to mitigate potential overfitting. The level of anomaly filtering is determined by the threshold parameter τ. It is important to note that the filtering process can be iterative in hierarchical anomaly detection, especially when multiple levels of granularity or complexity exist within the system. This iterative approach allows for a comprehensive examination of

Algorithm 1. FTL offline training and anomaly filtering.

Input: Network architecture; training set $\mathcal{T}$; validation set $\mathcal{V}$; anomaly threshold τ; NN hyperparameters (learning rate η; batch size n_{batch}; maximum number of epochs n_{epoch}; regularization parameter λ; early-stopping criterion k).
Standardize data in the training and validation set;
Initialize iterations $i \leftarrow 0$ and initial parameters ω^0;
while $i \leqslant n_{\text{epoch}}$ **do**
  for $j = 1, \cdots, n_{\text{batch}}$ **do**
    $\omega^{i+1} \leftarrow \omega^i$; ▷ update weights
    if $l^i_{val} > \forall_{i \geqslant k} l^{i-k}_{val}$ **then**
      $\omega^* \leftarrow \omega^{i-k}$; ▷ early stopping
      break;
    end if
  end for
end while
Set $\Phi \leftarrow \Phi(\omega^*)$, $\Psi \leftarrow \Psi(\omega^*)$; ▷ save operators
for $\mathbf{v} \in \mathcal{T}$ **do**
  if $\epsilon_r(\mathbf{v}) > \tau$ **then**
    $\mathcal{T} \leftarrow \mathcal{T} \backslash \mathbf{v}$; ▷ filter out anomalies
  end if
end for
output: Encoder Φ; decoder Ψ; refined training set $\mathcal{T}$.

anomalies across different levels of abstraction, ensuring a thorough identification process. Following the removal of anomalous samples, we proceed to retrain the FTL network using the normal samples. This retraining step aims to improve the model quality in preparation for subsequent tasks. The trained model possesses transferable capabilities that enables acquiring low-dimensional representations from spatially and/or temporally relevant systems when provided with a few new samples. This function notably benefits problems with limited data availability or in the study of intricate systems where collecting a substantial number of samples is infeasible.

Algorithm 2 outlines the procedure of leveraging the pretrained FTL model to adapt to a new dataset through transfer learning. After standardizing the data, the initial parameters obtained from the trained model in algorithm 1 are utilized as the starting configuration. During training, the network weights (and biases) ω are iteratively updated to optimize reconstruction accuracy for the new data samples. Once the loss function converges to an acceptable threshold for reconstruction, anomaly filtering can be similarly applied. This transfer learning approach is particularly effective for online tasks involving spatially or temporally correlated snapshots, significantly reducing computational costs while ensuring efficient model adaptation. More importantly, the utilization of an existing trained model enables convergence acceleration, even with a minimal number of samples, which would otherwise present challenges if starting the training process from scratch from limited measurements.

4. Results

We consider a high-dimensional spatiotemporal system $\mathbf{V}(\mathbf{x}, t; \mu)$ governed by first-principle PDEs parameterized by μ, and $\mathbf{x}$ and t are spatial and temporal coordinates. Our goal is to develop a data-driven, transferable ROM that extracts plasma spatial correlations while preserving the dynamics of system in the latent space for a generalized configurations of mode structures. The simulation data used are based on the reconstructed equilibrium of DIII-D experimental discharges.

4.1. Reconstruction and anomaly detection

In this experiment, we set the free parameter $d = 3$ for the dimension of the latent space, and focus on studying the perturbed electrostatic potential in the kink modes. Generally, the structure of the kink mode exhibits variability contingent upon a number of factors including magnetic field strength, field line tilt (referred to as the 'q-profile' as a function of radial position), and the pressure profile. The ideal MHD internal kink mode is a current-driven instability with $m = n = 1$, where m and n are the poloidal and toroidal mode number. In toroidal devices (e.g. tokamaks), the internal kink mode couples with higher-order poloidal harmonics, thereby altering the stability. The mode structure peaks at the rational surface, characterized by the safety factor $q = 1$ in tokamaks. The universal feature of the kink mode is the dominant $m = 1$ component in the region of $q < 1$, accompanied by a subdominant $m = 2$ component in the

Algorithm 2. FTL online fine-tuning.

Input: Network architecture; online set $\mathcal{S}$; encoder Φ; decoder Ψ; NN hyperparameters (learning rate η; batch size n_{batch}; maximum number of epochs n_{epoch}; regularization parameter λ; early-stopping criterion k); (optional) anomaly threshold τ .

Randomly split the online set $\mathcal{S}$ into training $\mathcal{S}_{tr}$ and validation set $\mathcal{S}_{val}$;
Standardize data in the training and validation set;
Initialize iterations $i \leftarrow 0$;
Set encoder parameters $\omega_e = \omega(\Phi)$;
Initialize decoder parameters $\omega_d^0 = \omega(\Psi)$;
while $i \leqslant n_{\text{epoch}}$ **do**
  for $j = 1, \cdots, n_{\text{batch}}$ **do**
    if $l_{val}^i > \forall_{i \geqslant k} l_{val}^{i-k}$ **then**
      $\omega_d^* \leftarrow \omega_d^{i-k}$;
      break;
    end if
  end for
end while
Define $\mathbf{Z} \leftarrow \Phi(\mathbf{V}_{\mathcal{S}}(\mathbf{x}))$; ▷ encode variables
$\hat{\mathbf{V}}_{\mathcal{S}}(\mathbf{x}) \leftarrow \Psi(\omega_d^*) \circ \Phi(\mathbf{V}_{\mathcal{S}}(\mathbf{x}))$; ▷ decode embeddings
procedure OPTIONAL Anomaly set $\mathcal{A}$
  for $\mathbf{v} \in \mathcal{S}$ **do**
    if $\epsilon_r(\mathbf{v}) > \tau$ **then**
      $\mathcal{A} \leftarrow \mathcal{A} \cup \mathbf{v}$;
    end if
  end for
end procedure

Output: Embeddings $\mathbf{Z}$; reconstructed snapshots $\hat{\mathbf{V}}_{\mathcal{S}}(\mathbf{x})$; (optional) anomaly set $\mathcal{A}$.

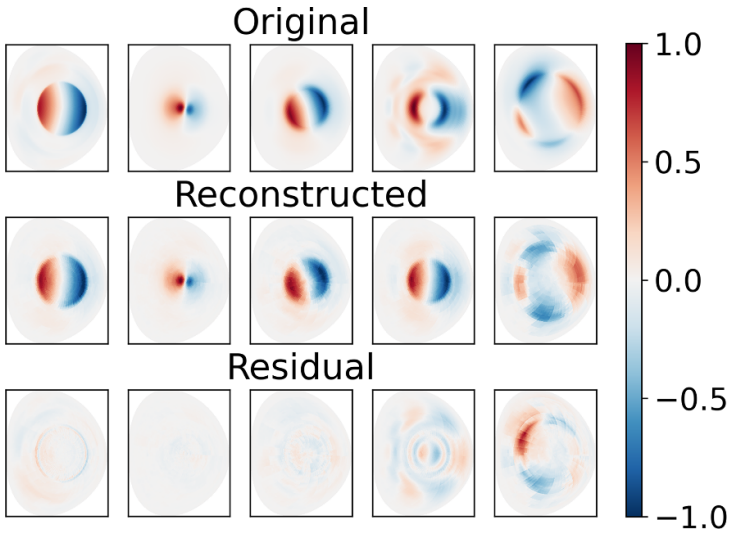

Figure 3. Comparison between the original and reconstructed kink modes for five samples from the test data. A larger reconstruction residual is observed for atypical mode structures.

$1 < q < 2$ region. The $m = 0$ component remains constant when approaching to the magnetic axis. We focus our study on the dominant $m = (0, 1, 2)$ components in this paper while neglecting higher-order components for physical analysis.

