Deep learning for Gaussian process soft X-ray tomography model selection in the ASDEX Upgrade tokamak

F. Matos,[1, a] J. Svensson,[2] A. Pavone,[2] T Odstrčil,[3] and F. Jenko[1]

[1]*Max Planck Institute for Plasma Physics, Boltzmannstr. 2, 85748 Garching, Germany*

[2]*Max Planck Institute for Plasma Physics, Wendelsteinstr. 1, 17491 Greifswald, Germany*

[3]*Plasma Science and Fusion Center, Massachusetts Institute of Technology, Cambridge, MA, USA*

(Dated: 19 October 2020)

Gaussian process tomography (GPT) is a method used for obtaining real-time tomographic reconstructions of the plasma emissivity profile in tokamaks, given some model for the underlying physical processes involved. GPT can also be used, thanks to Bayesian formalism, to perform model selection – i.e., comparing different models and choosing the one with maximum evidence. However, the computations involved in this particular step may become slow for data with high dimensionality, especially when comparing the evidence for many different models. Using measurements collected by the Soft X-ray (SXR) diagnostic in the ASDEX Upgrade tokamak, we train a convolutional neural network (CNN) to map SXR tomographic projections to the corresponding GPT model whose evidence is highest. We then compare the network's results, and the time required to calculate them, with those obtained through analytical Bayesian formalism. In addition, we use the network's classifications to produce tomographic reconstructions of the plasma emissivity profile.

[a] Electronic mail: francisco.matos@ipp.mpg.de

I. INTRODUCTION

Computed tomography generally refers to the process of imaging the interior of a body through indirect measurements. In many applications, this is achieved by focusing penetrating radiation on an object of interest from several directions and measuring the resulting decrease in radiation intensity on the opposite side (due to absorption by the body itself). Use of this information, the so-called *projection* of the object, allows one to reconstruct its internal properties[1].

In the case of radiative bodies, an alternative way to determine their properties is to perform cross-sectional imaging by treating the emitted radiation itself as a projection[2]. In the field of nuclear fusion, this procedure is employed in many tokamaks for the reconstruction of plasma emissivity profiles[3]. More specifically, in the ASDEX Upgrade tokamak, such imaging can be done with information from the soft X-Ray (SXR) diagnostic, which measures the line-integrated radiation emitted by the plasma along several lines of sight; these can be used to perform tomographic reconstruction (or inversion) of the plasma emissivity profile. Knowledge of this is useful for exploring magnetohydrodynamic phenomena, in addition to studying accumulation of impurities inside the plasma (particularly tungsten) due to their large contribution to the total amount of radiation[4].

Several techniques exist for solving the tomography problem[5]. One approach is to use regularization-based algorithms, namely Tikhonov[6,7]- and minimum Fisher-based techniques[8]. More recently, work has also been done using machine learning methods, namely deep neural networks[9–11], that are trained to create new reconstructions based on existing ones.

Yet another method is Gaussian process tomography (GPT)[12]. GPT is an established method for performing tomographic inversion on many different types of physical distributions, that are modeled as posterior Gaussian distributions in a Bayesian setting. Computing a posterior first requires specifying a prior distribution, which encodes one's assumptions about the underlying physical process before any measurements of it are taken. The posterior can then be computed based on that prior, and on an observation (measurement) of the data generated by the physical process. The prior itself can either be a fixed distribution, or be drawn from a family of different models.

Knowing the posterior, GPT guarantees that one can obtain the most likely (maximum *a posteriori*, MAP) estimate for the tomographic reconstruction as well as its associated error values. More interestingly, however, through Bayesian inference, GPT prescribes a way to estimate the evidence for different models, through a process known as Bayesian model selection. This procedure can be of particular importance in cases where the choice of prior might have a strong effect on the results of the tomographic inversion.

Unfortunately, in a neural network, there are no guarantees[13] about whether the reconstructions obtained correspond to the MAP estimate of the underlying distribution, and there is no direct way, in standard Deep Neural Networks, such as convolutional neural networks, to obtain uncertainty estimates on the outputs. Bayesian neural networks[14,15] and generative adversarial networks (GANs)[16] can generate probability distributions for their outputs; however, they can be computationally expensive and, in the case of GANs, difficult to train[17].

On the other hand, neural networks essentially store whatever function they have learned (through their training process) in their weights, making the inference process for new data very fast. With GPT, computing the MAP estimate based on a fixed model is also sufficiently fast for real-time purposes. This does not necessarily hold true, however, when performing bayesian model selection, since the process requires a series of additional computational steps, namely matrix inversions or using non-linear optimizers, which can be time-consuming, especially for data with a high dimensionality.

Thus, we propose an approach where we train a convolutional neural network (CNN) to learn the GPT model selection procedure. To do this, we take SXR measurement samples from several ASDEX Upgrade shots and, through Bayesian model selection formalism, compute for each data point the corresponding model (out of a set of possible, pre-defined ones) with the highest evidence. We then train the CNN to reproduce this step, i.e., to map measurements to their highest evidence model. Finally, through the GPT framework, we compute the tomographic reconstruction of the plasma emissivity profile for each measurement, given the most likely models predicted by the CNN.

This paper is organized as follows. Section II gives an overview of the problem of tomography, in particular soft x-ray tomography ASDEX Upgrade tokamak, and the existing techniques to solve it, including a review of GPT with bayesian model selection. Section III details the data we collected, the formulation of our problem, and the model proposed to solve it. Section IV details the direct results of the neural network classification, and the tomographic reconstructions obtained based on them; section V describes and discusses our conclusions.

II. BACKGROUND

A. Computed Tomography

The purpose of tomography is to reconstruct the internal (either two- or three-dimensional) properties of a given body from non-local measurements. Radon showed[18] that a 2D distribution can be retrieved from an infinite set of line-integrated measurements. In practical applications, the number of available measurements is always finite, but it is nevertheless possible to produce accurate reconstructions from a discrete set of measurements[19]. Tomographic algorithms can achieve this by taking many *projections* of the object of interest from different directions[1]. Mathematically, a projection is a function that computes the line-integrated absorbency (or, in the case of fusion plasmas, emissivity) of a body along several paths or lines of sight (LOSs) as

$$P_\theta(t) = \int\limits_{L(\theta,t)} G(x,y)\,dL \tag{1}$$

where t is a point in the projection domain, $L(\theta,t)$ is the LOS crossing the body mapping to t (along a direction given by an angle θ), and $G(x,y)$ is the two-dimensional physical distribution of interest (see Figure 1).

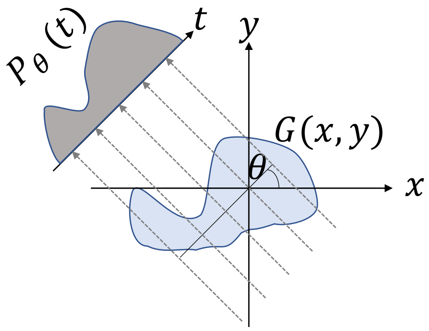


FIG. 1: An illustrated projection, P_θ, measured along an angle θ. The blue area $G(x,y)$ is the cross section of interest, and is being traversed by radiation. Because different rays traverse different areas of the object, the value at each point t in the projection space will be different. Figure reprinted with permission from F. Matos, "Deep Learning for Plasma Tomography" M.Sc. Thesis (University of Lisbon, 2016)[20].

