Representation learning algorithms for inferring machine independent latent features in pedestals in JET and AUG

A.E. Järvinen,[1,2] A. Kit,[2] Y.R.J. Poels,[3,4] S. Wiesen,[5] V. Menkovski,[4] L. Frassinetti,[6] M. Dunne,[7] the ASDEX Upgrade team,[a)] and JET Contributors[b)]

[1)] *VTT Technical Research Centre of Finland, FI-02044 VTT, Finland*

[2)] *University of Helsinki, FI-00014 Helsinki, Finland*

[3)] *École Polytechnique Fédérale de Lausanne, Swiss Plasma Center, CH-1015, Lausanne, Switzerland*

[4)] *Eindhoven University of Technology, Mathematics and Computer Science, NL-5600MB Eindhoven, The Netherlands*

[5)] *Forschungszentrum Jülich GmbH, Institut für Energie- und Klimaforschung - Plasmaphysik, DE-52425 Jülich, Germany*

[6)] *Division of Fusion Plasma Physics, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden*

[7)] *Max-Planck-Institut für Plasmaphysik, D-85748 Garching, Germany*

(*Electronic mail: aaro.jarvinen@vtt.fi)

(Dated: 14 February 2024)

Variational autoencoder (VAE) based representation learning algorithms are explored for their capability to disentangle tokamak size dependence from other dependencies in a dataset of thousands of observed pedestal electron density and temperature profiles from JET and ASDEX-Upgrade tokamaks. Representation learning aims to establish a useful representation that characterizes the dataset. In the context of magnetic confinement fusion devices, a useful representation could be considered to map the high dimensional observations to a manifold that represents the actual degrees of freedom of the plasma scenario. A desired property for these representations is organization of the information into disentangled variables, enabling interpretation of the latent variables as representations of semantically meaningful characteristics of the data. The representation learning algorithms in this work are based on VAE that encodes the pedestal profile information into a reduced dimensionality latent space and learns to reconstruct the full profile information given the latent representation. Attaching an auxiliary regression objective for the machine control parameter configuration, broadly following the architecture of the domain invariant variational autoencoder (DIVA), the model learns to associate device control parameters with the latent representation. With this multimachine dataset, the representation does encode density scaling with device size that is qualitatively consistent with Greenwald density limit scaling. However, if the major radius of the device is given through a common regression objective with the other machine control parameters, the latent state of the representation struggles to clearly disentangle the device size from changes of the other machine control parameters. When separating the device size as an independent latent variable with dedicated regression objectives, similar to separation of domain and class labels in the original DIVA publication, the latent space becomes well organized as a function of the device size.

I. INTRODUCTION

The standard high performance tokamak scenarios are based on the high confinement mode (H-mode) operation with a self-organized transport barrier and plasma pedestal at the edge[1,2]. As the heat transport in tokamak core plasmas is generally observed to be stiff, showing strong increase with temperature gradients exceeding a critical value, the overall achievable plasma performance becomes directly dependent on the boundary condition provided by the pedestal, such as the pedestal top pressure, p_{ped}. For example, simulations for ITER plasmas indicate fusion power to scale as p_{ped}^2[3]. Furthermore, in future fusion reactors, these high performance pedestals must be compatible with the necessary power exhaust measures to guarantee sufficiently long component duty cycles[4]. However, due to the different scalings of the pedestal and power exhaust physics from present-day fusion devices to reactor-scale facilities, a present-day fusion device cannot conclusively demonstrate reactor-relevant integration of high performance pedestal with a power exhaust solution. Therefore, accurate and fast predictions for pedestal plasmas are needed to design and operate the future tokamak fusion reactors.

Due to the multiple relevant physical processes, and spatial and temporal scales, predicting pedestal plasma states is extremely challenging, and pedestal transport and stability are very active topics of research[5–9]. Hence, reduced models assuming MHD-constrained pedestal pressure profiles provide presently the standard approach for projecting pedestal performance between scenarios and devices[10–13]. Even though these models have been very successful in predicting pedestal pressures in many presently operating tokamaks, such a reduced model approach does increase the prediction risk when extrapolating to a new domain, such as was observed when the carbon wall at JET was changed to the ITER-like wall[2],

or when scaling the device size towards the reactor-scale devices and integrating radiative power exhaust to the pedestal solution. On the other hand, a more complete, high physics fidelity predictive model for pedestal and core-edge integration would be expected to be computationally prohibitively demanding for agile design optimization tasks or real-time and control applications. Advancement of deep learning algorithms for high-performance computers has opened a path to a data-driven approach to overcome these gaps in predictive capability and model throughput[14–16].

In this work, representation learning algorithms are explored for pedestal profile measurements from the JET and ASDEX-Upgrade (AUG) tokamaks[17]. Representation learning algorithms aim to learn a useful representation that characterizes the training data. In the context of magnetic confinement fusion devices, a useful representation learning algorithm could be considered to compress the high dimensional experimental observations to a lower dimensional space that represents the actual degrees of freedom of the plasma scenarios, such that the high dimensional observations can be reconstructed with little loss from the low dimensional representation. More generally, state representation learning (SRL) algorithms aim to learn representations that are in low dimension, evolve in time, and are influenced by actions of an agent[17]. There could be several attractive applications for such representations, including rector and scenario design by supplementing reduced model assumptions in systems codes with fast, high-fidelity integrated scenario predictions and allowing computationally efficient optimization of feed forward scenario trajectories. Furthermore, such approaches could flexibly leverage previous and current information from the system to essentially open a pathway towards plasma scenario digital twins with plasma state awareness and enable agility for timely and informed decision, similar to Kalman Filtering in control applications[18]. In this work, the representation learning algorithms are based on variational autoencoders (VAE)[19]. Similar approaches have been previously discussed, for example, in the context of divertor detachment models[20], pedestals[21], and disruption predictions[22].

Building on top of the previous study by Kit et al.[21], the capabilities of these models to infer machine size scaling in large databases of experimentally observed pedestal plasmas in JET and AUG are investigated. One of the primary challenges of data-driven algorithms is that generally these are not expected to extrapolate beyond the data that was used for training the model. This work aims to address this challenge by investigating methods to disentangle the learned representation of the latent variable models into machine independent and dependent features. Theoretically, the machine independent features can extract the information that extrapolates between fusion devices of various scales. A related study by Kit et al.[23] investigates state representation learning algorithms that aim to capture the dynamical evolution of the plasmas by training a forward model together with the representation.

II. DATASET

The data used to explore the deep learning models in this work originates from experimental measurements of outer mid-plane (OMP) electron density, n_e, and temperature, T_e, profiles. For JET these are obtained from the High Resolution Thomson Scattering (HRTS) diagnostic[24] and for AUG the profiles produced by the Integrated Data Analysis (IDA) were used[25]. The HRTS measurements are used for JET without including profile fits and dedicated instrument function deconvolution procedures, such that a certain level of profile smearing is present in the data[26]. Only ITER-like wall (ILW) JET Pulse Numbers (JPN) are included in this work, which are all obtained after the improvement of the optical design of the laser input system. Following the improvement of the optical design, the profile smearing caused by the instrument function was reduced to a level of a relatively small correction rather than substantially changing the profile gradient and width[26]. The individual profile measurements are assumed to have negligibly small measurement error relative to the plasma fluctuation between individual time slices. This is expected to be a well justified assumption for the plasmas in this study.

It is also quite usual in the HRTS measurements to see relatively high scatter in T_e values in the scrape-off layer (SOL) as measured n_e values are low. To avoid feeding these large scatter T_e values to the deep learning algorithms, a heuristic approach was taken such that for any HRTS measurements where n_e is lower than 5×10^{18} m^{-3} and T_e is higher than 50 eV, the T_e value is replaced by $T_{e,\,\mathrm{corr}} = (n_e \times 10^{-18}$ m$^{-3}) \times 10$ eV. Such a procedure biases the data at low T_e values in the SOL. However, for the time being this was considered a better approach than confusing the model with a large scatter ranging between a few eV and values beyond keV, when it is known from scrape-off layer (SOL) physics that the expected values are significantly below keV. In future, the plan is to improve these SOL corrections with information from SOL diagnostics and boundary models.