The specifics of the GTC simulation setup are detailed in section 2.2. In all the simulations, the initial conditions are set to $\delta n_e(t = 0) = n_{\text{init}}(r)\cos(\theta - \zeta)$ with $n_{\text{init}}(r)$ the radial envelope. The envelope peaks at $r = a/2$, where a denotes the minor radius at the plasma wall. And all other quantities are set to 0 initially. We train our FTL model using 1605 samples of simulated kink modes obtained from GTC simulations, and test it on another set of 201 snapshots from unseen data. Note that there is no temporal correlation between the kink mode samples studied in this section, and we investigate dynamical behaviors in section 4.2. Figure 3 shows five examples of the reconstructed kink modes and their corresponding residuals relative to

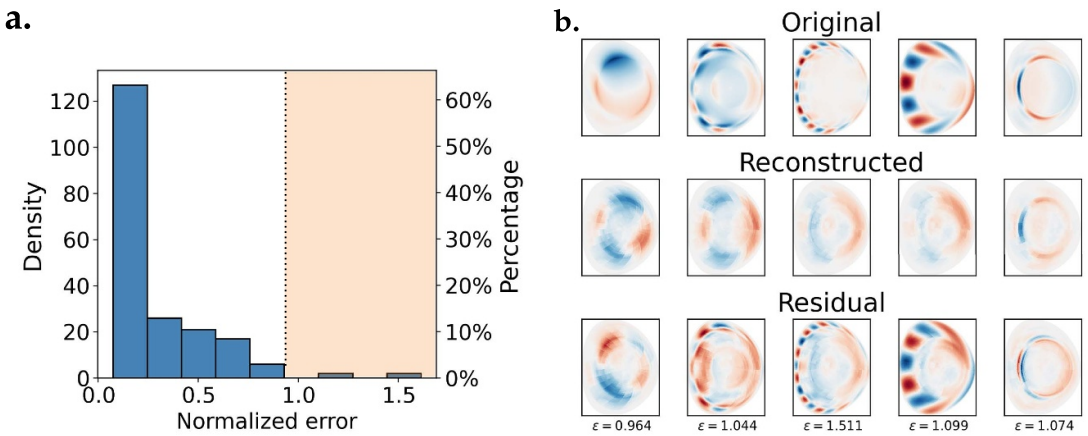

Figure 4. Detected anomalous modes using trained FTL network on the test set: (a) a histogram of sample density vs. normalized error highlighting anomalies in the tail of the distribution, (b) the original, reconstruction, and residual for five kink mode samples in the highlighted in the anomalous (shaded) region of (a), threshold parameter $\tau = 0.95$.

the original data. General kink mode structures including mode radial location and the different phase angles are well captured via the decoder. The reconstruction error is bigger for atypical structures, such as a dominant $m = 2$ structure or a dominant structure near the edge. Unstable kink modes are observed in the first four simulations. The fourth snapshot shows a less common scenario as the 'reverse-shear' equilibrium. With increasing radius, the q-profile first decreases to a minimum value ($q_{\min}$), and then increases. If $q_{\min} < 1$, there exist two radial positions at which $q = 1$, and the $m = 1$ structure is mainly located between the two $q = 1$ points. The FTL model discovers the structure at the region outside the $q = q_{\min}$ flux surface and disregards the structures near the magnetic axis. The fifth instance is a stable case where typical kink structures are absent, resulting in a larger reconstruction error as anticipated.

As the kink modes are produced from GTC simulation, certain output samples are subjected to subsequent reevaluation and filtration processes. Certain anomalous (non-physical) modes may be caused by an unrealistic profile or an unconverged EFIT equilibrium used as the input of a GTC simulation. By undergoing extensive training using a substantial dataset comprising numerous kink modes obtained from simulation, the FTL model demonstrates proficiency in reconstructing typical mode structures. This capability enables it to function as an effective filtering mechanism for identifying anomalous modes by examining the deviations when comparing the decoded outputs and original modes, as elucidated in algorithm 1. To construct the training dataset, we have cleaned the data via expert labeling and automatic scripts using empirical criteria to exclude the stable or numerical diverging cases. Figure 4(a) presents a histogram illustrating the distribution of normalized reconstruction errors. More than 60% of the test samples resemble the input data with a relative error smaller than 0.2. The few samples collected in the shaded region that have relatively larger reconstruction errors are shown in figure 4(b). One of the prevalent anomalies observed is the edge effect, characterized by the aggregation of small-scale structures at the boundary of the poloidal plane. We find these cases are not the physical modes yet failed to be identified by the previous empirical methods. Similarly, the model FTL effectively reconstructs the multi-scale structures of AEs by preserving a higher dimensional latent space. Anomalous AEs are identified through algorithm 1, as illustrated in figure A2.

This outcome demonstrates FTL's capacity to discern pollution within the collected data, thereby enhancing the performance of other models trained on the same data. Similarly, the trained model can be utilized to identify anomalous modes in real-world experiments, where unexpected structures manifest as outliers in the distribution, detected through the algorithm.

4.2. Learning nonlinear dynamics

We extend the study of kink modes derived from the linear GTC simulation in section 4.1 to a nonlinear GTC simulation with kinetic effects. Due to the inclusion of kinetic ion effect, the mode structure rotation with a finite frequency can be observed after the eigenmode is established in the linear phase. This leads to a trajectory rather than a fixed point in the latent space. We notice an abrupt tweak in the modes when the system evolves to the nonlinear phase, and the structure rotates at different speed at different radial positions (i.e. flow shear) after zonal flow is generated. The flow shear caused by the zonal flow is the main reason for the nonlinear transition and saturation of the mode amplitude. Due to the presence of flow shear, the mode

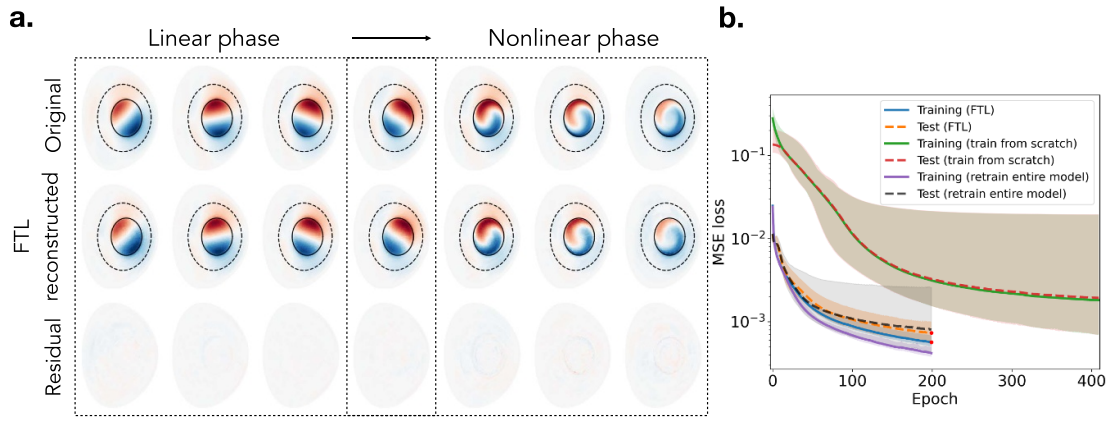

Figure 5. (a) Simulated Kink modes with kinetic effects and reconstructed nonlinear mode structures at $t = 0.073, 0.135, 0.186, 0.223, 0.250, 0.272, 0.295$ ms using the trained FTL model. (b) Training and test error of the three models vs. epoch number. The solid (training) and dash (test) curves represent results averaged over using 20 different random realizations. The shaded region indicates the distribution of the variance of the reconstruction error in the multiple runs. Early stopping (red dots) is used during the NN training to prevent overfitting.

structures are tilted, accompanied by variations in the relative amplitude of different m components when compared to the linear stage.