By computing several tomographic projections with different directions (i.e. different values of θ), it is possible to reconstruct $G(x,y)$. For an exact reconstruction based only on the projections, an infinite number of them would need to be obtained. However, the problem is highly ill-posed[21], since small changes in projection space can translate into large changes in the tomographic reconstructions. Furthermore, in many settings such as nuclear fusion experiments, it is difficult, or impossible, to obtain more than a handful of such projections, making the problem under-determined – that is, the dimensionality of the reconstruction grid is much larger than that of the projection, ultimately resulting in an infinite number of solutions (reconstructions) that can fit the data. For these reasons, in most tomography applications in fusion plasmas, some additional information, in the form of assumptions about the function $G(x,y)$, must be introduced in order to obtain a tomographic reconstruction.

B. SXR tomography at ASDEX Upgrade

In the ASDEX Upgrade Tokamak, the Soft X-ray (SXR) diagnostic[22] consists of eight pinhole cameras that measure the total radiation emitted by the plasma along 208 different volumes of sight (VOSs)[4]. We considered the extent of the VOSs in the toroidal and poloidal directions of the tokamak to be minimal, and treated them instead as lines of sight (LOSs). In addition, we also ignored the fact that the LOSs in the same camera array partially overlap. Based on this, the measurements collected by the individual cameras correspond to a single projection of the underlying plasma emissivity distribution, which is computed at 208 discrete positions, in a poloidal plane. In terms of the poloidal coordinates (R,z) of the 2D tokamak cross-section, the total brightness, b_i, incident on a single detector, i, is given by[23]

$$b_i = \int_r G(R,z)\, dr \tag{2}$$

where $G(R,z)$ is the plasma emissivity distribution (in W/m^3) and r is the LOS corresponding to b_i (Fig. 2).

By discretizing equation 2, one obtains the plasma emissivity distribution at a finite number of positions (or pixels) along a tomographic reconstruction grid. In this case, the incident radiation on a single detector, assuming an associated noise, ξ, is[24]

$$b_i = \sum_{j=1}^{n} M_{i,j} g_j + \xi_i \qquad i \in 1,..,208. \tag{3}$$

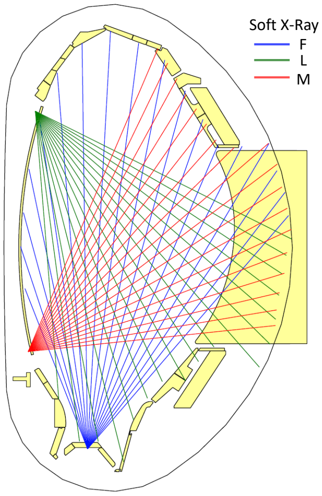


FIG. 2: ASDEX Upgrade cross-section, and partial schema of the SXR measurement system, with three cameras (F, L and M) shown (plot obtained with *diaggeom*)

From now on, we will denote the set of values b_i, that is, a set of 208 line-integrated SXR measurements of the plasma emissivity taken at a certain point in time, as the plasma's tomographic projection in that instant. We denote equation 3 as the *forward model* of the problem. Here, n corresponds to the total number of pixels on a tomographic reconstruction grid, whereas $M_{i,j}$ is the discretization of the function $M(R,z)$ in equation 2, mapping the relative contribution of pixel j of that grid to measurement i of the projection. The actual values of M were pre-defined and contingent on the geometry and configuration of the sensors inside the machine, which can vary between different shot campaigns. Consequently, we denote M as the *geometric matrix*. The goal of a tomographic reconstruction algorithm is then to solve the ill-posed problem by using the tomographic projection (i.e., the 208 measurements b_i) and some *a priori* knowledge about the plasma to find a suitable tomographic reconstruction g that satisfies equation 3.

C. Regularization-based Methods

To solve the ill-posed problem, traditional tomographic algorithms use regularization techniques, usually based on assumptions regarding the smoothness of the plasma emissivity profile, that constrain the space of possible solutions. Such algorithms, however, are often computationally expensive and typically can only be used for post-experimental tomographic reconstruction, due to computational time constraints. In addition, the quality of the reconstructions is highly dependent on assumptions made about the data[4]. Generally, those assumptions are encoded into the reconstructions through the use of Tikhonov regularization. In this case, computing the tomographic reconstruction of the plasma emissivity profile becomes a matter of finding a reconstruction $\hat{g}$ such that

$$\hat{g} = \arg\min_{g}(||Mg - b||^2 + \Lambda O(g)) \tag{4}$$

where $O(g)$ is a penalty term that encodes information about expected properties of the target plasma distribution, multiplied by a regularizing parameter Λ that controls the regularization strength[25]. There are several options for the choice of the regularization term O; typical choices are the Laplace operator, which favors smooth solutions, and minimum Fisher information[26], that favors solutions that are mostly flat in low-intensity regions, and peaked in high-intensity ones.

D. Deep Learning-based Methods

Recent work has applied deep learning algorithms to the tomographic problem, namely by using de-convolutional neural networks to produce tomographic reconstructions taking measurement data as input[10,11,27]. This is achieved by training the networks on reconstructions that have been previously computed using standard tomographic algorithms. Generically, in a deep learning setting, a Deep Neural Network is *trained* to learn a function mapping an input x into its target output y[28]; that is,

$$y = y(x, \theta) \tag{5}$$

where θ denotes the neural network's parameters, i.e. its weights and biases. The training process consists in finding an optimal value for θ that minimizes the mismatch between the network's outputs and their corresponding labels.

In our setup, training a deep neural network to produce tomographic reconstructions would have required training it with measurements from the SXR diagnostic and pre-computed reconstructions, produced by other algorithms (namely, regularization-based ones). The expectation would then have been that the parameters θ computed during training would have converged to values such that if new, unseen data were fed into the network, it would be capable of *generalizing* to outside of its training set. However, even assuming good generalization capacity of a neural network, it is at most as good as whatever data it has been trained on. In other words, should existing tomograms have had errors, a neural network would have learned to reproduce them.

E. Gaussian Process Tomography

Another alternative is to use bayesian probability theory to produce tomographic reconstructions, by treating the underlying unknown plasma emissivity distribution as a *Gaussian process*. Evaluating that process along a discrete set of points (the tomographic reconstruction grid) yields a multi-dimensional Gaussian distribution.

By definition, in the Gaussian process framework, one assumes that multiple solutions for the tomographic reconstruction exist, in a Gaussian distribution of possible solutions. Treating the tomography problem with this framework allows using Bayesian formalism, which guarantees that the most likely solution for the tomographic reconstruction (i.e. the maximum *a posteriori*, or MAP, estimate), subject to some assumptions about the underlying physical and data distributions, can be computed through Bayes's formula[29],

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{6}$$

In the Gaussian process tomography (GPT) setting, the terms in the formula are multivariate probability distributions, which are assumed to be Gaussian. Each of those distributions is specified by a vector of means (which we denote μ), whose entries are the individual means of each random variable in the multivariate distribution, and a covariance matrix, which we denote Σ, where each entry denotes the pair-wise covariance between those same variables.