It is also well known that the magnetic equilibrium reconstruction is often not sufficiently accurate to locate the separatrix within the profile sufficiently well considering the profile gradient scale lengths. To obtain a common separatrix alignment procedure for the investigated plasmas, the separatrix is shifted to the location where T_e equals 100 eV for all plasmas for both JET and AUG, which is of the order of the expected values for conduction limited SOL transport in H-mode conditions. However, this is an oversimplification and will be re-evaluated in future studies, as all dependencies on power densities, SOL conditions, and device sizes are neglected by this approach. Nevertheless, such a procedure does align the separatrix approximately to the location where the radially steep gradients of the pedestal T_e begin and provides a pragmatic approach to align 10^4 profiles with a batch algorithm. Reaching such level of robustness with a more sophisticated separatrix alignment algorithms is expected to require substantially more work than was allocated here. Furthermore, varying $T_{e,\,\mathrm{SEP}}$ within $20 - 30$ %, as observed in SOL simulations of deuterium and nitrogen injection scans at JET[27], does not affect the determination of the key pedestal parameters

significantly[28]. It is acknowledged that by doing this, the original uncertainty in the equilibrium reconstruction is replaced with the physics uncertainty of the separatrix realignment procedure. However, the latter is considered more attractive here as it provides a controlled bias to the data that is consistent between the data points, whereas the equilibrium reconstruction uncertainty magnitude and direction is expected to be more uncertain between the data points when combining thousands of measured profiles from more than one fusion devices. All profiles were mapped to a common normalized poloidal flux, Ψ_N, grid of 30 points uniformly spaced between 0.85 and 1.05.

Only deuterium plasmas without significant impurity seeding, resonant magnetic perturbations, kicks or pellets were considered in this study. For JET, the plasma discharge numbers are selected from the JET pedestal database[2], whereas for AUG, the algorithm selects plasma discharges that have been flagged as useful in the operating journal, which in practice only filters out plasma discharges that failed. Furthermore, only plasma time slices with core T_e exceeding 2 keV are selected. The total resulting number of plasma discharges in the set of data used for training the deep learning algorithms is 1280 for JET and 3724 for AUG. From these plasmas, time periods where the total heating power exceeds 3 MW for JET and 2 MW for AUG were selected for analysis. To reduce the size of the dataset, the total database of measured profiles was downsampled by randomly selecting ten time slices from each investigated time window as well as choosing the three time slices with the highest pedestal pressure within each time window. The intention of the former approach is to approximately retain the statistical distribution of the full data while downsampling, whereas the intention of the latter is to make sure that the highest pressure profiles right before Edge Localized Modes (ELMs) are retained in the training database. The downsampling procedure also helps to balance the dataset as there are significantly more profiles in a given AUG time window relative to a given JET time window. Since there is no effort to filter ELMs or identify ELM cycles in this study, the resulting scatter in the profile data is relatively large and encoded as stochastic variation by the model in this study. In future studies, methods to identify fractions of ELM cycles automatically from the data will be investigated. The dataset reduction is expected to help with agile algorithm testing. The total resulting number of plasma time slices is 55087, as the average number of time slices retained for a plasma discharge is a bit more than 11.

The following input machine control parameters were selected for the model in this work (Table I): (1) toroidal magnetic field, B_T, (2) safety factor at 95 % flux surface, q_{95}, (3) total heating power normalized by the Martin LH-threshold scaling[29], P_{TOT}/P_{LH}, (4) total deuterium injection rate, D_{TOT}, (5) major radius, R, (6) aspect ratio, $A = R/a$, where a is the minor radius, (7) elongation, κ, (8) upper triangularity, δ^u, and (9) lower triangularity, δ^l, of the plasma. Even though the Martin LH-threshold scaling might not provide a fully accurate LH-threshold value for all plasmas considered in this study, the scaling provides a pragmatic approach to normalize the total heating power as two devices of different scales

TABLE I. Overall ranges, means, and standard deviations, σ, of the applied machine control inputs in the deep learning model training in this work. $|B_T|$ stands for the absolute magnitude of the toroidal magnetic field in units of Tesla, $|q_{95}|$ for the absolute value of the safety factor at 95 % flux surface, P_{TOT}/P_{LH} for the total heating power normalized with the Martin LH-threshold scaling[29], D_{TOT} for the total deuterium gas injection rate in units of 10^{22} e/s, R for the major radius of the center of the plasma in units of meters, $A = R/a$ for the aspect ratio, κ for the elongation, δ^u for the upper triangularity, and δ^l for the lower triangularity of the plasma.

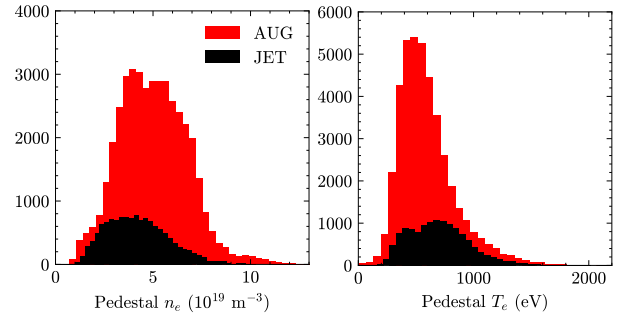|	JET				AUG			
Parameter	Min	Max	Mean	σ	Min	Max	Mean	σ
$|B_T|$ (T)	1.0	3.7	2.4	0.5	1.5	3.1	2.4	0.3
$|q_{95}|$	2.4	9.2	3.5	0.5	2.3	12.0	5.0	1.1
P_{TOT}/P_{LH}	0.2	7	1.9	0.8	0.5	12	2.6	1.3
D_{TOT} (10^{22} e/s)	0	19	2.3	2.2	0	10	0.8	1.0
R (m)	2.7	3.0	2.9	0.02	1.5	1.7	1.61	0.01
$A = R/a$	2.8	3.4	3.1	0.1	2.7	3.8	3.3	0.1
κ	1.27	1.83	1.67	0.04	1.09	2.0	1.67	0.06
δ^u	0.05	0.51	0.20	0.08	0.0	0.55	0.13	0.09
δ^l	0.04	0.50	0.31	0.04	0.0	0.59	0.42	0.07

FIG. 1. n_e and T_e distributions near the pedestal top in the dataset. The y-axis represents the number of samples in the bin.

are considered. As can be seen in Table I, the average plasma in the investigated database is an H-mode at about 2.4 T with a moderately low q_{95} of $3.5 - 5.0$ and moderately low upper triangularity of $0.1 - 0.2$. Figure 1 illustrates the distributions of the n_e and T_e values near the pedestal top in this dataset. The dataset used for training the model is split into training, validation, and test sets in proportions of 70%, 20%, and 10%. The data is further normalized by subtracting the mean and dividing by the standard deviation of the training set. In addition to this, two representative plasmas that have been documented in previous publications[13,30] are specifically selected to be retained outside the dataset used for training the model and applied for prediction testing. In the remainder of this manuscript, these will be called the holdout plasmas. For AUG, the plasma discharge number 33173 was selected, which is further discussed in[13]. For JET, the plasma discharge number 96202 was selected, which is further discussed in[30]. The representative input machine control parameters for the two holdout plasma discharges are shown in Table II. These plasmas represent relatively standard operational points and, therefore, test the models capability in interpolating between

previously observed points within the data distribution.

TABLE II. Representative machine control parameters for the two holdout plasma discharges.

	$\|B_\mathrm{T}\|$ (T)	q_{95}	$P_\mathrm{TOT}/P_\mathrm{LH}$	D_TOT (10^{22} e/s)	R (m)	A	κ	δ^u	δ^l
AUG	2.45	4.1	3.5	0.9	1.62	3.3	1.65	0.07	0.46
JET	2.3	3.5	1.7	0.8	2.91	3.1	1.63	0.13	0.33

III. MACHINE SIZE REGRESSION IN A REPRESENTATION LEARNING MODEL

Algorithms to learn a representation with tokamak size regression are explored. All approaches are based on the variational autoencoder (VAE)[19], which encodes the representation for the observed plasma profile information in this work (Fig. 2). In Section III A, a standard VAE is first trained with the dataset to demonstrate that the representation learning model has sufficient capacity to reconstruct the profiles well. Once this is demonstrated, Section III B explores an approach similar to domain invariant variational autoencoder (DIVA)[31] to connect the machine control configurations with the learned representation, building on top of the previous work by Kit et al.[21]. However, as will be discussed in Section III B, when connecting the device size together with the machine control parameters to the same auxiliary learning objective, the resulting device size regression is mixed with changes of the other parameters. This leads to large reconstruction error for the device size, indicating that the device size dependent features are not well disentangled from the size independent features. Therefore, in Section III C, a separate latent space is dedicated to the device size regression, which is observed to facilitate disentangled encoding of the size dependent features into this latent variable.