Leveraging the network trained on linear modes from section 4.1, we illustrate that the FTL model is able to embed the modes within the latent space and reconstruct the modes accommodating the rotational behaviors in the linear phase. The detailed implementation of the transfer learning process is provided in algorithm 2. The nonlinear evolution of the mode structures exhibits distinctive features, for which transfer learning can be used to accelerate the training process using a limited set of data. We fine tune the decoder via retraining the FTL network on 128 snapshots randomly sampled on the data from the nonlinear simulation. Figure 5(a) shows the reconstructed modes from the test set, which are temporally correlated to the training set while unseen by the retrained network. The decoded mode structures present strong consistency with the original data, displaying only minor discrepancies. As seen in figure 5(b), the mean-square-error (MSE) loss shows rapid convergence during the network training process for both the training set and test sets using FTL, reaching values below 1×10^{-3} following 100 epochs. Early stopping is employed to prevent overfitting for generalized performance. The convolutional encoder explores the manifold through learning the variances of the mode structures over a large number of samples, and provides approximate optimal embeddings in such latent space that significantly reduce the online learning time for MHD instabilities. The fine-tuned FTL model then adeptly reproduces the mode rotation in the linear stage, captures the structure variation at the transition point, and accurately depicts the mode structure evolution. We compare the reconstruction accuracy of our model with that of principle component analysis (PCA) in appendix. Notably, our encoder–decoder architecture in FTL demonstrates more stable performance by approximately optimizing the low-dimensional embeddings across the phase transition from linear to nonlinear regimes. In contrast, PCA tends to prioritize the reconstruction of dominant linear mode patterns, on which its principle component basis is constructed, at the expense of accurately capturing the patterns of nonlinear modes.

4.3. Computational evaluation

We evaluate the accuracy and computational cost of the FTL model in reconstructing modes and compare its performance to two other ML models: (i) a Baseline model, which involves training an encoder–decoder model from scratch using the same network architecture, and (ii) an Enhanced model, which retrains the encoder–decoder model using the previous trained model obtained on the equilibrium data in section 4.1. Leveraging the pre-trained model, both the Enhanced model and FTL model exhibit more rapid loss convergence compared to the Baseline model. Additionally, FTL outperforms the Enhanced model at the same number of training epochs $n_{\text{epoch}} = 200$, by achieving a smaller test error and less uncertainty variation in the reconstruction results. This can be attributed to the increased generalizability gained from the pre-trained encoder when trained with a bigger training set, while the retrained autoencoder tends to overfit to a smaller size of training set, leading to higher error when the model is evaluated on hold-out data. The training and test errors, as well as total trainable parameter counts of the three models are presented in table 1. As the number of trainable parameters drops by 49.8%, a significant amount of computational workload and memory usage is reduced during back-propagation and optimization, saving the storage requirements for the gradients and optimizer states. FTL demonstrates advantageous performance compared

Table 1. Comparison of the evaluation of three ML models (Baseline, Enhanced, and FTL) on the training and test sets at a fixed number of epochs, $n_{\text{epoch}} = 200$. Both the mean (μ) and standard deviation (σ) of the mean-square error (MSE) are presented for the training and test loss. The last column reports the total number of trainable parameters in each model.

Model	Training MSE		Test MSE		Trainable parameter count
	μ	σ	μ	σ	
Baseline model	3.14×10^{-3}	4.19×10^{-3}	3.27×10^{-3}	4.21×10^{-3}	424 052
Enhanced model	4.19×10^{-4}	2.97×10^{-5}	8.05×10^{-4}	4.72×10^{-4}	424 052
FTL model	5.68×10^{-4}	5.06×10^{-5}	$\mathbf{7.32 \times 10^{-4}}$	$\mathbf{1.15 \times 10^{-4}}$	**212 817**

Note: Bold indicates the performance of the FTL model.

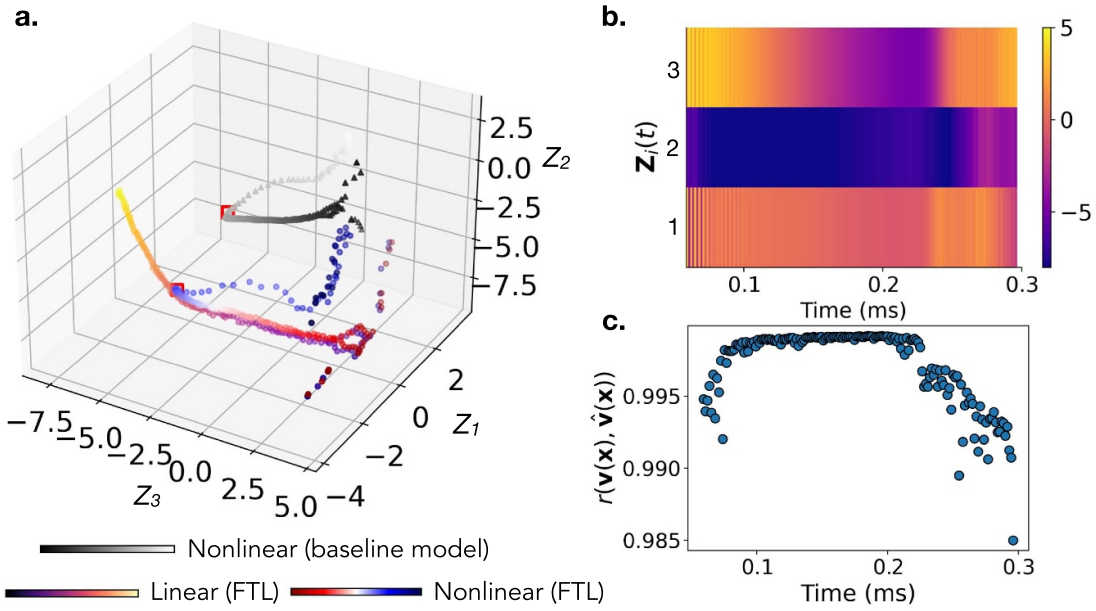

Figure 6. (a) Dynamics trajectory of the plasma state evolution in the latent space with time mapped as color. A bifurcation point (marked by red square) is observed when the modes transit from linear to nonlinear. The nonlinear trajectory developed from FTL shows a similar pattern with the same transition point when compared to the result obtained from the baseline model. (b) latent space vector vs. time. (c) Correlation coefficient of reconstructed modes and ground truth over the time window, $\Delta t_s = 0.5R_0/C_s = 1.483\,\mu s$.

to the other two models in terms of reconstruction accuracy, stability, and computational efficiency during the online training phase. Although the encoder–decoder framework is general for any suitable autoencoders, our findings indicate the convolutional autoencoder with regularization delivers the best performance. A comparison with variational autoencoders (VAEs) is detailed in appendix , where we investigate the effect of regularization of KL divergence on reconstruction accuracy.

4.4. Bifurcation analysis

As the kink instabilities progress from the linear to nonlinear phase, significant alterations occur in both the characteristics of the mode structures and the rotational state. We investigate the temporal evolution of the kink modes within the low-dimensional embeddings, $\mathbf{Z}(t) = \mathbf{\Phi}(\mathbf{V}(\mathbf{x}, \mathbf{t}; \boldsymbol{\mu}))$ where the spatial domain is compressed to $\mathbb{R}^d$ for each snapshot.

Specifically, we study two cases including one simulation derived from the linear GTC simulation with kinetic effects and the other from the GTC simulation with kinetic and nonlinear effects. The linear and nonlinear simulations start at identical initial conditions. Our investigation focuses on the bifurcation point in the temporal evolution of the kink modes, with the exclusion of the first snapshots under initial effects. Although the mapping between the latent space coordinates and physical space structures is nonlinear through the encoding layers, which results in the complexity in attributing the effect of a single coordinate, it remains interesting to visualize and examine the relationship between the trajectories in the latent space and their real-space counterparts. Figure 6(a) shows the dynamic trajectories of the kink mode evolution in the three-dimensional embedded space, highlighting where the bifurcation occurs between the linear and nonlinear phase of the mode evolution. In the linear stage of the nonlinear simulation, the effect of the nonlinear terms is small, resulting in a system that closely resembles the linear simulation. This can be confirmed by the overlap of the first half of the linear and nonlinear trajectories.