In Bayes's theorem, the term $P(A)$ is called the *prior*. In GPT, by denoting the underlying plasma emissivity as e, the prior distribution $P(e) \sim \mathcal{N}(\mu_{pr}, \Sigma_{pr})$ encodes existing assumptions about the physical emission process, without observing any data (SXR measurements); this is equivalent to the assumptions one might encode in the regularization parameter in traditional tomography (Equation 4). Each random variable in the prior distribution is also Gaussian, and corresponds to the prior plasma emissivity e at each point x in the tomographic reconstruction grid.

Here, the prior mean has a size equal to that of the tomographic reconstruction grid, n. Intuitively, the prior covariance matrix Σ_{pr} encodes information about the expected smoothness of the plasma emissivity. The entries in the covariance matrix are computed for all pairs of points in the reconstruction grid through a prior covariance function. One covariance function generally used in Gaussian process regression is the squared exponential[30]; in using this function, the prior covariance between a pair of points x_1 and x_2 in the tomographic reconstruction grid becomes

$$\mathrm{cov}(x_1,x_2) = \theta_f^2 \exp\left(-\frac{d(x_1,x_2)}{2\theta_x^2}\right), \tag{7}$$

where $d(x_1,x_2)$ is a distance metric between points x_1 and x_2. The prior covariance is only dependent on that distance and on $\theta = \{\theta_f, \theta_x\}$, which are the model's *hyperparameters* and are common to all points in the reconstruction grid. The parameter θ_f controls the prior variance of the plasma emissivity at a given location in the reconstruction grid, whereas the parameter θ_x, usually referred to as the *length scale*, controls the extent to which points at a certain distance from each other in the reconstruction grid are correlated. Models where the length scale is large yield high correlations even between grid points which are far apart, while smaller length scales yield covariance matrices where only points which are closer to each other are significantly correlated.

With these definitions, the prior becomes a probability distribution for the plasma emissivity, e, subject to the model's hyperparameters, i.e., $P(e|\theta)$, before any data, that is, a tomographic projection, has been observed. The prior can then be updated by multiplying it with the likelihood of the data d (as per Bayes' theorem), yielding the posterior distribution, $P(e|d,\theta)$:

$$P(e|d,\theta) = \frac{P(d|e,\theta)P(e|\theta)}{P(d|\theta)} \tag{8}$$

The denominator in Bayes's theorem is known as the model *evidence* or *marginal likelihood*; if one is merely computing the posterior $P(e|d,\theta)$, it can be ignored, as it is just a normalizing constant. Interestingly, however, one can use this term to compare several different models (each with their own prior), and choose the one which best fits the data[12]. In this case, one assumes a *hyperprior*, from which different possible priors (individually specified by different hyperparameters) are sampled. The evidence can then be computed for different models – a process that is referred to as *marginalization* – and the model with the highest evidence can be selected[31]. Calculating this requires an evaluation of the integral[32] over e

$$P(d|\theta) = \int P(d|e,\theta)P(e|\theta)\,de \tag{9}$$

which in many cases is analytically intractable. However, in our case, the prior for an emissivity distribution is a multivariate Gaussian, defined as

$$P(e|\theta) = (2\pi)^{-\frac{k}{2}}|\Sigma_{pr}|^{-\frac{1}{2}}\exp(-\frac{1}{2}(\mu_{pr}-e)^T\Sigma_{pr}^{-1}(\mu_{pr}-e)), \tag{10}$$

where k denotes the number of variables in the prior distribution (that is, the number of pixels in a reconstruction grid). In addition, we assume a data distribution which is also Gaussian, $P(d|\theta) \sim \mathcal{N}(\mu_d, K+\Sigma_d)$[12]. The mean μ_d of the data distribution is merely the value of the measurements in a projection. The data co-variance has two components: matrix K denotes the (noise-free) co-variance values, and is a linear transformation of the prior covariance Σ_{pr} (imposed on the plasma emissivity) into measurement (data) space, given by $K = M\Sigma_{pr}M^T$, where M is the geometric matrix defined in equation 3. The other component, Σ_d, is a diagonal matrix whose non-zero entries are the absolute noise values, ξ, of each measurement in a projection; we assume the noise values are independent from each other. By assuming that this noise is also Gaussian, the logarithm of the integral in Equation 9 can be analytically calculated as[32]

$$\log(P(d|\theta)) = -\frac{1}{2}\left(m\log(2\pi) + \log|K+\Sigma_d| + (\mu_d - f_L)^T(K+\Sigma_d)^{-1}(\mu_d - f_L)\right) \tag{11}$$

where m is the number of SXR measurements in a tomographic projection and f_L is the mapping of the prior mean, μ_{pr}, (imposed in reconstruction space) onto measurement space, given by $f_L = M\cdot\mu_{pr}$.

The marginalization procedure is particularly useful because the trade-off between model complexity and data fit is automatic – the model for which the evidence score is highest is always the simplest model that can explain the data, an embodiment of the Occam's Razor principle[33]. In addition, the model evidence is also a function of the variance σ^2 of the data (through matrix Σ_d), which means that it is possible to treat the expected projection error as an additional hyperparameter of the model to be tuned; this can be done, for example, by treating the data variance as a fraction of the measured value of SXR radiation in the tomographic projection, with the value of the fraction constituting an additional model hyperparameter. This means that, through the

Gaussian process tomography framework, one can estimate not only the most likely model for the underlying plasma emissivity distribution, but also the most likely model for the error values of the data(though this is no replacement for a calibration of the detectors with a known source).

Once the most likely model is selected, and applying Bayes's formula, the posterior mean, μ_{post}, and posterior covariance, Σ_{post}, as a function of the prior mean μ_{pr} and prior covariance Σ_{pr} for that model, are respectively given by[34]

$$\mu_{post} = \mu_{pr} + \Sigma_{pr} M^T (K + \Sigma_d)^{-1} (d - f_L) \tag{12}$$

and

$$\Sigma_{post} = \Sigma_{pr} - \Sigma_{pr} M^T (K + \Sigma_d)^{-1} M \Sigma_{pr}. \tag{13}$$

By computing the posterior distribution $P(e|d,\theta) \sim \mathscr{N}(\mu_{post}, \Sigma_{post})$, one can then produce tomographic reconstructions either by sampling from $P(e|d,\theta)$, or simply by taking the mean of that distribution as the tomographic reconstruction (because the distribution is Gaussian, the mean corresponds to the maximum *a posteriori* estimate). In addition, one can directly obtain uncertainties for the tomographic reconstruction from the diagonal values of the posterior covariance matrix, which correspond to the individual posterior variances of each pixel in the reconstruction grid.

The drawback of the marginalization procedure, however, is its potential computational complexity. First, the calculation of the evidence term involves a series of matrix multiplications and an inversion, which can be cumbersome particularly in our setting because of the dimensionality of the data, which generates very large matrices. Matrix K in Equation 11 can be previously computed and kept in memory when performing model selection (which we do). However, in our setting, we treat the underlying error as a fraction of the data, and therefore, the values appearing in matrix Σ_d change with every new data point. As a result, the matrix determinant and inversion in Equation 11 must be re-computed for every new point, which is the main reason behind the high cost of the Bayesian optimization procedure. Furthermore, the evidence must be computed for all models that are taken into consideration. When each model has several hyperparameters, the number of possible models to evaluate can become very large, which means that finding the optimal one can be time-consuming. For practical purposes, this limits the number of models which can be evaluated and thus, potentially limits the quality of the results.