A. VAE

A VAE is a generative model, defining a joint distribution over the observed data, x, and latent variables, z, as $p_\theta(x, z) = p(z)p_\theta(x|z)$. Learning representation of the data means to learn the posterior distribution $p_{\theta 1}(z|x)$ that maps the observations to distributions of latent variables. The generative side of the model aims to model the dataset, x, as a conditional distribution given the latent representation, z, $p_{\theta 2}(x|z)$. In this work, this is implemented using encoder, $q_\Phi(z|x)$, and decoder, $p_{\theta 2}(x|z)$, distribution, parameterized by convolutional neural networks (Fig. 2). The $q_\Phi(z|x)$ represents an amortized variational approximation of the intractable true posterior, $p_{\theta 1}(z|x)$. The VAE model architecture in this work is shown in Table III. The encoder consists of convolutional layers that take the radial n_e and T_e profiles as input. The output from the second convolutional layer is transformed into mean and standard deviation of 12 latent variables through two dense layers. The decoder consists of transposed

convolutional layers. The last transposed convolution layer provides two output channels, which are transformed into two radial profiles of 30 points each through a dense layer.

TABLE III. The VAE model architecture. The inputs are experimentally measured 1D profiles of n_e and T_e mapped to a static Ψ_N grid of 30 points uniformly spaced between 0.85 and 1.05. These are the two channels with 30 points each that enter the convolutional layers at the first level of the encoder. Conv. and Transp. Conv. stand for 1D convolution and transposed convolution layers. The convolution layer parameters are (number of input channels, number of output channels, kernel width, stride). The encoder to state has two components as the VAE parameterize the mean and standard deviation of distribution. Since the last Trans.Conv. layer outputs two channels, the last dense layer transforms the output to two spatial profiles of 30 points each.

Model Component	Layers	Activation
Encoder	Conv. (2, 4, 4, 2)	ReLU
	Conv. (4, 8, 4, 2)	None
Encoder to latent state	Dense (48, 12), Dense (48, 12)	None
Decoder	Dense (12, 48)	None
	Transp. Conv. (48, 16, 5, 3)	ReLU
	Transp. Conv. (16, 8, 5, 3)	ReLU
	Transp. Conv. (8, 4, 6, 3)	ReLU
	Transp. Conv. (4, 2, 6, 3)	ReLU
Output	Dense (165, 30)	None

The learning objective is to minimize the reconstruction error as well as Kullback-Leibler (KL) divergence between the encoder distribution and a given model prior, chosen here as $\mathcal{N}(0, 1)$:

$$\begin{aligned}\mathscr{L}_\mathrm{VAE} = &-\alpha_\mathrm{obs.}\mathbb{E}_{q_\Phi(z|x)}\left[\log p_\theta(x|z)\right] \\ &+\beta_\mathrm{prior}\mathrm{KL}\left(q_\Phi(z|x)||\mathcal{N}(0, 1)\right).\end{aligned} \quad (1)$$

The second term constraints the model to learn a latent representations that resembles a Gaussian hypersphere. This constraint aims to regularize the learned representation towards continuity. The $\mathbb{E}_{q_\Phi(z|x)}$ stands for expection of the loss integrated over the amortized variational approximations of the posterior. In practice, the reconstruction error is obtained by sampling a reconstruction from the VAE latent posterior distribution and computing the mean squared error. β_prior multiplier is used to control the strength of the KL constraint, as motivated by the β-VAE by Higgins et al.[32]. Similarly, $\alpha_\mathrm{obs.}$ multiplier is added in front of the reconstruction error to also adjust the strength of the loss term. In addition to the standard objective in equation (1), a physics regularization term is added to the objective function $\propto \beta_\mathrm{press}||n_eT_e - \hat{n}_e\hat{T}_e||_2$ to add a penalty if the reconstructed static pressure, $\hat{n}_e\hat{T}_e$, deviates from the measured static pressure, n_eT_e. The intention is that this term encourages reconstruction of physically mutually consistent n_e and T_e profiles. Such a model was trained for 100 epochs with the adam optimizer[33] using a learning rate of 0.01 and batch size of 1024. The loss multipliers are shown in Table IV. These hyperparameters were manually selected based on manual testing of various hyperparameter values. In future, a more careful hyperparameter optimization studies will be performed. Figure 3 illustrates the reconstruction quality for the holdout plasmas and their representative
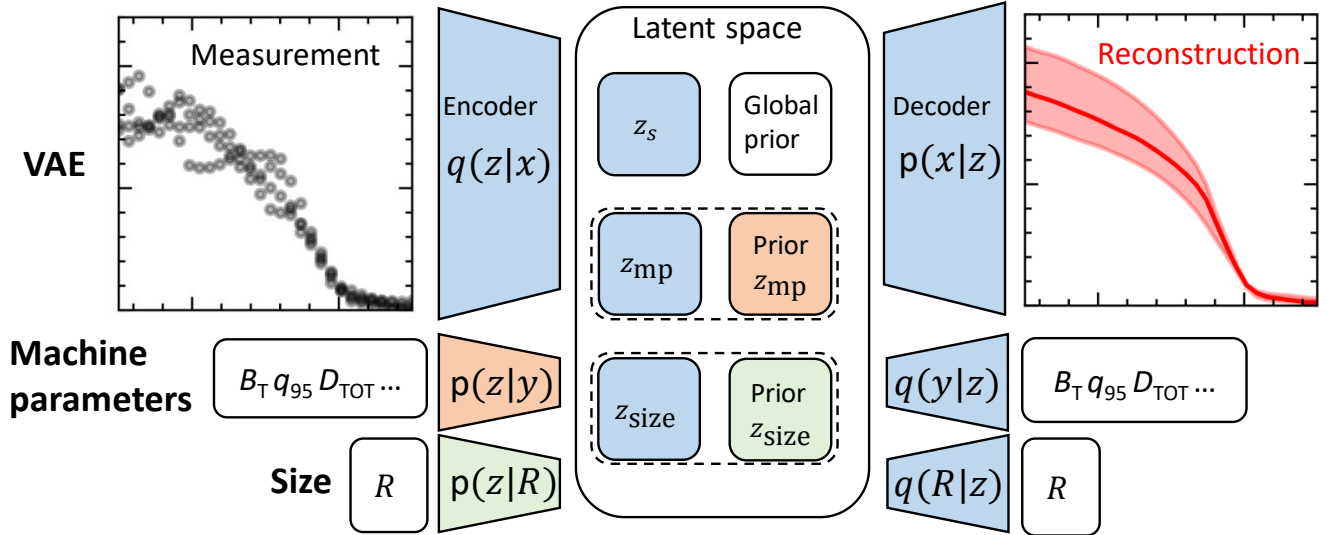
FIG. 2. Schematic illustration of the model components in this work. In Section III A, only the VAE component is active and only the global prior is applied to the latent space components z. The global prior stands here for the chosen VAE prior of $\mathcal{N}(0,1)$. There is no separation of the different components of z in the studies in Section III A. In Section III B, the machine parameter regression model is activated. The device size, R, is provided among the other machine parameters. The colors indicate that the prior regression module encodes a prior distribution for z_{mp}. It should be noted here that the auxiliary regression module, $q(y|z)$, decodes the machine parameter configurations based on the z_{mp}, not based on the prior. In Section III C, the device size dependence is given a dedicated regression and latent space. The detailed network architectures are documented in Tables III, VI. The dashed lines highlight that the regression model provided priors only apply to z_{mp} or z_{size}. The global prior always applies to all components of z, but the individual components have their own multipliers, β, in front of the term.

TABLE IV. Loss hyperparameters in the VAE model.

$\alpha_{obs.}$	β_{prior}	β_{press}
100	0.01	0.01

locations within a 2D cut of the learned 12 dimensional latent representation. The overall reconstruction errors are very small throughout the profiles, which indicates that the present model setup has the necessary capacity to represent the profiles in the investigated dataset (Table V).

TABLE V. The mean absolute errors of the reconstructions for the test dataset. n_e values are in units of 10^{19} m^{-3} and the T_e values are in units of eV. $<>$ represents profile integral, PED stands for pedestal top, and SEP for separatrix.