In the early stage of the simulation, different eigenmodes co-exist and the trajectory shows how the dominant kink eigenmode structure prevails. The compact oscillation between $t = 30R_0/C_s$ and $t = 35R_0/C_s$ may correspond to a physical mode characterized by high frequency and low growth rate. Here, R_0/C_s is used as the time unit, where the major radius at the magnetic axis is $R_0 = 1.78\,\mathrm{m}$, and the ion sound speed is $C_s = 6.01 \times 10^5 \mathrm{m\,s^{-1}}$. The subsequent smooth evolution in both the trajectories corresponds to the rotation of the most unstable eigenmode in the physical space. When the bifurcation occurs at $t = 76.5R_0/C_s$, the linear trajectory continues to evolve on the consistent path while the nonlinear case deviates from the track sharply and starts to exhibit disturbed behaviors in the latent space. Simultaneously, we observe a transition in the mode rotation direction in the physical space of the nonlinear simulation, shifting from clockwise to counterclockwise at the tipping point. Note that a closer examination in the latent space in figure 6(a) shows that FTL resembles the evolution in the low-dimensional trajectory of the baseline model whose autoencoder is trained from scratch without any a prior embedding information. The next change in the direction of the nonlinear trajectory may correspond to an increase of the mode tilt. At the later stage, the counterclockwise rotation transits to a minor oscillatory mode where we observe a 'reverse' in the trajectory. Overall, examining the dynamics of the latent space proves highly illustrative, revealing distinct physical stages that may not be readily discernible in the real physical space.

The FTL model effectively encodes the coherent mode structures while retaining the intrinsic dynamics within the learned manifold. Figure 6(b) illustrates the dynamic changes of the embeddings for $\mathbf{Z}_i$ over the time domain. The first two components, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ represent the positive and negative variations, respectively, of the latent variables; the third component $\mathbf{Z}_3$ captures the transitional points of the instabilities, i.e. when the modes evolve to a stable rotational state at $t = 0.10$ ms after the initial perturbations, and the transitional point at bifurcation. The Pearson correlation coefficient r computed between the original snapshot $\mathbf{V}(\mathbf{x})$ and corresponding reconstructed data $\hat{\mathbf{V}}(\mathbf{x})$ across the entire temporal domain is depicted in figure 6(c), with each snapshot data being centered and scale-invariant,

$$r = \frac{\sum_{i=1}^{n}\left(\mathbf{V}(\mathbf{x})_i - \overline{\mathbf{V}(\mathbf{x})}\right)\left(\hat{\mathbf{V}}(\mathbf{x})_i - \overline{\hat{\mathbf{V}}(\mathbf{x})}\right)}{\sqrt{\sum_{i=1}^{n}\left(\mathbf{V}(\mathbf{x})_i - \overline{\mathbf{V}(\mathbf{x})}\right)^2}\sqrt{\sum_{i=1}^{n}\left(\hat{\mathbf{V}}(\mathbf{x})_i - \overline{\hat{\mathbf{V}}(\mathbf{x})}\right)^2}}, \tag{13}$$

where $\overline{\mathbf{V}(\mathbf{x})}$ and $\overline{\hat{\mathbf{V}}(\mathbf{x})}$ refer to the mean of original snapshots and reconstructed data correspondingly. A high linear correlation is observed when eigenmode is formed in the well-defined linear phase, and the coefficient r begins to decrease from 0.998 to 0.985 subsequent to the transition point. The linear eigenmode structure is relatively easy to estimate, given the nature of its simple dynamics while we utilize the FTL model to fine tune with temporally correlated data. The decrease in the correlation aligns with the emergence of mode structures predominantly influenced by the nonlinear effects. Nevertheless, despite this reduction, the reconstruction quality remains acceptable.

4.5. Towards small or sparse datasets

In this section, we evaluate the performance of the FTL model on a limited number of snapshots, which targets real-time tasks for online learning. This is crucial for applications where only a small dataset, characterized by a short temporal span or sparse sampling, is available and expeditious decisions-making is imperative. Extending the FTL architecture, we retrained the network on 17 snapshots collected from the simulation with a larger sampling interval $\Delta t_s = 5R_0/C_s$, while maintaining the same initial conditions as in section 4.2.

Figure 7(a) shows the original and reconstructed snapshots of the 2D kink mode structures, capturing the transitional behaviors at the bifurcation point and the subsequent nonlinear evolution of modes. The fine-tuned FTL network effectively reconstruct the coherent structures in both the linear and nonlinear phases, exhibiting minor degradation in clarity and sharpness. The Fourier decomposition provides a more quantitative comparison; the Fourier coefficients in figure 7(b) show that the FTL reconstruction qualitatively preserves the amplitude and phase of the first three Fourier harmonics. Although the reconstructed structure exhibits minor zigzags, their impact on the physical analysis is negligible. It is remarkable that the reconstructed structures reproduce the critical changes of the Fourier components at the magnetic axis point, i.e. $q = 1$ and $q = 2$ surfaces precisely, despite that a small amount of loss is observed in the higher order mode.

4.6. Understanding tipping points in latent space

Interpretation the tipping points in the latent space involves understanding significant transitions or critical thresholds that occur within the underlying data representation. These tipping points may signify abrupt physical state changes, phase transitions, or critical events in the system being studied. Analyzing these points

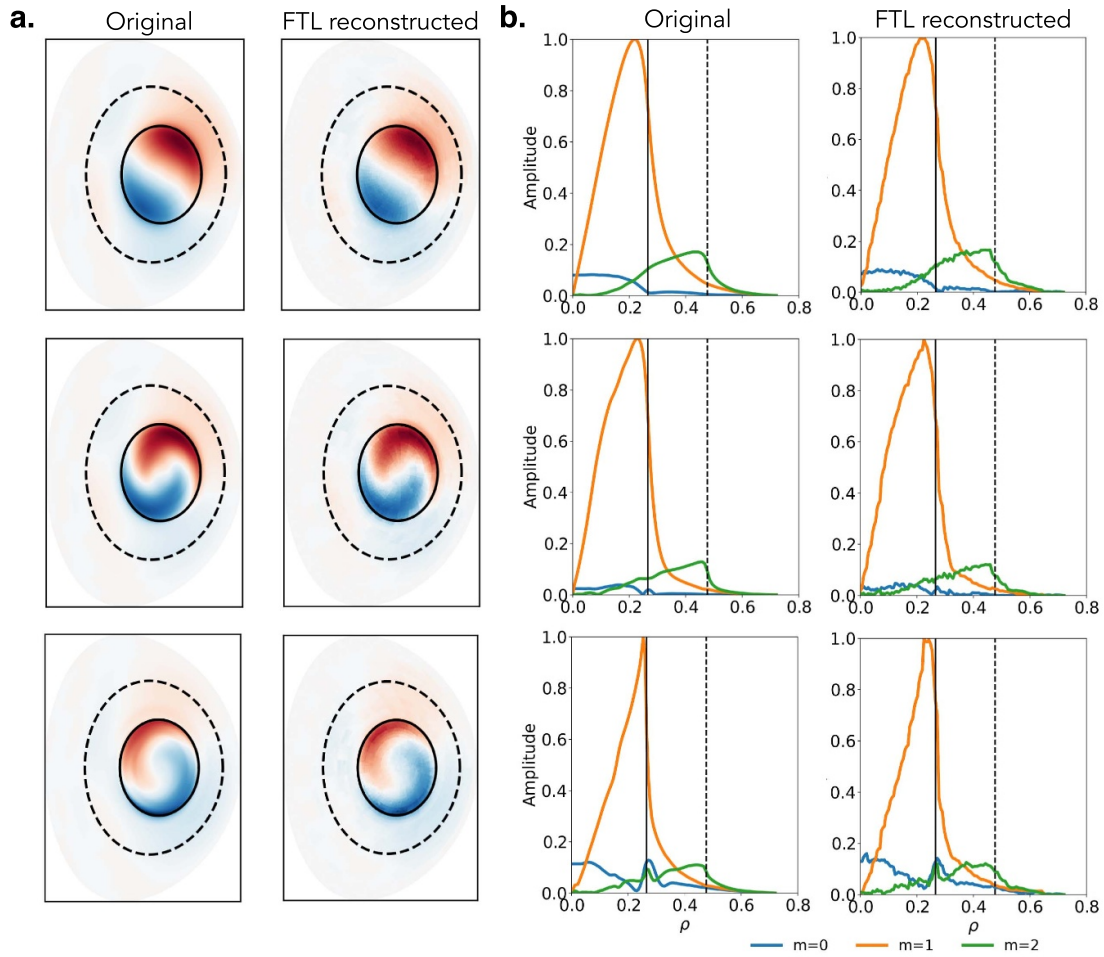

Figure 7. 2D mode structures and radial structures of the Fourier modes elucidating the transition from linear to nonlinear phase: (a) the original snapshots and reconstructed snapshots at $t = 0.222, 0.237, 0.281$ ms, (b) the amplitude of first three Fourier coefficients ($m = 0, 1, 2$) along the radius on the poloidal plane at the respective time. The solid and dash lines mark the q profiles of $q = 1$ and $q = 2$ along the radius of the poloidal plane.

can provide insights into the inherent dynamics, identify regime shifts, or reveal features of importance in the data distribution.