We therefore propose to bypass the need for analytical marginalization, by training a classifier (in this case, a convolutional neural network) to automatically choose the most likely model (out of several pre-defined ones) for the tomographic projection data collected by the ASDEX Upgrade SXR system.

This has potentially several advantages. On one hand, a Gaussian process model, while potentially having priors and posteriors with many dimensions, can be fully specified by its much smaller set of hyperparameters. In practice, this allows for parameterizing a distribution of high dimensionality with only a few variables. In the case of this work, this means that neural networks will learn to map tomographic projections to a lower-dimensional space (of dimensionality equal to the number of models under consideration). This should facilitate the network's learning process, allowing for easier generalization when compared with deep learning methods that attempt to map projections directly into a reconstruction space of larger dimensions. On the other hand, for potential real-time applications, this method potentially speeds up Gaussian process tomography, since it bypasses the marginalization procedure.

III. METHODS

A. Soft X-Ray Data

For this work, we had at our disposal a collection of 112 ASDEX Upgrade shots, totalling 127528 data points (208-dimensional tomographic projections), with each dimension corresponding to a specific detector in the Soft X-ray (SXR) system. The projections come from the downsampled signal of the SXR diagnostic, at a sampling rate of $250Hz$. The dataset also contains an error model, which assigns every measurement in every projection an estimated error value; we develop this topic in Section III B. In many cases the SXR detectors can be damaged and yield completely erroneous measurements, such as for example negative brightness; in these cases, the measurement is simply considered to be faulty. A sample projection can be seen in Figure 3.

We also possessed a geometric matrix M that maps the relative contribution of each pixel in a $60 \times 40 (2400)$-dimensional tomographic reconstruction grid to each of the 208 SXR measurements in a projection. Each pixel in the grid has a pair of poloidal coordinates (R, z), based on the poloidal dimensions of ASDEX Upgrade; the tomographic reconstruction is computed on this grid. The geometric matrix itself was computed based on the physical layout of the SXR sensors in the ASDEX Upgrade vessel, and holds for all shots in our dataset.

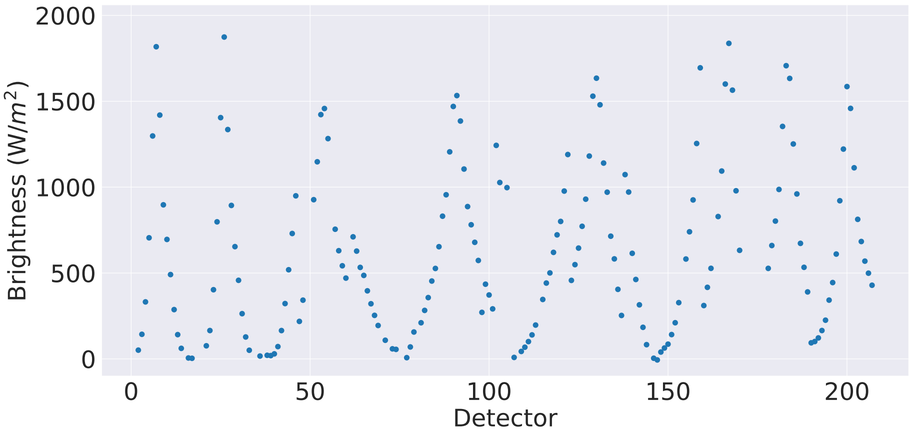


FIG. 3: Sample tomographic projection (SXR measurements) from the ASDEX Upgrade tokamak, taken from shot #30294 at $t = 5,8691s$. Faulty measurements have been removed.

B. Dataset Generation

Before training the neural network classifier, we generated its training and validation dataset by individually computing, for all the measured tomographic projections, the most likely model from which those projections were sampled. To that end, the first task was to define different models and their respective priors (individually specified by their specific set of hyperparameters) and then, through Equation 11, compare them based on their evidence.

As is typical in Gaussian process regression tasks[32], for all models, we defined the prior means, μ_{pr}, as vectors of zeros, of size 2400 (the size of the reconstruction grid). We computed the prior covariance matrices Σ_{pr} using a squared exponential function as defined in equation 7; in essence, this covariance function encodes our belief that correlations between pixels on the tomographic reconstruction grid will decay exponentially as the distance between those points increases. We computed the distance between pairs of points in the reconstruction grid (expressed in terms of their poloidal coordinates) using the Euclidean definition, i.e., $d(x_1,x_2) = \sqrt{(R_1 - R_2)^2 + (z_1 - z_2)^2}$. The covariance function (Equation 7) has only two parameters: θ_f, the individual variance of single pixels, and θ_x, the length scale which controls the extent of the correlation between pixels in the reconstruction grid. Different models have priors specified by different values of these hyperparemeters, but they all use the same definition of co-variance function and distance.

Finally, we defined, for each model, our assumptions regarding the data distribution associated with that model. For all models, we discarded measurements that had been previously labeled as faulty. In practice, this meant that, when evaluating the evidence for models, and when computing the maximum *a posteriori* (MAP) estimate for the plasma emissivity, some of the 208 measurements of each projection were not used. We treated the remaining (non-faulty) measurements in the projections as the mean values μ_d of the data distribution.

The individual variances, σ^2, of the variables in the data distribution correspond to the entries in the diagonal of matrix Σ_d of equation 11, and represent the uncertainties in the measurements. We computed the values σ^2 as fractions of the measurement values themselves; those fractions depend on a scaling factor θ_{err} which is multiplied by the measurements, and that constitutes an additional hyperparameter for the models under consideration. We assumed this value to be global, i.e., for any given model, we assume that the scaling factor is the same for all measurements in a projection.

Formalizing, we iteratively computed, for each individual data point (i.e. projection), and from a set of pre-defined models for the plasma emissivity and data distributions that might have generated that projection, the highest-evidence model – that is, through equation 11 we looked for $\hat{\theta} = (\hat{\theta}_f, \hat{\theta}_x, \hat{\theta}_{err})$ such that

$$\hat{\theta} = \arg\max_{\theta} \log P(d|\theta),$$

where $P(d|\theta)$ is the model evidence from Equation 8. We searched for the ideal hyperparameters (i.e. the hyperparemeters that specify the highest-evidence model) in a grid by assuming a uniform hyper-prior (all models were considered to be equally likely), and computed the model evidences at several discrete positions in the hyper-prior space. The question was then, what positions in the hyper-prior space should one evaluate the models' evidences on? This required taking several factors into account.