$<n_e>$	$<T_e>$	$n_{e, PED}$	$T_{e, PED}$	$n_{e, SEP}$	$T_{e, SEP}$
0.1	12	0.1	19	0.1	1

Since there are no other learning objectives than the reconstruction error and the overall shape of the latent representation through the KL constraint, the information may be entangled between the latent variables in this learned representation. To generate a useful representation, a desired property is organization of the information into disentangled latent variables, enabling interpretation of the latent variables as representations of certain, semantically-meaningful characteristics of the dataset[31,34–37]. To achieve disentangled latent vari-

ables, semi-supervision of the learning task is explored[31,35,37]. By introducing auxiliary learning objectives to a selection of the latent variables, the learning algorithm is given an incentive to encode the semantic information related to the learning objective to those latent variables. Attaching the auxiliary learning objective to the machine parameter configuration of the plasma discharge, a connection is also established between the machine configuration and the state of the plasma. The first approach, discussed in Section III B, follows broadly the structure of the domain invariant variational autoencoder (DIVA)[31]. In Section III C, this approach is extended by separating a dedicated latent space to represent the scaling with the machine size. The latter approach is actually closer to the original DIVA publication, where the machine size can be considered to represent a class label and the machine parameter configuration a domain label. However, the machine size in this work is treated as a continuous rather than discrete variable.

B. DIVA

The original DIVA publication by Ilse et al. discusses splitting the latent space of a VAE to three sub-spaces: (1) domain, (2) class, and (3) residual variations[31]. The motivation is to disentangle the latent representation to domain variant and invariant features, such that the domain invariant features could potentially be applicable to new, previously unseen domains.
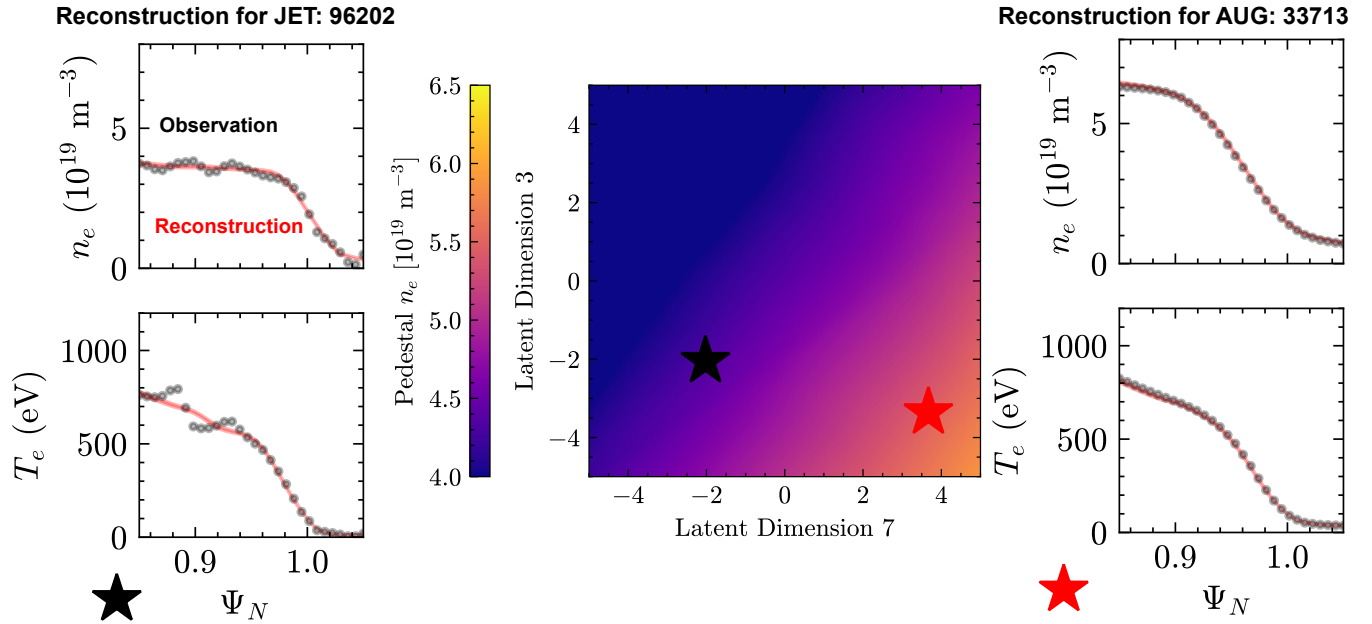
FIG. 3. Example of VAE reconstructed (red) n_e and T_e profiles for selected measured time slices (black dots) of the holdout plasma discharges and their representative locations for a 2D cut of the latent space. The selection of latent dimensions 3 and 7 is arbitrary here and any other 2D cut could have been selected equally well. The color coding of the latent space indicates the inferred pedestal top electron density value. Since the latent space color coding is established for an average latent space representation for the profiles within the training set, the exact pedestal density values are not same as those for the reconstructed holdout plasma discharges.

In this work, this approach is applied for JET and AUG experimental data with the aim to learn a representation for which certain latent features are independent of the device size. Fundamentally, the idea is that if it is possible to disentangle the features that are dependent and independent on device size, these tools could be used to infer fundamental device size scalings in the experimental databases and potentially also to build machine learning models that are transferable to new device sizes with relatively small amounts of training data at scale.

TABLE VI. The architecture of the machine parameter prior regression and auxiliary regression and their connection to the VAE latent space. The input dimensionality of the machine parameter vector is nine. The latent space dimensionality allocated for the prior regression is nine and the stochastic latent space is allocated three dimensions, such that the total dimensionality of $\boldsymbol{z}$ is twelve and the structure of the VAE (Table III) does not change other than the encoder to latent space layers being split to two different latent spaces.

Model Component	Layers	Activation
	Dense (9, 30)	ReLU
Prior regressor	Dense (30, 30)	ReLU
	Dense (30, 48)	None
Prior regressor to latent state	Dense (48, 9), Dense (48, 9)	None
Decoder	Dense (9, 30)	ReLU
Output	Dense (30, 9)	None

In this first test, all of the nine machine parameters, $\boldsymbol{y}_{\mathrm{mp}}$, (Table I) considered in this study are given as an input to a prior regression distribution, $p_{\mathrm{mp,\,prior}}(\boldsymbol{z}_{\mathrm{mp}}|\boldsymbol{y}_{\mathrm{mp}})$, parameterized via a fully connected neural network (Table VI). The

learning objective (eq. 1) is modified by a reconstruction loss for the machine parameter configuration and a KL constraint between the prior regressor distribution and the VAE posterior for the machine parameter latent space, $q_{\Phi}(\boldsymbol{z}_{\mathrm{mp}}|\boldsymbol{x})$:

$$\mathcal{L}_{\mathrm{DIVA\text{-}like}} = \mathcal{L}_{\mathrm{VAE}}$$
$$-\alpha_{\mathrm{mp}}\mathbb{E}_{q_{\Phi(\boldsymbol{z}_{\mathrm{mp}}|\boldsymbol{x})}}\left[\log q_{\theta,\,\mathrm{mp}}(\boldsymbol{y}_{\mathrm{mp}}|\boldsymbol{z}_{\mathrm{mp}})\right] \quad (2)$$
$$+\beta_{\mathrm{mp}}\mathrm{KL}(q_{\Phi}(\boldsymbol{z}_{\mathrm{mp}}|\boldsymbol{x})||p_{\mathrm{mp,\,prior}}(\boldsymbol{z}_{\mathrm{mp}}|\boldsymbol{y}_{\mathrm{mp}})).$$