We compare the first three Fourier modes in both their real and imaginary parts in figure 8. Periodicity is observed in all the five components: the real parts of $\mathscr{F}^0(\mathbf{v}), \mathscr{F}^1(\mathbf{v}), \mathscr{F}^2(\mathbf{v})$, and the imaginary parts of $\mathscr{F}^1(\mathbf{v}), \mathscr{F}^2(\mathbf{v})$. Figure 9 illustrates the latent space trajectory associated with a linear simulation spanning an extended time evolution. The mode structure in the physical space exhibits a rotational period, which in the latent space corresponds to a closed loop. In traditional methods, an eigenmode can be represented by separating the oscillation over time, the toroidal harmonics, and the poloidal harmonics by carrying out a Fourier transformation. The FTL embeddings provide a low-dimensional representation of the eigenmode, i.e. a visualization of the linear loop trajectory in the encoded latent space. The trajectory in figure 9 also exhibits dynamical characteristics in addition to the periodic mode rotation. For example, the abrupt direction changes observed at the two flagged points relate to the amplitude of the standing wave component in the physical space.

The identified tipping points in the latent space are depicted in both physical space and the Fourier domain, demonstrating their correspondence. Figure 10(a) depicts the tipping points observed in the latent space are close to normalized spatial maxima and minima for all the snapshots in the extended linear simulation. The phase shift occurs at approximately $t = 0.35$ ms and $t = 0.65$ ms in the Fourier domain, as in figure 10(b), and signifies the transition point where the real-space structure exhibits a distinct change. Specifically, this shift indicates the moment when the pure cosinusoidal wave $\cos(m\theta)$ becomes predominant, while the sinusoidal pattern diminishes. The potential can be generally expressed as

$$\mathbf{v} = \boldsymbol{\phi}(\rho, \theta, t) = \sum_{m=0} \left[\left(\hat{\phi}_{R,m}(\rho, t) + \mathrm{i}\hat{\phi}_{I,m}(\rho, t) \right) \mathrm{e}^{\mathrm{i}m\theta} + \left(\hat{\phi}_{R,m}(\rho, t) - \mathrm{i}\hat{\phi}_{I,m}(\rho, t) \right) \mathrm{e}^{-\mathrm{i}m\theta} \right], \tag{14}$$

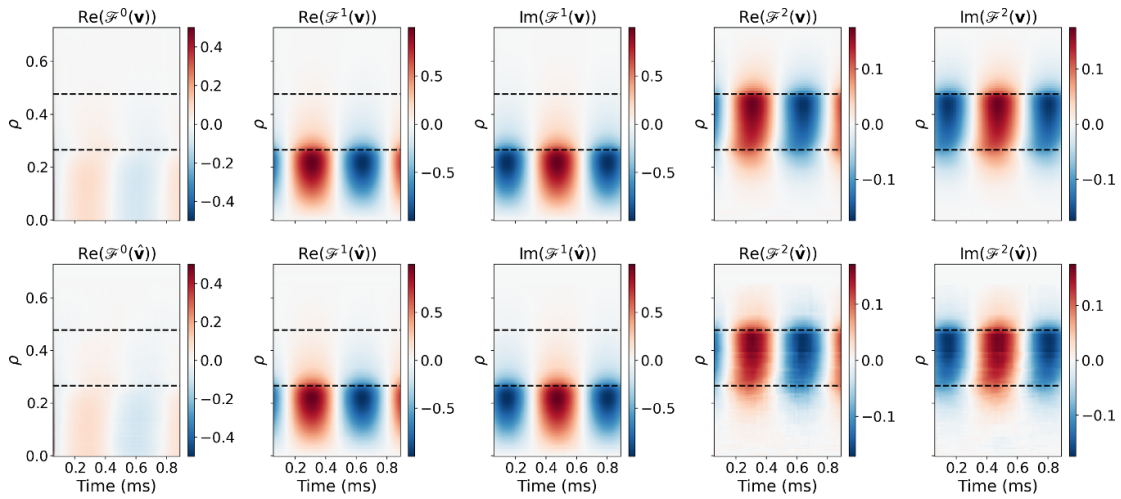

Figure 8. Real and Imaginary parts of the Fourier modes for all the snapshots. The reconstructed results (bottom) of the Fourier modes shown good alignment with those obtained from the original GTC data (top). The dash lines mark the q profiles of $q = 1, 2$ along the radius of the poloidal plane.

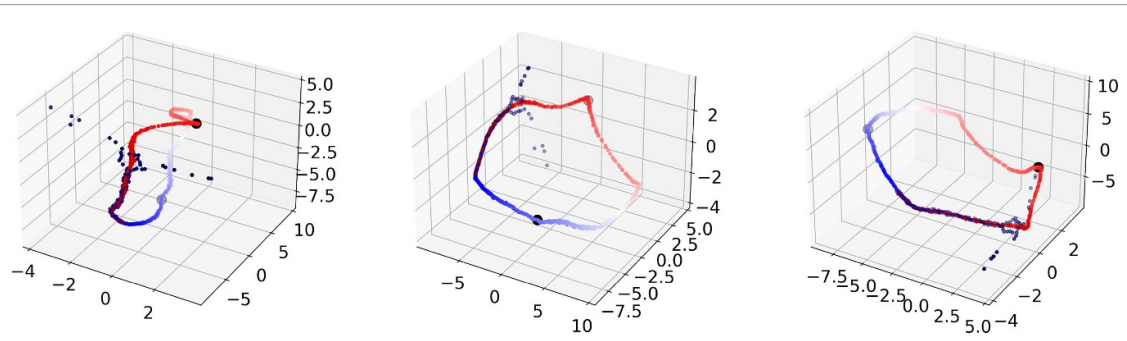

Figure 9. Dynamic trajectory of one linear case evolution displayed from three different perspectives in the latent space. At $t = 0.665$ ms, a complete evolution period is observed, culminating in the formation of a closed loop. Two tipping points at $t = 0.355, 0.651$ ms are highlighted as large black circles, denoting points of notable inflection observed in the evolutionary trajectory.

where $\phi(\rho, \theta, t)$ is a real number, and the subscripts R and I indicate the real and imaginary parts. For a certain mode number m, when the imaginary Fourier component is 0 at $t = t_0$, we have $\phi(\rho, \theta, t_0) = 2\hat{\phi}_{R,m} \cos(m\theta)$. Note that in general $\hat{\phi}_{I,m}(\rho, t)$ may not be 0 for all r and all m numbers. In this particular case, the imaginary parts vanish almost altogether near $t = 0.35$ ms and $t = 0.65$ ms, thus creating the two special tipping points.

5. Conclusion and discussion

We have developed FTL, an encoder–decoder based transfer learning model to represent high-dimensional snapshots of MHD modes in low-dimensional embeddings. Our approach has demonstrated its effectiveness in mode reconstruction, anomaly detection, and revealing transitional dynamical features for unstable kink modes. Compared to the Baseline and Enhanced models, the proposed FTL delivers greater stability and higher reconstruction accuracy while using less computation and memory. The transferable architecture of FTL, capable of extracting essential information from spatially and/or temporally relevant problems, even with limited datasets, enhances the model's adaptability and efficiency to a spectrum of MHD problems. The analysis of plasma dynamics in the latent space enables correlations of dynamic transitions with real and frequency domain behaviors, and adds a layer of physical interpretability to the FTL model.