The first requirement was the expected nature of the plasma emission process itself. A previous analysis of the measurement data, and of existing tomographic reconstructions from ASDEX Upgrade[4], showed that the plasma emissivity has a wide dynamic range for different regions of the plasma, with emissivity in the plasma core being up to 3 orders of magnitude higher than in the pedestal. Likewise, in some periods of some shots, the maximum radiation value in the reconstruction grid was in the order of magnitude of $10^2 Wm^{-3}$, while in other phases, it could be as large as $10^5 W/m^{-3}$. Thus, we considered this range in emissivities a good region to explore possible values for the hyperparameter θ_f. In addition, ASDEX Upgrade has a minor radius $a = 0.5m$ (horizontally) and $b = 0.8m$ (vertically)[35]; given this and the size of our reconstruction grid, we assumed that a good region of the hyper-prior in which to evaluate the evidence for certain values of θ_x ranged, in the limit, from 0 (no correlation at all between pixels) to 1.6. For the hyparameter θ_{err} we assumed that, in the limit, it could range from 0 (no noise in the tomographic projections) to 1 (all of the measured brightness corresponded to noise).

The second requirement related to the training process for neural networks. For this work, we wanted to train a neural network to perform a classification task – to learn to map measurements to the the most likely model. Typically, in a machine learning classification setting, care should be taken such that training samples fed to a network are reasonably balanced with respect to their different classes; that is, a good training practice is that one class not be too over-represented in the data when compared to others. In our setting, achieving this balance required experimenting with different potential evaluation positions in the hyperparameter search grid. This comes at the cost of leaving out some grid positions for which some data points might have had higher evidence scores.

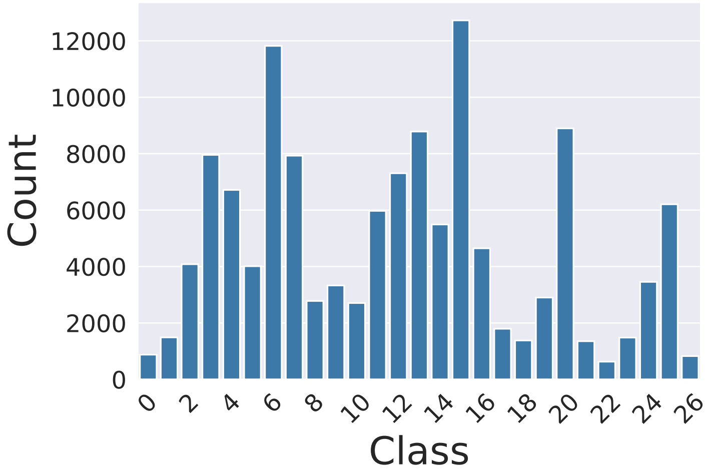


FIG. 4: Number of data samples (from all available shots) mapping to each class (set of hyperparameter values), after the marginalization procedure.

In practice, considering these requirements, we performed several evaluations through trial and error of the hyper-prior at different positions; we settled on a $3 \times 3 \times 3$ grid, where the points correspond to $\theta_f = \{500, 1000, 2500\}$, $\theta_x = \{.15, .175, .2\}$ and $\theta_{err} = \{.5, .75., 1\}$, which corresponds to 27 Gaussian process models. Performing the Bayesian model selection procedure on all projections in our dataset using the models parameterized by these values of $(\theta_f, \theta_x, \theta_{err})$ yielded a relative balance in terms of the amount of data samples mapping to each of the 27 possible classes (points on the hyperparameter grid); this can be seen in Figure 4, which shows the number of points mapping to each class. Computing the evidence for different models for all tomographic projections took a total of 48h.

This dataset – i.e., the mappings between tomographic projections and the class to which their highest-evidence model belongs (out of 27 possible ones) – was then used to train and test the neural network classifier. The choice of modelling the task with a classifier, instead of treating it as a regression problem, has one motivation: the loss function to use, and how to model the network outputs.

If performing regression, one option would have been to have a separate model output for each hyperparameter and combine them with a mean squared error loss. However, it is not obvious that this would work correctly, because the evidence term depends on all hyperparameters together. For example, a target output $y_t = (\theta_f = j, \theta_x = k, \theta_{err} = l)$ could be approximated by the network as $y_n = (\theta_f = j, \theta_x = 0.9k, \theta_{err} = l)$. Computing the mean squared error between y_t and y_n would yield a potentially good score because on average because the hyperparameter values are similar in both cases; however, there is no such guarantee for the evidence score, which could be very different from one case to the other. In fact, it is for this very reason that when performing classification we use a single output with 27 possible categories, instead of a separate output (and loss function) for each hyperparameter, each with 3 possible categories. The alternative would have been to use a loss function based on the evidence score itself; however that would have been computationally infeasible because it would require the evaluation of Equation 11 for each network gradient update.

C. Deep Learning Model

Several possibilities exist when it comes to modelling deep neural network architectures. For our purposes (the learning of the Bayesian model selection procedure) we opted to use a convolutional neural network (CNN). CNNs are widely used for signal processing tasks, due to their ability to efficiently detect spatial correlations in data, which is what we expected to find in our SXR measurements. The model we used is, with regards to its architecture, inspired by the network for classification of images described in[36], popularly known as the VGG network. We designed the model using the Keras framework for deep learning[37].

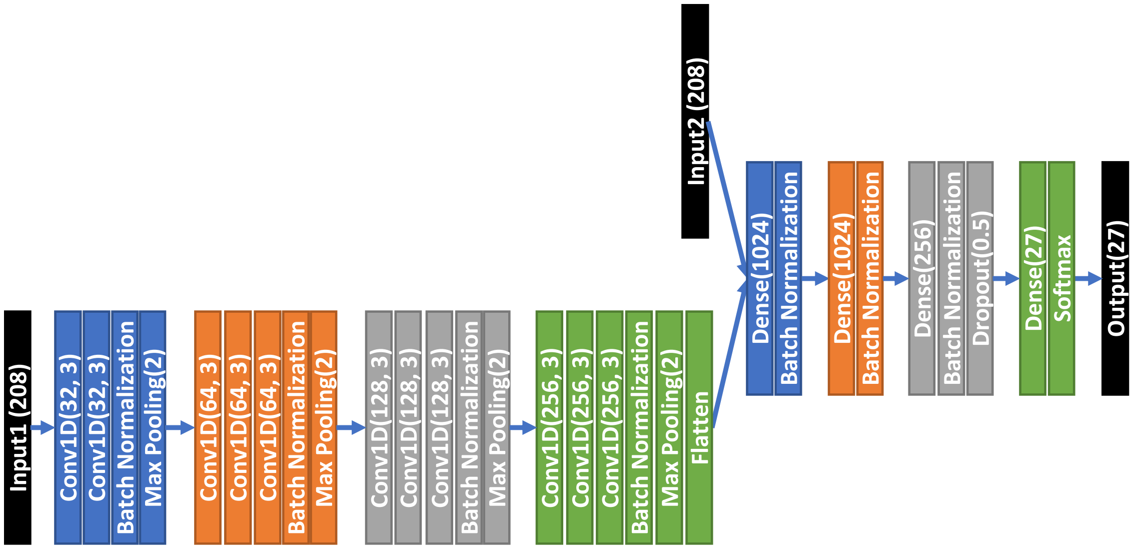


FIG. 5: Schematic of the deep learning model used for this work.