This learning objective aims to encode information from the machine parameter configuration into the latent variable, $\boldsymbol{z}_y$, together with the observed n_e and T_e profile information. As the neural network generated prior distribution depends conditionally on the machine parameter configuration, the algorithm is given incentive to encode such a latent representation that also conditionally depend on the machine parameter configuration. Furthermore, the machine parameter reconstruction objective also incentives the algorithm to encode such a latent representation that the information content is sufficient to reconstruct the machine parameter configuration. Therefore, correlations with plasma state and machine parameters would be expected to be preferentially encoded to $\boldsymbol{z}_{\mathrm{mp}}$, while any residual variations would be expected to be encoded to the residual stochastic part of the latent space, $\boldsymbol{z}_{\mathrm{s}}$. To retain the same latent space dimensionality as in the initial VAE example, $\boldsymbol{z}_{\mathrm{mp}} = 9$ and $\boldsymbol{z}_{\mathrm{s}} = 3$ were selected in this study. For the original VAE model, the only change is that for those nine latent dimensions, the prior distribution is given primarily by the neural network model rather than the static $\mathcal{N}(0,1)$ prior.
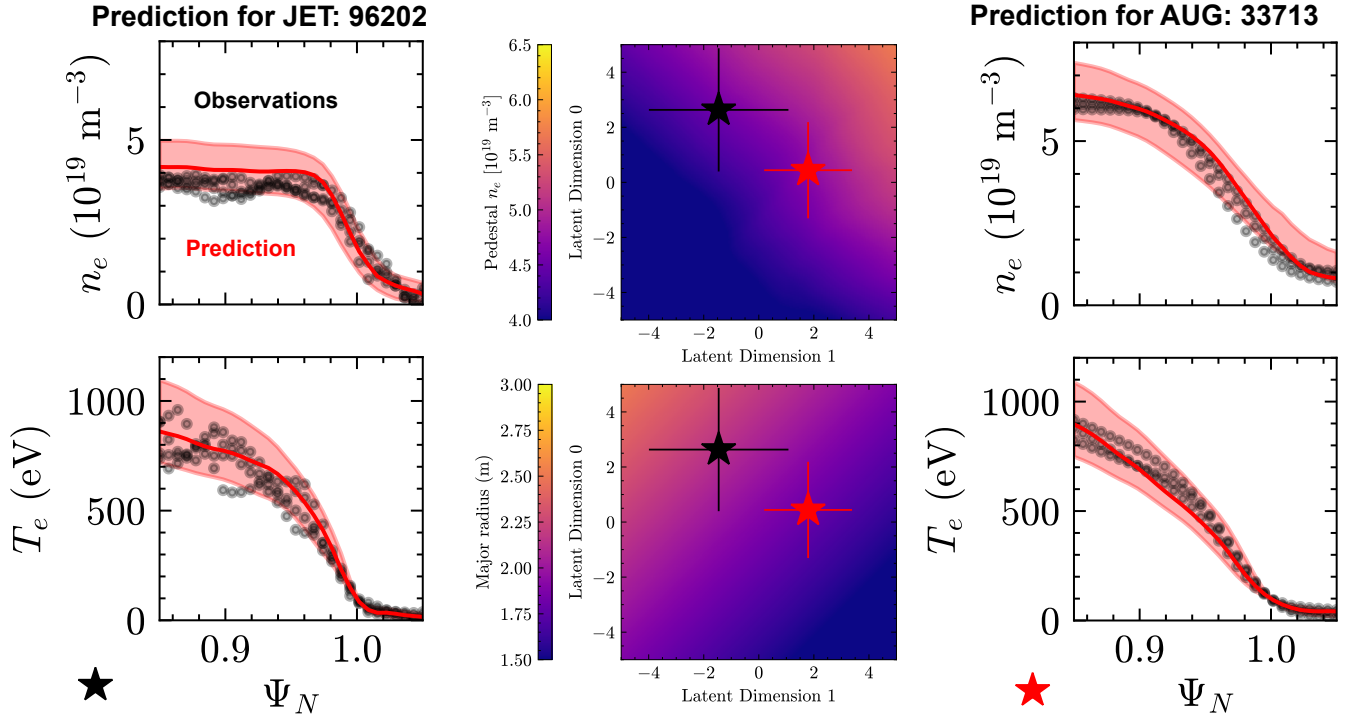
FIG. 4. Illustration of the performance of the DIVA-like model for the holdout plasma discharges and their respective locations in a 2D cut of the latent space, as predicted by the machine parameter prior regression. The contour colors show the inferred pedestal n_e (upper) and R (lower). The error bars illustrate one standard deviation of the distribution mapped by the prior regression module and the profile uncertainties are obtained through collecting conditionally generated samples. The solid red lines in the profile figures show reconstructions for the means of the latent distributions. A few observed time slices are shown for each of the holdout plasmas. The reason to show multiple time slices here is that those time slices represent the temporal fluctuation of the plasmas. When testing the prediction performance of the model, it is meaningful to compare the predicted distribution (red shaded area) to the scatter of the experimentally observed profiles. In Figure 3, the aim is to demonstrate the reconstruction capacity of the VAE for individual time slices.

For the remaining three latent dimension, the prior is simply the static $\mathcal{N}(0,1)$. The chosen latent space dimensionality is small relative to the dimensionality of the profiles, but not relative to the dimensionality of the machine parameters. This is intended, as the machine parameters are supposed to provide the independent control parameters of the system. If this would be exactly the case, then one would not be able to compres the information any more than is the dimensionality of the control parameter vector. In practice this is not the necessarily the case as the experimental databases have cross-correlations within them.

TABLE VII. Loss hyperparameters in the DIVA model. The $\beta_{\text{prior mp}}$ stands for the multiplier in front of the global prior term, while the $\beta_{\text{mp prior reg.}}$ stands for the multiplier in front of the prior regression term.

$\alpha_{\text{obs.}}$	α_{mp}	$\beta_{\text{prior stoch.}}$	$\beta_{\text{prior mp}}$	$\beta_{\text{mp prior reg.}}$	$\beta_{\text{press.}}$
1000	500	0.5	0.00001	1.0	0.01

The loss multipliers are shown in Table VII. The $\beta_{\text{prior mp}}$ stands for a small static prior loss on $\mathcal{N}(0,1)$ applied for the machine parameter latent space. This is a regularizing term

in this model setup. The multiplier value is set low as this loss should not dominate over the prior regression loss for the model to work as intended. Overall these hyperparameters resulted from simple manual scan and represent values that lead to reasonable model performance. A more systematic hyperparameter optimization would probably improve the model performance further but was not considered in this study.

Figure 4 illustrates prediction performance of the trained model for the holdout plasmas. By providing the machine control parameters of the holdout plasmas through the prior regression module, a conditional distribution is obtained in the latent representation. This conditional distribution is showed with stars and error bars in Figure 4. From this conditional distribution, full profile samples can be generated with the decoder of the VAE. As can be seen, the generated profiles are broadly in line with the observations, although the detailed profile shapes are not necessarily fully reproduced and the prediction uncertainties are quite large. One of the leading reasons for both of these shortcomings of the model is presumably the fact that ELM cycles are averaged by the model in this work. Therefore, stochastic variation below the ELM fluctuations would not be expected through the generative model

with this model setup. In future studies, focus will be given to more directly encode ELM information into the latent representation, which is expected to improve the prediction quality.

For a more quantitative performance test, mean absolute errors for reconstructions and conditional generation or prediction for the test set were calculated (Tables VIII, IX). There is no noticeable loss of reconstruction accuracy in the VAE model even though some of the VAE capacity is allocated for the additional training objectives (Tables V, VIII). The conditional generation or prediction error is about a factor of 5 – 10 larger than the reconstruction error (Table VIII), which is expected as solely based on the machine parameters there is no information about the fraction of the ELM cycle at which the target profile is measured. Considering this challenge, the prediction error is also relatively small. Actually, comparing to Figure 4, it can be seen that the prediction error is of the order of the standard deviation of the predicted distribution, which is expected. Therefore, when predicting an individual time slice with this model, one would not expect the model to be more accurate than this. Same arguments apply for the reconstructions of the machine parameters (Table IX). In future, further development of the model to encode information about the ELM cycle is expected to reduce both the prediction error and the machine parameter reconstruction error.

TABLE VIII. Mean absolute errors of the reconstructions and prediction for the test dataset. n_e values are in units of 10^{19} m^{-3} and the T_e values are in units of eV. $<>$ represents profile integral, PED stands for pedestal top, and SEP for separatrix.

	$<n_e>$	$<T_e>$	$n_{e,\,\text{PED}}$	$T_{e,\,\text{PED}}$	$n_{e,\,\text{SEP}}$	$T_{e,\,\text{SEP}}$
Reconstruction	0.1	13	0.1	21	0.1	1
Prediction	0.8	97	0.9	152	0.8	5

TABLE IX. Mean absolute errors of the reconstructions of the machine parameters for the test dataset. The names of the parameters and units are given the in caption of Table I.