Limitations and future directions. Despite the substantial advances enabled by the FTL model, there are several aspects to be improved upon. Whereas this work primarily focuses on mode structures, with snapshot amplitudes being normalized to enhance anomaly detection capabilities, future work will endeavor to incorporate amplitude as well as their growth rates into the FTL model. Moreover, we envisage leveraging the versatile FTL architecture to encompass multiple physical modes including NTMs thereby expanding the model's capability. One limitation lies in generality of the current hyperparameter settings of FTL's

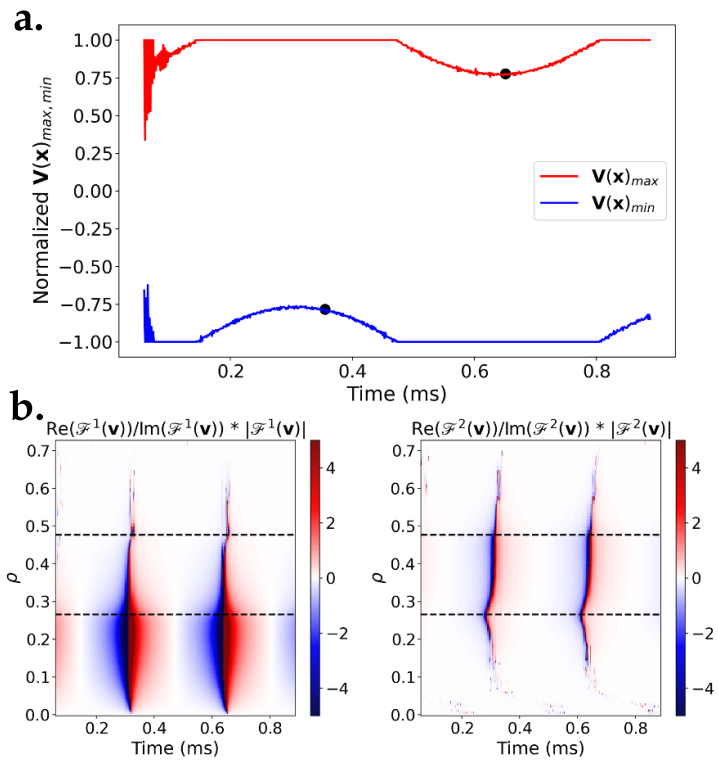

Figure 10. The correspondence of the identified tipping points (latent space) in physical space and Fourier domain: (a) normalized spatial maxima and minima for each snapshot in the time domain. (b) the ratio of the real and imaginary parts weighted by the amplitude of the Fourier modes. A phase shift is detected at the same time regime of the transition points as in figure 9.

architecture model. Each physical problem necessitates its own hyperpameter tuning, and exploration of network settings, including a network size, depth, activation functions, learning rate, optimizers, regularization methods and loss functions.

Another avenue for future direction involves addressing concerns regarding the applicability of the current FTL model to unstructured data or irregular geometries, as it is currently designed for structured data with regular geometry. We note that FTL can be extended by replacing the convolutional kernels with graph convolutions in graph neural network in order to process irregular data represented as graphs. For specific physics-constrained problems, the loss functions can be formulated with the residual component with respect to the first-principled PDEs, as demonstrated in the physics-informed neural networks [50, 51]. Additionally, we envision a hybrid approach that integrates FTL with other surrogate models, e.g. SGTC, to develop data-driven embedding methodologies for inferring plasma dynamics. The collaborative framework aims to accurately predict mode structures and their growth rate from constrained encoding, facilitating a deeper understanding and more precise forecasting of disruptions in large-scale plasma systems.

Ultimately, we look to extend FTL and develop predictive models to plasma transitions for active plasma control through data assimilation. The anticipated future work involves model predictive control [52, 53] built on FTL and its application in real tokamak experiments, wherein the ongoing study of bifurcation prediction in complex systems necessitates comprehending how system behavior evolves with varying parameters. As bifurcations mark significant shifts in system dynamics, accurate prediction entails a thorough exploration of parameter space to pinpoint transitional regimes. This endeavor will reveal critical conditions triggering bifurcations, providing a comprehensive understanding into the complex systems and guiding strategies for control or mitigation.

Data availability statement

All data needed to evaluate the conclusions in the paper are present in the paper. Additional data related to this paper are available from the corresponding author upon reasonable request. The source codes that were used for the findings of this study will be openly available in the GitHub at https://github.com/zhbai/FTL [54].

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

Z B, W T, L O and S W gratefully acknowledge support from the U.S. Department of Energy, Office of Science, SciDAC/Advanced Scientific Computing Research under Award Number DE-AC02-05CH11231. X W, W T and Z L acknowledge support from the U.S. Department of Energy, SciDAC/Fusion Energy Sciences and the INCITE award. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract Number DE-AC02-05CH11231.

Conflict of interest

The authors declare no competing interests.

Appendix. Supplementary results

A.1. Benchmarking results using PCA

As a benchmark, we perform principal component analysis (PCA) [55] on the kink modes, and determine the rank r threshold based on the convergence of energy over an increasing number of PCA modes [56, 57]. In figure A1(a), the singular values indicate the energy of each component. We keep $r = 3$ principle components, which contain 97% of the total variances of the kink mode data. In contrast to autoencoders, which capture mode variances by imposing constraints on the latent space through nonlinear mapping using NNs, PCA projects the data onto a basis formed by the dominant eigenvectors of its covariance matrix. Figure A1(b) shows the relative $\mathcal{L}_2$ error of the PCA and FTL reconstructions over the entire time window as in figure 6. We observe that FTL demonstrates greater performance stability than PCA, especially in the nonlinear regime following the phase transition, where it effectively reduces the relative L_2 error from 0.3 to 0.1. With an equivalent number of latent variables, FTL exhibits a superior capacity in capturing the variances of mode structures when compared to linear projection-based method such as PCA while identifying the bifurcation point.

A.2. Reconstruction of AEs

We extend the FTL model to reconstruct AEs collected from GTC simulation. AEs are a kind of electromagnetic oscillations commonly seen in neutral beam injection (NBI) heating plasmas. The AE instability can be driven by the wave-particle resonance of fast ions (alpha particles or NBI generated ions). The dataset was originally collected from experiments to predict the excitation of different AE types [58]. These experimental shots have been simulated by GTC to study the instability drive and fast ion transport. The AE mode structures exhibit significant differences from kink mode structures. The radial mode location does not depend on the local q value. Additionally, the poloidal mode number m is generally about 10 for DIII-D tokamak. Except for the beta-induced AEs and reversed-shear AEs , most of the modes in the dataset exhibit strong ballooning structures (coupling between different m numbers). The extension to AEs demonstrates the generalization capabilities of our model.

Specifically, we apply the model to a total of 533 AE equilibrium cases, which are randomly divided into 426 training samples and 107 test samples. Due to the presence of more complex structures on the poloidal plane characterized by dense concentrations near the edge, we lift the latent dimension to $d = 55$, guided by the accumulated energy retained in the system as determined through PCA. Figure A2 illustrates the GTC simulated AEs, reconstructed AEs, and the residuals from the reconstruction using our model. Notably, the multi-scaled structures are captured in the low-dimensional embeddings and then decompressed through the decoder network. Overall, the FTL model demonstrates good reconstruction performance for most AE cases, particularly on the low field side (right half of the poloidal plane), where the large eddies are prevalent. We notice that the high relative error in the high field side (left half of the poloidal plane) can be attributed to the presence of finer mode structures. Physically, these fine structures are caused by the phase shift of different m harmonics at varying radial locations. However, the limited size of the dataset may not provide sufficient features for the model to fully capture the variances and effectively learn these small-scale structures in AEs. We also propose that incorporating mathematical expressions of the mode structures into the loss function may enhance the FTL reconstruction of the fine structures. Further improvement of the AE reconstruction involves physics-informed terms for regularization, and will be studied in future work.