The network itself receives two inputs: a tomographic projection (208 SXR measurements, fed as input 1 in Figure 5), and a corresponding mask of ones and zeros (taken from the existing error model in our dataset) corresponding to input 2 in Figure 5, that gives information regarding which measurements in the projection are assumed to be faulty. The network uses a series of convolutional layers followed by max pooling layers to process high-level features in the measurement data. The output of those layers is then combined with the information in the error mask and processed in the last layers of the network, which are standard fully-connected layers. We also used batch normalization[38] layers to speed up training, and dropout in the final layer[39] to increase the network's capacity for generalization outside of its training set. We used the rectified linear unit (ReLU) activation function throughout the entire network apart from the last layer, which uses a softmax function, because we modelled the network output as probabilities over 27 possible classes, which must add up to 1. For the same reason, we used categorical cross-entropy as loss function. We used the Adam optimizer[40], and left all optimizer hyperparameters at their default values.

IV. RESULTS

We here performed two separate assessments. First, we evaluated the accuracy of the neural network's fit of the individual projections to their highest-evidence models. Then, based on the highest-probability class determined (by the network) for each data point, we computed the corresponding maximum *a posteriori* estimate of the tomographic profile, and measured the fit of those reconstructions to the data by projecting them back into measurement space (through the forward model in equation 3), obtaining their *back-projections*. We then measured the deviation between those back-projections and the original tomographic projections.

A. Neural Network

To increase the robustness of our methods, we opted to train an ensemble of neural networks (of equal architectures), using the k-fold cross-validation strategy[41]. K-fold cross validation is useful to determine whether the choice of the train/test split has biased whatever results have been obtained, or whether the results can be assumed to hold independently of the data split. We opted to divide our data into k= 10 folds – that is, we trained 10 networks with different overlapping splits of train data, and tested them on non-overlapping validation splits. We trained the networks for 50 epochs, and ran them on an NVIDIA Quadro RTX 5000 Graphics Processing Unit (GPU). The total training time for the whole ensemble was 1h, while total prediction time for the validation data was $41,62$s.

As the networks performed a 27-way classification, we used top-k categorical accuracy as a metric for network classification quality. We now follow with a brief explanation of this metric.

Each data point x (corresponding to a tomographic projection) in our dataset was assigned a label, y_{label}, denoting for which of the 27 model classes the evidence was highest. A classifier learns, through the training process, to compute the probability of that point belonging to a certain class $P(C(x) = c)$, where c can take one out of 27 possible values; we denote the vector containing the probabilities of belonging to each of those classes y_{pred}. We further define y_{pred_k} as the $k-$th most likely class given by a classifier for x; for example, for y_{pred_1}, one would get

$$y_{pred_1} = \arg\max_c P(C(x) = c) = \arg\max_c y_{pred}$$

whereas for c_{27} one would have

$$y_{pred_{27}} = \arg\min_{c} P(C(x) = c) = \arg\min_{c} y_{pred}.$$

Based on this, the top-k accuracy metric then calculates for each data point:

$$acc_k(x) = \begin{cases} 1 & \text{if } y_{label} \subset \{y_{pred_1}, ...y_{pred_k}\}. \\ 0 & \text{otherwise.} \end{cases}$$

We then computed the categorical accuracy metric on the validation data for the 27 different values of k. Because we opted for a cross-validation train and test strategy (with an ensemble of 10 classifiers), we show the mean value and standard deviation of the top-k accuracy across all members of the ensemble. The results of the metric can be seen in Figure 6 (up to k=27) and Table I (up to k=5). In Table I we show the results only up to k=5 for ease of comprehension.

	K				
	1	2	3	4	5
mean	0.509	0.783	0.903	0.955	0.977
st. dev.	0.041	0.038	0.033	0.016	0.01

TABLE I: Accuracy mean and standard deviation across ensemble of 10 neural networks, for validation data, up to top-5 accuracy.

An analysis of Table I and Figure 6 shows that the ensemble of 10 neural networks achieves very good results on the classification task, with a mean top-5 accuracy score of 0.976 (out of a maximum score of 1) for validation data. This means that for any data point, the correct prior is found within the top-5 most likely outputs predicted by the network in $97,7\%$ of cases. In practice, if one is exclusively interested in finding the single, most likely, prior, this result reduces the search space for the right hyperparameters from 27 classes to 5. Should one be interested only in comparing different models for certain physical distributions, this result also allows for quickly estimating which priors are more or less likely. Furthermore, the standard deviation of the accuracy score demonstrates consistently low values, indicating that the choice of train/test split for our data did not significantly bias the achieved results; all neural networks in the ensemble behave similarly, even if tested on different data.

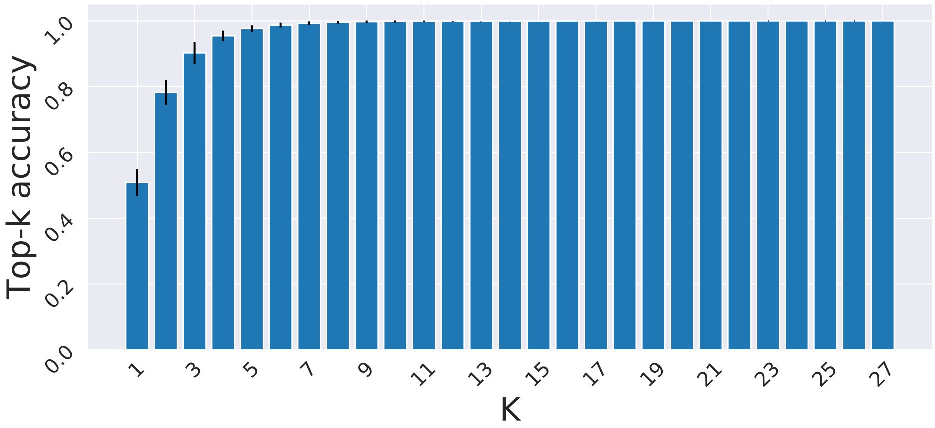


FIG. 6: Top-k accuracy (up to k=27) for validation data. The blue bars indicate the mean accuracy across the ensemble of 10 networks, while the smaller black bars indicate the accuracy's standard deviation across the ensemble (for each k)

B. Sample Reconstructions

In addition to evaluating the neural network's capacity for classification purposes, we also produced and evaluated tomographic reconstructions. To that end we took, for each data point in the validation dataset, the most likely class prediction given by the neural network; this class prediction maps to one of the models we have previously defined. We then computed, based on the class prediction and on equations 12 and 13, the posterior mean and covariance for each data point. We took the posterior means (i.e. the maximum *a posteriori* estimates) as the tomographic reconstructions of the plasma emissivity - each mean was a 2400-dimensional vector, where each entry μ_j denotes the most likely value for the plasma emissivity in a point j in the reconstruction grid. The posterior covariances allowed us to determine the error of the tomographic reconstruction, by taking the diagonal of the covariance matrix, which corresponds to the individual variance σ^2_{post} of each pixel in the reconstruction grid; we converted the value of that variance into a percentage error by once again taking advantage of the $3-\sigma$ rule, and computing said percentage, for pixel j, as

$$\%_{err_j} = 3\frac{\sqrt{\sigma^2_{post_j}}}{\mu_j} \times 100\%$$

Two sample results can be seen in figures 7 and 8. For the reconstruction error, we show only points where the percentage error was found to be below 100%. Notice how in figure 8, despite the value of σ_f being 2500, a reconstruction with a much larger maximum intensity can still be produced. Furthermore, in both cases, the reconstruction error values are noticeably lower in the center of the grid. We explain this with two factors: the larger number of LOS covering that region, which lower the uncertainty in the reconstructed values, and the higher intensity of the plasma emissivity, which lowers the relative error.