$|B_T|$ (T)	$|q_{95}|$	$P_{\text{TOT}}/P_{\text{LH}}$	D_{TOT} $(10^{22}$ e/s)	R (m)	A	κ	δ^u	δ^l
0.4	0.7	0.7	0.8	0.24	0.07	0.04	0.07	0.04

As a test of the capability of the generative model to predict the impact of R on the pedestal n_e and T_e, the following experiment was conducted. The machine parameter inputs were set to representative values near the mean of the training set: $|B_T| = 2.25$ T, $q_{95} = 4.7$, $P_{\text{TOT}}/P_{\text{LH}} = 2.4$, $D_{\text{TOT}} = 10^{22}$ e/s, $A = 3.2$, $\kappa = 1.67$, $\delta^u = 0.14$, $\delta^l = 0.4$. Then R was scanned from 1.6 to 2.9 m (Fig. 5). It is observed that the model does encode a scaling of reduced pedestal density as R is increased, which is consistent with the scaling of the empirical Greenwald density limit in tokamaks[38], $n_{\text{GW}} = \frac{I_p}{\pi a^2} \propto \frac{B_T}{R}$, where the latter relation assumes constant plasma shape and q_{95} when scaling the device size. However, as can be observed from the reconstructed R, the internal representation of the model is not fully consistent with the scan of the major radius (Fig. 5). As there are more AUG time slices than JET time slices in the training database, it seems that the average represen-
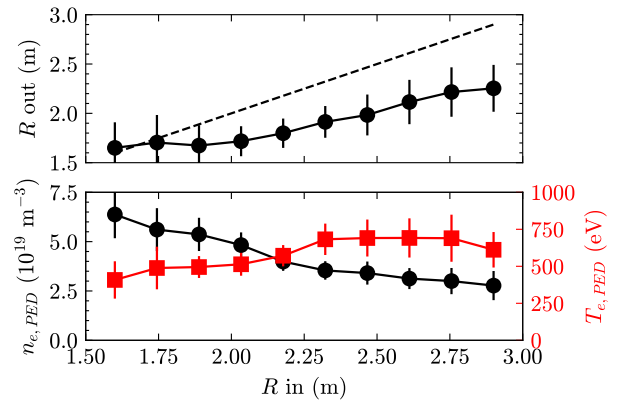

FIG. 5. Test of conditional generation for various device sizes while keeping other input parameters constant. Upper row shows the reconstructed R, the dashed line shows the line with no reconstruction error. The lower figure shows the predicted $n_{e,\,\text{PED}}$ (black) and $T_{e,\,\text{PED}}$ (red).

tation is weighted towards smaller device sizes. As in this model construction, all the machine parameters are encoded and reconstructed through a common fully connected neural network, the machine parameters end up being entangled with each other. Therefore, even though the model does learn the correlation of reduction of average pedestal top n_e with device size, this correlation is mixed with changes of other parameters, such that the internal state of the model is entangled. As a result, when the machine parameters are reconstructed from latent distribution, the machine parameter reconstruction model mixes the original drivers of the posterior latent distribution, leading to large reconstruction error on the device size. Therefore, although the results are already quite promising in terms of connecting machine parameters to the learned representation, the information organization is not yet quite as disentangled as would be needed to fully disentangle the machine size dependence from the other dependencies in the dataset. Therefore, in Section III C, a separate latent space is dedicated for the machine size.

C. DIVA with dedicated latent space for device size

A dedicated latent space of dimensionality of one was allocated for the machine size regression model. To retain the same latent space dimensionality as previously, the latent space dimensions were set as $z_{\text{mp}} = 8$, $z_{\text{size}} = 1$ and $z_s = 3$. The machine parameter and size regression architectures are similar to that presented in Table VI. The primary difference is that the input, output, and latent space dimensionality of the machine parameter regression is reduced to eight, and the representative dimensionality for the size regression is one. The loss is similar to eq. (2) with the machine size reconstruction loss and KL constraint separated from the other machine parameters. The loss multipliers are shown in Table X. As before, these hyperparameters are obtained through a manual search and a more systematic hyperparameter optimiza-
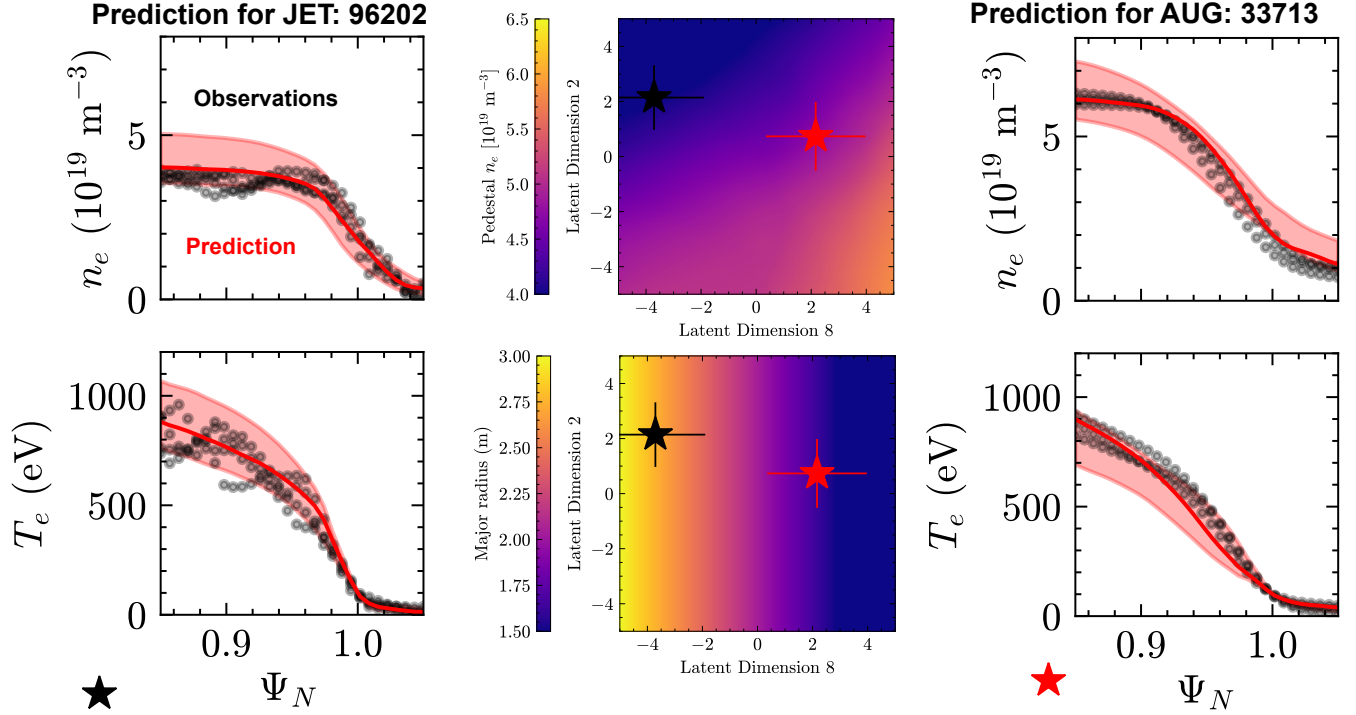
FIG. 6. Illustration of the performance of the DIVA-like model with the dedicated latent space for size dependence for the holdout plasma discharges and their respective locations in a 2D cut of the latent space, as predicted by the machine parameter prior regression. The contour colors show the inferred pedestal n_e (upper) and R (lower). The error bars illustrate one standard deviation of the distribution mapped by the prior regression module and the profile uncertainties are obtained through collecting conditionally generated samples. The solid red lines in the profile figures show reconstructions for the means of the latent distributions. A few observed time slices are shown for each of the holdout plasmas. Latent dimension eight is associated with device size, but the y-axis could have been plotted for any of the other machine parameter dependent latent dimensions, which do not have access to device size information directly.

tion could potentially improve the model performance further.

TABLE X. Loss hyperparameters in the DIVA model with a dedicated latent space for device size.

$\alpha_{obs.}$	α_{mp}	α_{size}	$\beta_{prior\ stoch.}$	$\beta_{prior\ mp/size}$	$\beta_{mp/size\ prior\ reg.}$	$\beta_{press.}$
1000	500	500	0.5	10^{-5}	1.0	0.01

A clear qualitative feature that can be seen with this model setup is that the R scaling is now well organized into latent dimensions eight (Fig. 6). As the latent dimension eight is scanned from 4 to -4, the inferred tokamak major radius increases from mid-sized 1.5 m to large 3.0 m. Also the Greenwald like density scaling can now clearly be seen when comparing the upper and lower latent space contour plots in Figure 6, where the smaller device size is associated with higher pedestal n_e values.

As the model becomes more regularized through splitting of the latent space, there is a small, less than factor of 2, increase of the VAE reconstruction error (Tables VIII, XI). The prediction errors remain almost the same between the models with and without the separate latent space for device size (Tables VIII, XI). For most of the machine parameters, recon-

struction errors remain the same with the exception that the reconstruction error on R is reduced by a factor of 2, as might be expected (Table IX, XII).

TABLE XI. Mean absolute errors of the reconstructions and prediction for the test dataset. n_e values are in units of 10^{19} m^{-3} and the T_e values are in units of eV. $<>$ represents profile integral, PED stands for pedestal top, and SEP for separatrix.