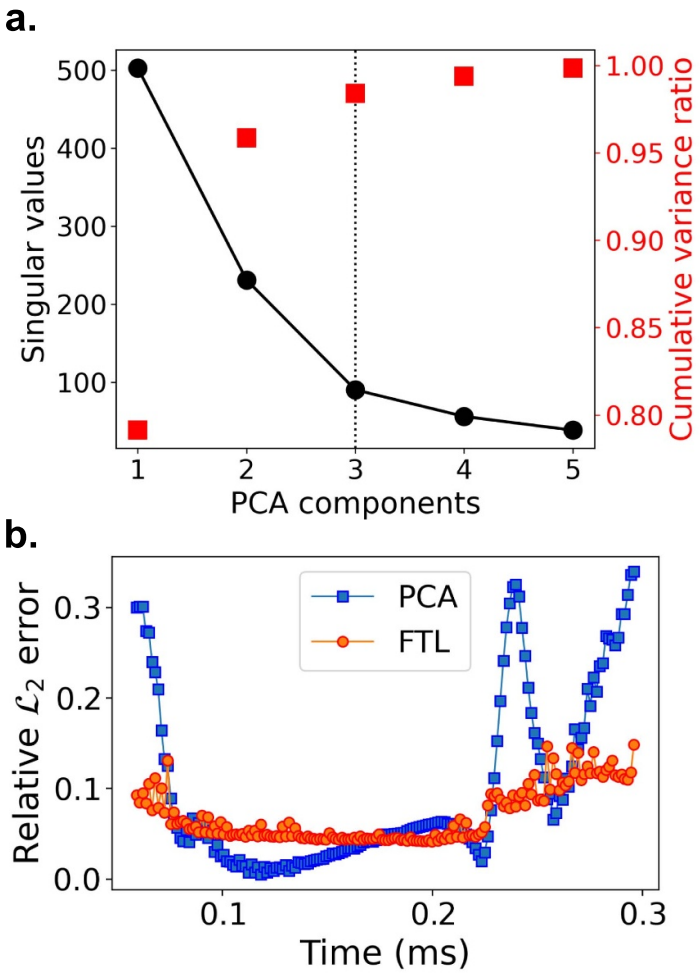

Figure A1. Principal component analysis (PCA) performed on the kink instabilities. (a) singular values and cumulative variance ratio vs. the principal components. (b) relative $\mathcal{L}_2$ error vs. time for the reconstruction accuracy comparision between PCA and our apporach FTL.

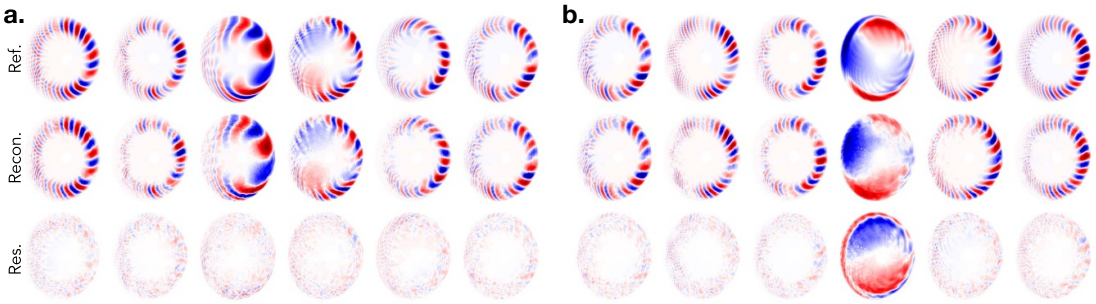

Figure A2. Original, reconstructed Alfvén eigenmodes and the corresponding residuals using our approach FTL: (a) training set, (b) test set. The fourth case in panel (b) is an anomalous sample identified based on the residual detected following algorithm 1.

A.3. Relaxation of KL divergence in VAEs

We present the results of beta-VAEs (β-VAEs) [59, 60], exploring the effects of varying the weights of the KL divergence term. As a generative model, VAE aims to approximate a complex posterior distribution, the distribution of the latent variables z with a distribution of $q(z|v)$, typically as a multivariate Gaussian with parameters, i.e. mean $\mu(v)$ and variance $\sigma(v)$, which are modeled and predicted by the encoder Φ. In addition to the reconstruction loss, KL (Kullback-Leibler) divergence is included in the loss function to approximate the learned latent distribution $q(z|v)$ to be close to the prior distribution $p(z)$ of a standard

normal distribution $\mathcal{N}(0, I)$,

$$D_{\mathrm{KL}}\left(q_{\Phi}\left(z|v\right)||p\left(z\right)\right) = \beta \cdot \frac{1}{2}\sum_{i=1}^{d}\left(\sigma_i^2 + \mu_i^2 - \log\sigma_i^2 - 1\right), \tag{A.1}$$

and

$$l_{\beta\text{-VAE}} = \mathbb{E}_{q_{\Phi}(z|v)}\left[-\log p_{\Psi}\left(v|z\right)\right] + D_{\mathrm{KL}}\left(q_{\Phi}\left(z|v\right)||p\left(z\right)\right). \tag{A.2}$$

Here, q_{Φ} denotes the approximate posterior distribution through the encoder, p_{Ψ} is the likelihood function of the decoder, and $q_{\Phi}(z|v)||p(z)$ represents the relative entropy between the posterior and prior distributions. This penalty imposes constraints between the learned and target distributions, and that the generated samples resemble those from the prior distribution. Due to the KL divergence term, there is a trade-off between minimizing the reconstruction loss and staying close to the prior. We investigate the effect of KL divergence β through annealing by gradually increasing its weight β. As shown in figure A3, although a larger $\beta = 10^{-3}$ has a potentially faster decay within the first 60 epochs, a smaller β achieves a lower MSE loss on both the training and test datasets. This implies relaxing KL regularization allows the encoder and decoder to focus more on minimizing the reconstruction error and ensuring $p(v|z)$ captures the data well rather than strictly regularizing the latent space distribution. It may be attributed to the sharp transitions and significant variability in the mode structures, which pose challenges for the VAE in generating new samples that align with the prior distribution. Consequently, incorporating regularization as in equation (11) with an early stopping criterion in the baseline model of our architecture shows improved training and generalization performance, as seen in figure A3.

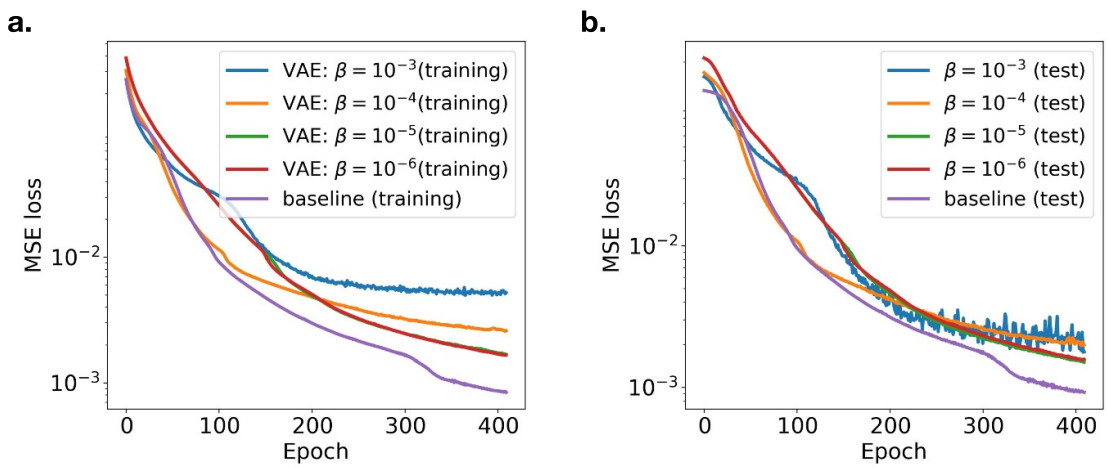

Figure A3. Mean squared error (MSE) vs. the number of epochs: (a) training set and (b) test set using VAEs with varied $\beta = 10^{-3}, 10^{-4}, 10^{-5}$, and 10^{-6} for regularization with KL divergence.