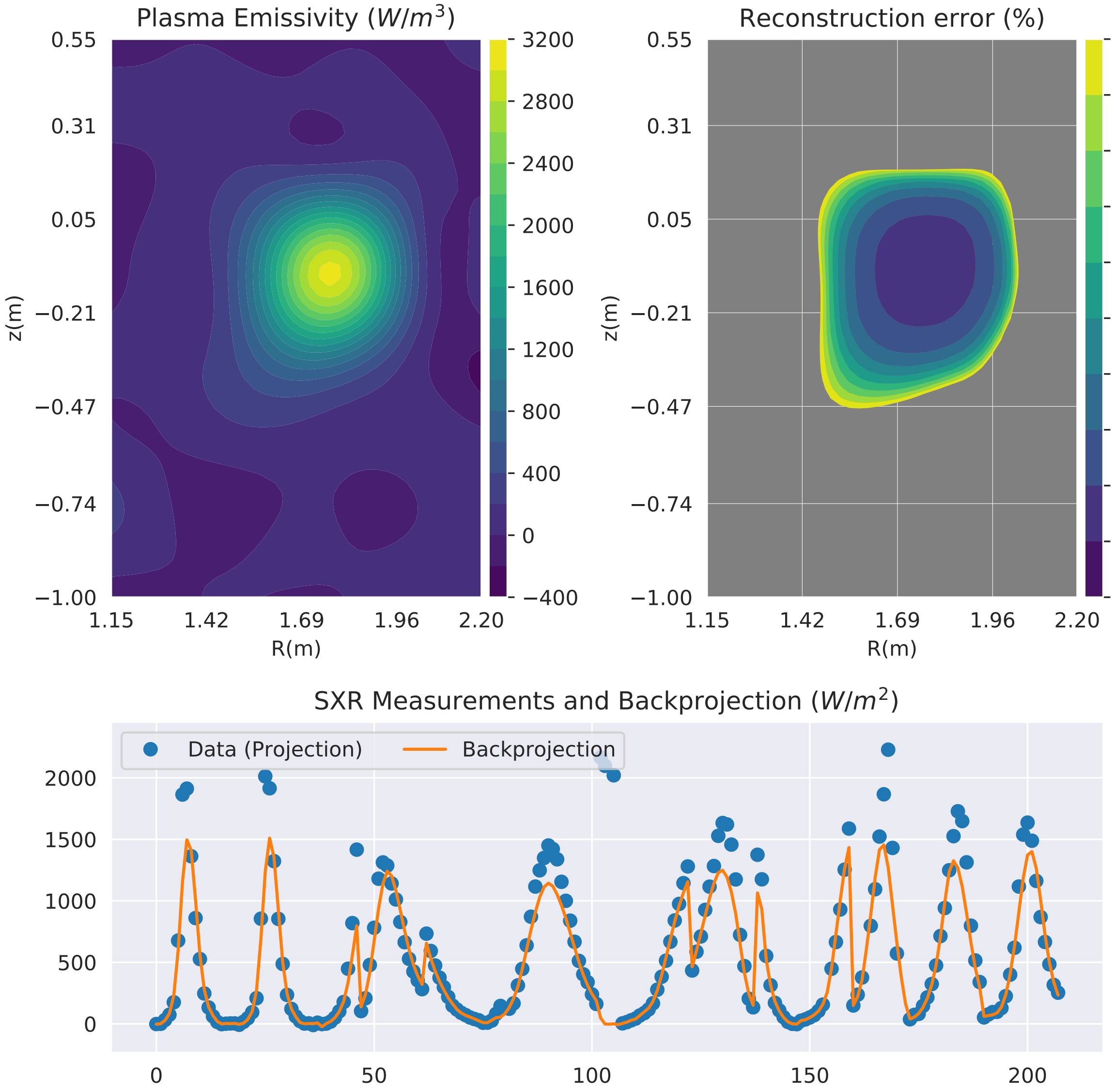


FIG. 7: Sample tomographic reconstruction and error, and comparison between the SXR measurement and the back-projected reconstruction, for ASDEX Upgrade shot #30857, t=$4,0441s$. The determined model hyperparameters by the classifier were $\theta_{err} = 0,75, \theta_f = 500, \theta_x = 0,175$; 200 measurements (out of 208) were used for this reconstruction.