	$<n_e>$	$<T_e>$	$n_{e,\ PED}$	$T_{e,\ PED}$	$n_{e,\ SEP}$	$T_{e,\ SEP}$
Reconstruction	0.1	21	0.1	28	0.1	2
Prediction	0.8	98	0.9	151	0.7	5

TABLE XII. Mean absolute errors of the reconstructions of the machine parameters for the test dataset. The names of the parameters and units are given the in caption of Table I.

$|B_{TF}|$ (T)	$|q_{95}|$	P_{TOT}/P_{LH}	D_{TOT} (10^{22} e/s)	R (m)	A	κ	δ^u	δ^l
0.4	0.7	0.7	0.7	0.14	0.07	0.04	0.06	0.04

Finally, the test on the capability of the generative model to predict depence of pedestal n_e and T_e on machine size shows
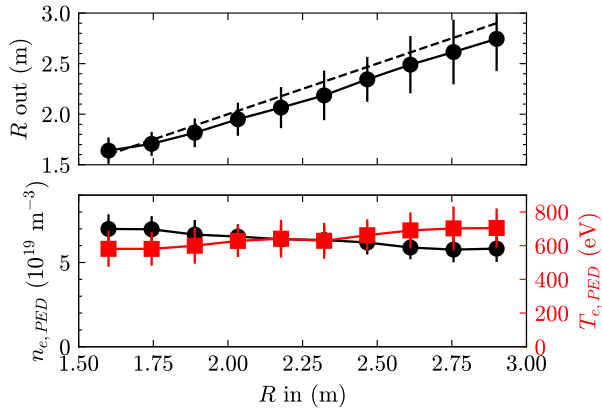
FIG. 7. Test of conditional generation for various device sizes while keeping other input parameters constant. Upper row shows the reconstructed R, the dashed line shows the line with no reconstruction error. The lower figure shows the predicted $n_{\mathrm{e,\,PED}}$ (black) and $T_{\mathrm{e,\,PED}}$ (red).

that the internal representation now captures the size dependence significantly better (Figs. 5, 6). As the input R approaches JET sized values, the internal state also approaches these numbers, which was not the case when R was mixed in the same latent space with the other machine parameters (Figs. 5, 6). Furthermore, similar test were done by scanning the size from the latent representation of each of the hold-out plasma discharges. When scanning from the AUG case, the dependence was very similar to that observed in Figure 7. From the JET case, the dependence was significantly weaker, which indicates that further research is needed to fully understand how the size dependence emerges globally in the latent space.

IV. DISCUSSION

Variational autoencoder based representation learning algorithms are explored for datasets of thousands of experimentally observed pedestal n_e and T_e profiles at JET and AUG tokamaks. Representation learning aims to establish a useful representation that characterizes the dataset. In the context of magnetic confinement fusion devices, a useful representation could be considered to map the high dimensional observations to a lower dimensional manifold that represents the actual degrees of freedom of the plasma scenarios. In this work, the capability of these algorithms to infer tokamak size dependence in multimachine databases is investigated.

Using VAE based representation learning models and following broadly the architecture of DIVA to connect the machine parameter configuration with the learned representation, the model learns joint probability distributions that associate a certain device size and machine control configuration with a certain pedestal n_e and T_e. Since ELM fluctuations are treated as stochastic variations of the profiles and not taken into account in the model design, the prediction variances for profiles and machine parameters remain quite large. Given this short-

coming, the performances for predicting pedestal n_e and T_e profiles based on the machine control configuration and device size within the test sets are quite good.

It was observed that when connecting the device size together with the machine control parameters to a common regression objective and latent space, the model ended up entangling the machine parameters within each other. As a result, conditional latent distributions with a certain device size were mixed with other device size and machine parameter configurations. By separating the device size to a dedicated regression objective and latent space, similar to separation of domain and class labels in the original DIVA publication[31], the device size information becomes well organized and disentangled from the other machine parameters. This highlights the importance of properly designing the learning algorithm structure appropriately to appropriately encode the information into semantically meaningful latent variables.

For these multimachine datasets, the representation does encode density scaling with a device size that is qualitatively consistent with Greenwald density limit scaling[38]. Since the training dataset consists of data from AUG and JET, it should be noted that there are no training observations at values in between the two extremes. Therefore, for a good fraction of the scan in Figure 7, the model is interpolating in a domain where there is no training data, and this interpolation seems quite stable, which is not necessarily given when operating with overparameterized models capable for overfitting. The features that the model learns to be independent on device size, could in principle extrapolate beyond the training domain. In future studies, such interpolation and extrapolation performance could be explored by training the model on JET and TCV, for example, and testing the interpolation performance for AUG, or extrapolation performance by training on TCV and AUG and extrapolating to JET. How much real data is needed at scale and whether simulations can be used to supplement the real data at scale are interesting research questions to be addressed in future.

DATA AVAILABILITY STATEMENT

The experimental data used for training the deep learning models in this work is stored at the data storage facilities of AUG and JET and the authors do not have permission to make this data publicly available. However, the training data preparation routines will be provided on request such that anyone with access to the data can regenerate the training dataset. The Python code encompassing the deep learning algorithms are available in GitHub at https://github.com/DIGIfusion/latent-state-modeling.

[1]H. Urano, "Pedestal structure in H-mode plasmas," Nuclear Fusion **54**, 116001 (2014).

[2]L. Frassinetti, S. Saarelma, G. Verdoolaege, M. Groth, J. Hillesheim, P. Bilkova, P. Bohm, M. Dunne, R. Fridström, E. Giovannozzi, F. Imbeaux, B. Labit, E. de la Luna, C. Maggi, M. Owsiak, R. Scannell, and JET contributors, "Pedestal structure, stability and scalings in JET-ILW: the EUROfusion JET-ILW pedestal database," Nuclear Fusion **61**, 016001 (2020).

[3]J. Kinsey, G. Staebler, J. Candy, R. Waltz, and R. Budny, "ITER predictions using the GYRO verified and experimentally validated trapped gyro-Landau fluid transport model," Nuclear Fusion **51**, 083001 (2011).

[4]A. Kallenbach, M. Bernert, R. Dux, L. Casali, T. Eich, L. Giannone, A. Herrmann, R. McDermott, A. Mlynek, H. W. Müller, F. Reimold, J. Schweinzer, M. Sertoli, G. Tardini, W. Treutterer, E. Viezzer, R. Wenninger, M. Wischmeier, and the ASDEX Upgrade Team, "Impurity seeding for tokamak power exhaust: from present devices via ITER to DEMO," Plasma Physics and Controlled Fusion **55**, 124041 (2013).

[5]C. Chang, S. Ku, A. Loarte, V. Parail, F. Köchl, M. Romanelli, R. Maingi, J.-W. Ahn, T. Gray, J. Hughes, B. LaBombard, T. Leonard, M. Makowski, and J. Terry, "Gyrokinetic projection of the divertor heat-flux width from present tokamaks to ITER," Nuclear Fusion **57**, 116023 (2017).

[6]M. Kotschenreuther, X. Liu, D. Hatch, S. Mahajan, L. Zheng, A. Diallo, R. Groebner, the DIII-D TEAM, J. Hillesheim, C. Maggi, C. Giroud, F. Koechl, V. Parail, S. Saarelma, E. Solano, A. Chankin, and JET Contributors, "Gyrokinetic analysis and simulation of pedestals to identify the culprits for energy losses using 'fingerprints'," Nuclear Fusion **59**, 096001 (2019).

[7]D. Hatch, M. Kotschenreuther, S. Mahajan, G. Merlo, A. Field, C. Giroud, J. Hillesheim, C. Maggi, C. P. von Thun, C. Roach, S. Saarelma, and JET Contributors, "Direct gyrokinetic comparison of pedestal transport in JET with carbon and ITER-like walls," Nuclear Fusion **59**, 086056 (2019).

[8]K. Barada, T. Rhodes, S. Haskey, R. Groebner, A. Diallo, S. Banerjee, L. Zeng, Z. Yan, J. Chen, F. Laggner, and G. Wang, "New understanding of inter-ELM pedestal turbulence, transport, and gradient behavior in the DIII-D tokamak," Nuclear Fusion **61**, 126037 (2021).

[9]H. Nyström, L. Frassinetti, S. Saarelma, G. Huijsmans, C. P. von Thun, C. Maggi, J. Hillesheim, and JET contributors, "Effect of resistivity on the pedestal MHD stability in JET," Nuclear Fusion **62**, 126045 (2022).