ORCID iDs

Zhe Bai https://orcid.org/0000-0002-3092-0903
Xishuo Wei https://orcid.org/0000-0001-7486-0407
William Tang https://orcid.org/0000-0002-8211-3636
Leonid Oliker https://orcid.org/0000-0002-7923-2896
Zhihong Lin https://orcid.org/0000-0003-2007-8983
Samuel Williams https://orcid.org/0000-0002-8327-5717

References

[1] McGuire K *et al* 1983 *Phys. Rev. Lett.* **50**
[2] Goeler S, Stodiek W and Sauthoff N 1974 *Phys. Rev. Lett.* **33** 1201–3
[3] De Blank H J 2008 *Fusion Sci. Technol.* **53** 122–34
[4] Carrera R, Hazeltine R D and Kotschenreuther M 1986 *Phys. Fluids* **29** 899–902

[5] ITER Physics Basis Editors, ITER Physics Expert Group Chairs and Co-Chairs, ITER Joint Central Team and Physics Integration Unit 1999 *Nucl. Fusion* **39** 2137
[6] ITER Physics Expert Group on Confinement and Transport and ITER Physics Expert Group on Confinement Modelling and Database and ITER Physics Basis Editors 1999 *Nucl. Fusion* **39** 2175
[7] Chen L, Lin Z and White R 2000 *Phys. Plasmas* **7** 3129–32
[8] Diamond P H, Itoh S, Itoh K and Hahm T 2005 *Plasma Phys. Control. Fusion* **47** R35
[9] Lucia D J, Beran P S and Silva W A 2004 *Prog. Aerosp. Sci.* **40** 51–117
[10] Quarteroni A *et al* 2014 *Reduced Order Methods for Modeling and Computational Reduction* vol 9 (Springer)
[11] Keiper W, Milde A and Volkwein S 2018 *Reduced-Order Modeling (ROM) for Simulation and Optimization: Powerful Algorithms as key Enablers for Scientific Computing* (Springer)
[12] Choi Y, Boncoraglio G, Anderson S, Amsallem D and Farhat C 2020 *J. Comput. Phys.* **423** 109787
[13] Noack B R, Morzynski M and Tadmor G 2011 *Reduced-Order Modelling for Flow Control* vol 528 (Springer)
[14] Duriez T, Brunton S L and Noack B R 2017 *Machine Learning Control-Taming Nonlinear Dynamics and Turbulence* vol 116 (Springer)
[15] Holmes P 2012 *Turbulence, Coherent Structures, Dynamical Systems and Symmetry* (Cambridge University Press)
[16] Benner P, Gugercin S and Willcox K 2015 *SIAM Rev.* **57** 483–531
[17] Taira K, Brunton S L, Dawson S T, Rowley C W, Colonius T, McKeon B J, Schmidt O T, Gordeyev S, Theofilis V and Ukeiley L S 2017 *AIAA J.* **55** 4013–41
[18] Schmid P J 2010 *J. Fluid Mech.* **656** 5–28
[19] Rowley C W, Mezić I, Bagheri S, Schlatter P and Henningson D S 2009 *J. Fluid Mech.* **641** 115–27
[20] Koopman B O 1931 *Proc. Natl Acad. Sci.* **17** 315–8
[21] Mezić I 2005 *Nonlinear Dyn.* **41** 309–25
[22] Bagheri S 2013 *J. Fluid Mech.* **726** 596–623
[23] Proctor J L, Brunton S L and Kutz J N 2016 *SIAM J. Appl. Dyn. Syst.* **15** 142–61
[24] Bai Z, Kaiser E, Proctor J L, Kutz J N and Brunton S L 2020 *AIAA J.* **58** 561–74
[25] Benner P, Goyal P, Kramer B, Peherstorfer B and Willcox K 2020 *Comput. Methods Appl. Mech. Eng.* **372** 113433
[26] Bai Z and Peng L 2021 *Adv. Model. Simul. Eng. Sci.* **8** 28
[27] Jardin S 2004 *J. Comput. Phys.* **200** 133–52
[28] Sovinec C, Glasser A, Gianakon T, Barnes D, Nebel R, Kruger S, Schnack D, Plimpton S, Tarditi A and Chu M 2004 *J. Comput. Phys.* **195** 355–86
[29] King J R, Burrell K H, Garofalo A M, Groebner R J, Kruger S, Pankin A Y and Snyder P B 2016 *Nucl. Fusion* **57** 022002
[30] Ferraro N and Jardin S C 2006 *Phys. Plasmas* **13** 092101
[31] Charlton L, Holmes J, Hicks H, Lynch V and Carreras B 1986 *J. Comput. Phys.* **63** 107–29
[32] Charlton L, Holmes J, Lynch V, Carreras B and Hender T 1990 *J. Comput. Phys.* **86** 270–93
[33] Lin Z, Hahm T S, Lee W, Tang W M and White R B 1998 *Science* **281** 1835–7
[34] Chang C S, Ku S and Weitzner H 2004 *Phys. Plasmas* **11** 2649–67
[35] Ku S, Chang C S and Diamond P H 2009 *Nucl. Fusion* **49** 115021
[36] Chen Y and Parker S E 2007 *J. Comput. Phys.* **220** 839–55
[37] Candy J and Waltz R 2003 *J. Comput. Phys.* **186** 545–81
[38] Jenko F, Dorland W, Kotschenreuther M and Rogers B 2000 *Phys. Plasmas* **7** 1904–10
[39] Görler T, Lapillonne X, Brunner S, Dannert T, Jenko F, Merz F and Told D 2011 *J. Comput. Phys.* **230** 7053–71
[40] Liu P, Wei X, Lin Z, Brochard G, Choi G J, Heidbrink W W, Nicolau J H and McKee G R 2022 *Phys. Rev. Lett.* **128** 185001
[41] Brochard G *et al* 2024 *Phys. Rev. Lett.* **132** 075101
[42] Kates-Harbeck J, Svyatkovskiy A and Tang W 2019 *Nature* **568** 526–31
[43] Degrave J *et al* 2022 *Nature* **602** 414–9
[44] Tang W *et al* 2023 *Contrib. Plasma Phys.* **63** e202200095
[45] Dong G, Wei X, Bao J, Brochard G, Lin Z and Tang W 2021 *Nucl. Fusion* **61** 126061
[46] Kadomtsev B B 1975 *Sov. Tech. Phys. Lett.* **1** 5
[47] Buttery R J *et al* 2000 *Plasma Phys. Control. Fusion* **42** B61
[48] Wei X, Choi G, Brochard G, Liu P, Lin Z and Bao J 2021 arXiv:2109.08873
[49] Brochard G *et al* 2022 *Nucl. Fusion* **62** 036021
[50] Raissi M, Perdikaris P and Karniadakis G E 2019 *J. Comput. Phys.* **378** 686–707
[51] Cai S, Mao Z, Wang Z, Yin M and Karniadakis G E 2021 *Acta Mech. Sin.* **37** 1727–38
[52] Maljaars B, Felici F, de Baar M and Steinbuch M 2015 *IFAC-PapersOnLine* **48** 314–21
[53] Tartaglione G, Ariola M, Bie W, Di Grazia L E, Mattei M and Mele A 2022 Plasma magnetic control for DEMO tokamak using MPC *2022 IEEE Conf. on Control Technology and Applications (CCTA)* (IEEE) pp 825–30
[54] Bai Z, Oliker L and Williams S 2024 FTL: transfer learning nonlinear plasma dynamic transitions in low dimensional embeddings (FTL) v1.0 (https://doi.org/10.11578/dc.20240621.3)
[55] Jolliffe I T and Cadima J 2016 *Phil. Trans. R. Soc.* A **374** 20150202
[56] Bai Z, Brunton S L, Brunton B W, Kutz J N, Kaiser E, Spohn A and Noack B R 2017 *Data-Driven Methods in Fluid Dynamics: Sparse Classification From Experimental Data* (Springer)
[57] Brunton S L and Kutz J N 2022 *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems and Control* (Cambridge University Press)
[58] Garcia A V, Jalalvand A, Steiner P, Rothstein A, Zeeland M V, Heidbrink W W and Kolemen E 2023 *Nucl. Fusion* **63** 126039
[59] Kingma D P 2013 arXiv:1312.6114
[60] Higgins I, Matthey L, Pal A, Burgess C P, Glorot X, Botvinick M M, Mohamed S and Lerchner A 2017 *Int. Conf. on Learning Representations*