[10]P. Snyder, N. Aiba, M. Beurskens, R. Groebner, L. Horton, A. Hubbard, J. Hughes, G. Huysmans, Y. Kamada, A. Kirk, C. Konz, A. Leonard, J. Lönnroth, C. Maggi, R. Maingi, T. Osborne, N. Oyama, A. Pankin, S. Saarelma, G. Saibene, J. Terry, H. Urano, and H. Wilson, "Pedestal stability comparison and ITER pedestal prediction," Nuclear Fusion **49**, 085035 (2009).

[11]P. Snyder, R. Groebner, J. Hughes, T. Osborne, M. Beurskens, A. Leonard, H. Wilson, and X. Xu, "A first-principles predictive model of the pedestal height and width: development, testing and ITER optimization with the EPED model," Nuclear Fusion **51**, 103016 (2011).

[12]S. Saarelma, L. Frassinetti, P. Bilkova, C. D. Challis, A. Chankin, R. Fridström, L. Garzotti, L. Horvath, C. F. Maggi, and JET Contributors, "Self-consistent pedestal prediction for JET-ILW in preparation of the DT campaign," Physics of Plasmas **26**, 072501 (2019), https://pubs.aip.org/aip/pop/article-pdf/doi/10.1063/1.5096870/16687623/072501_1_online.pdf.

[13]M. G. Dunne, S. Potzel, F. Reimold, M. Wischmeier, E. Wolfrum, L. Frassinetti, M. Beurskens, P. Bilkova, M. Cavedon, R. Fischer, B. Kurzan, F. M. Laggner, R. M. McDermott, G. Tardini, E. Trier, E. Viezzer, M. Willensdorfer, The EUROfusion MST1 Team, and The

ASDEX-Upgrade Team, "The role of the density profile in the ASDEX-Upgrade pedestal structure," Plasma Physics and Controlled Fusion **59**, 014017 (2017).

[14]O. Meneghini, S. Smith, P. Snyder, G. Staebler, J. Candy, E. Belli, L. Lao, M. Kostuk, T. Luce, T. Luda, J. Park, and F. Poli, "Self-consistent core-pedestal transport simulations with neural network accelerated models," Nuclear Fusion **57**, 086034 (2017).

[15]J. Citrin, S. Breton, F. Felici, F. Imbeaux, T. Aniel, J. Artaud, B. Baiocchi, C. Bourdelle, Y. Camenen, and J. Garcia, "Real-time capable first principle based modelling of tokamak turbulent transport," Nuclear Fusion **55**, 092001 (2015).

[16]K. L. van de Plassche, J. Citrin, C. Bourdelle, Y. Camenen, F. J. Casson, V. I. Dagnelie, F. Felici, A. Ho, S. Van Mulders, and JET Contributors, "Fast modeling of turbulent transport in fusion plasmas using neural networks," Physics of Plasmas **27**, 022310 (2020), https://pubs.aip.org/aip/pop/article-pdf/doi/10.1063/1.5134126/15772521/022310_1_online.pdf.

[17]T. Lesort, N. Díaz-Rodríguez, J.-F. Goudou, and D. Filliat, "State representation learning for control: An overview," Neural Networks **108**, 379–392 (2018).

[18]F. Felici, M. de Baar, and M. Steinbuch, "A dynamic state observer for real-time reconstruction of the tokamak plasma profile state and disturbances," in *2014 American Control Conference* (2014) pp. 4816–4823.

[19]D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," ICLR (2014), arXiv:1312.6114 [stat.ML].

[20]B. Zhu, M. Zhao, H. Bhatia, X.-q. Xu, P.-T. Bremer, W. Meyer, N. Li, and T. Rognlien, "Data-driven model for divertor plasma detachment prediction," Journal of Plasma Physics **88**, 895880504 (2022).

[21]A. Kit, A. Järvinen, S. Wiesen, Y. Poels, and L. Frassinetti, "Developing deep learning algorithms for inferring upstream separatrix density at JET," Nuclear Materials and Energy **34**, 101347 (2023).

[22]Y. Wei, J. Levesque, C. Hansen, M. Mauel, and G. Navratil, "A dimensionality reduction algorithm for mapping tokamak operational regimes using a variational autoencoder (VAE) neural network," Nuclear Fusion **61**, 126063 (2021).

[23]A. Kit, A. E. Järvinen, Y. R.J. Poels, S. Wiesen, V. Menkovski, R. Fischer, M. Dunne, and ASDEX-Upgrade Team, "On learning latent dynamics of the aug plasma state," Submitted to Physics of Plasmas (2023).

[24]R. Pasqualotto, P. Nielsen, C. Gowers, M. Beurskens, M. Kempenaars, T. Carlstrom, D. Johnson, and JET-EFDA Contributors, "High resolution Thomson scattering for Joint European Torus (JET)," Review of Scientific Instruments **75**, 3891–3893 (2004), https://pubs.aip.org/aip/rsi/article-pdf/75/10/3891/10994132/3891_1_online.pdf.

[25]R. Fischer, C. J. Fuchs, B. Kurzan, W. Suttrop, E. Wolfrum, and ASDEX Upgrade Team, "Integrated Data Analysis of Profile Diagnostics at ASDEX Upgrade," Fusion Science and Technology **58**, 675–684 (2010), https://doi.org/10.13182/FST10-110.

[26]L. Frassinetti, M. N. A. Beurskens, R. Scannell, T. H. Osborne, J. Flanagan, M. Kempenaars, M. Maslov, R. Pasqualotto, M. Walsh, and JET-EFDA Contributors, "Spatial resolution of the JET Thomson scattering system," Review of Scientific Instruments **83**, 013506 (2012), https://pubs.aip.org/aip/rsi/article-pdf/doi/10.1063/1.3673467/16073704/013506_1_online.pdf.

[27]J. Simpson, D. Moulton, C. Giroud, M. Groth, and G. Corrigan, "Using EDGE2D-EIRENE to simulate the effect of impurity seeding and fueling on the upstream electron separatrix temperature," Nuclear Materials and Energy **20**, 100599 (2019).

[28]L. Frassinetti, M. Dunne, U. Sheikh, S. Saarelma, C. Roach, E. Stefanikova, C. Maggi, L. Horvath, S. Pamela, E. de la Luna, E. Wolfrum, M. Bernert, P. Blanchard, B. Labit, A. Merle, L. Guimarais, S. Coda, H. Meyer, J. Hillesheim, the ASDEX Upgrade Team, JET Contributors, the TCV Team, and the EUROfusion MST1 Team, "Role of the pedestal position on the pedestal performance in AUG, JET-ILW and TCV and implications for ITER," Nuclear Fusion **59**, 076038 (2019).

[29]Y. R. Martin, T. Takizuka, and (and the ITPA CDBM H-mode Threshold Database Working Group), "Power requirement for accessing the H-mode in ITER," Journal of Physics: Conference Series **123**, 012033 (2008).

[30]L. Frassinetti, C. P. von Thun, B. Chapman, A. Fil, J. Hillesheim, L. Horvath, G. Huijsmans, H. Nyström, V. Parail, S. Saarelma, G. Szepesi, B. Viola, R. B. Morales, M. Dunne, A. Field, J. Flanagan, J. Fontdecaba, D. Hatch, B. Lomanowski, C. Maggi, S. Menmuir, S. Pamela, C. Roach,

E. Rachlew, E. Solano, and JET Contributors, "Role of the separatrix density in the pedestal performance in deuterium low triangularity JET-ILW plasmas and comparison with JET-C," Nuclear Fusion **61**, 126054 (2021).

[31] M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling, "DIVA: Domain Invariant Variational Autoencoders," in *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, Proceedings of Machine Learning Research, Vol. 121, edited by T. Arbel, I. Ben Ayed, M. de Bruijne, M. Descoteaux, H. Lombaert, and C. Pal (PMLR, 2020) pp. 322–348.

[32] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework," in *International Conference on Learning Representations* (2017).

[33] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," CoRR **abs/1412.6980** (2014).

[34] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: a review and new perspectives," IEEE transactions on pattern analysis and machine intelligence **35**, 1798–828 (2013).

[35] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations," in *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 97, edited by K. Chaudhuri and R. Salakhutdinov (PMLR, 2019) pp. 4114–4124.

[36] B. Dai and D. Wipf, "Diagnosing and Enhancing VAE Models," in *International Conference on Learning Representations* (2019).

[37] T. Joy, S. Schmon, P. Torr, S. N, and T. Rainforth, "Capturing Label Characteristics in VAEs," in *International Conference on Learning Representations* (2021).

[38] M. Greenwald, J. Terry, S. Wolfe, S. Ejima, M. Bell, S. Kaye, and G. Neilson, "A new look at density limits in tokamaks," Nuclear Fusion **28**, 2199 (1988).