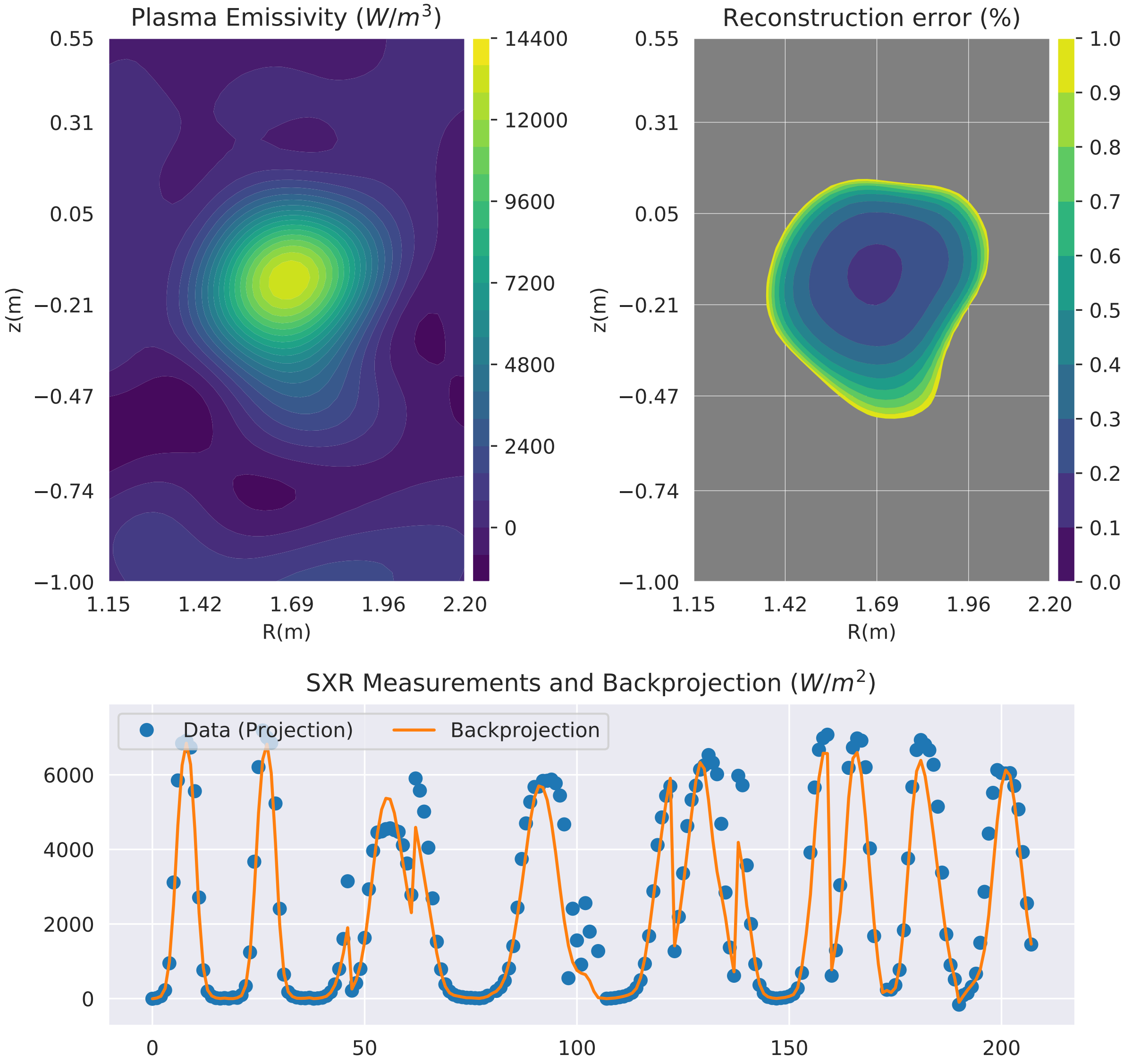


FIG. 8: Sample tomographic reconstruction and error, and comparison between the SXR measurement and the back-projected reconstruction, from ASDEX Upgrade shot #31238, t= $3,2641$. The determined model hyperparameters by the classifier were $\theta_{err} = 1,0, \theta_f = 2500, \theta_x = 0,175$. 199 measurements (out of 208) were used for this reconstruction.

C. Model complexity and data fit

To evaluate the quality of the models we performed, for each maximum *a posteriori* (MAP) tomographic reconstruction, a pass through the forward model defined in equation 3 to obtain the corresponding back-projection, i.e., the projection of the reconstruction back into measurement space. The marginalization procedure guarantees that, from the ensemble of models that is evaluated for a data point, the simplest model that can fit the data will be chosen, and we have shown that the proposed convolutional neural network can in most cases do this as well. However, a problem can arise if the ensemble of models from which we sample is itself mostly composed of overly simple or overly complex models.

Computing the backprojections allowed us to see how the obtained MAP estimates fit the original SXR data. Our expectation was that, if the models we defined were too complex, we would observe very tight fits of the data, with very low deviations between it and the backprojections. Conversely, if the models were too simple, we would tend to see large differences between the backprojections and the data.

To check this, we computed the percent deviation between the back-projections and the original tomographic projections, and did this for every measurement in every data point. We computed this value as

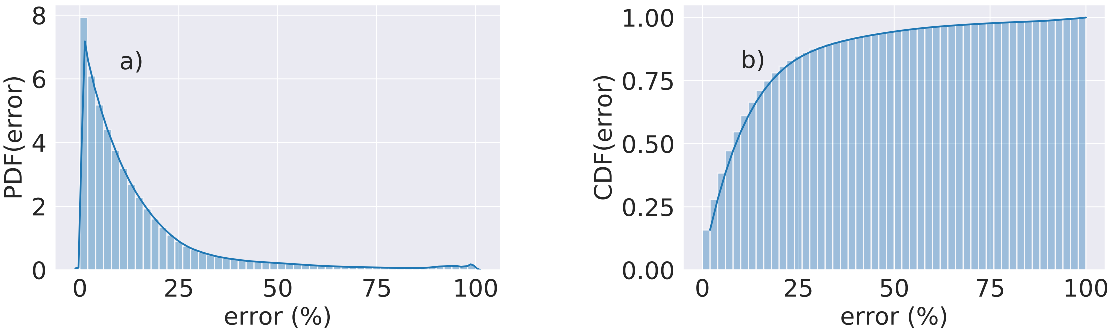
$$err = \frac{|backprojection - measurement|}{measurement} \times 100\%$$

.

The histograms in Figures 9a and 9b show the results of this evaluation, up to a deviation of 100%, a threshold which covers 99,38% of the validation data. Note that the deviation was computed by comparing the back-projection only with valid (non-faulty) measurements. On average, 91.25% of the 208 measurements in each data point were used to compute the tomographic reconstructions.

D. Discussion

Looking at the distribution of the deviations between backprojections and measurement data in Figure 9, one can see that most backprojected values have relatively low deviations from the data – 90% have a deviation lower than 50%. On the other hand, one can also see that some backprojected



Percent deviation values

Cumulative percent deviation values

FIG. 9: Cumulative distribution of the deviations between the tomographic reconstructions' back-projections into measurement space, and the data. $54,4\%$ of the individual back-projected measurements have a relative deviation lower than 10%; 93.83% have a deviation lower than 50%; and $99,38\%$ have a deviation lower than 100%.

values have larger deviations, and in fact, a few had an error greater than 100% (though they are not represented in the figure for ease of visualization; they represent only 0.6% of cases). We interpret this as a good result: on one hand, the low backprojection deviations indicate that the models chosen for the tomographic reconstructions have the necessary rigidity to constrain the solutions to mostly match the data, while at the same time being flexible enough to allow for large deviations, when not doing so would render overly complex models. We would like to point out the fact that many other types of models (other than the ones we used) can be chosen. For example, we used Euclidean distance and a single length scale for the covariance function. Nevertheless, it's possible to use more complex co-variance functions with different length scales in the R and z directions, or with a distance metric that uses the radial and angular coordinates of pixels in the reconstruction grid, to account for the fact that points in the same flux surface are considered to be highly correlated. In principle, it's always possible to define more complex models that fit the data better, and reduce the deviation between projection and backprojection; nevertheless, those models will not necessarily have the highest evidence when compared with simpler ones.

V. CONCLUSIONS

Gaussian process tomography makes it possible to obtain the most likely estimate for an unknown, potentially infinite-dimensional, quantity, given some assumptions about the underlying physical distribution and about the data generated by that distribution. The tomography problem, based on SXR measurement data from the ASDEX Upgrade tokamak, lends itself to investigation under this framework. If one assumes a fixed model for the behavior of the underlying physical distribution (i.e. the plasma emissivity) and for the data, for example by specifying the length-scales involved in the emission process and the expected fraction of noise in the measurements, Gaussian process tomography (GPT) inversion techniques readily yield the corresponding maximum *a posteriori* estimate of the plasma SXR emissivity in the two-dimensional tokamak cross-section.

Nevertheless, this raises the issue of what models one would like to assume in the first place. Through the Bayesian Occam's Razor principle, GPT answers this question by computing the evidence for different possible models, out of which the one with the highest score can then be selected. This can be useful if one wishes to test different assumptions regarding the data distribution: for example, what fraction of noise can be expected in the observations (measurements)? However, in a setting such as SXR tomography with ASDEX Upgrade data, this task can become cumbersome due to the dimensionality of the tomographic projections. This is further compounded when the number of models under evaluation is large.

For these reasons, we developed a novel method for automatic selection of the best model (out of 27 pre-defined ones) for the plasma SXR emissivity distribution and the corresponding data, for measurements from the ASDEX Upgrade tokamak. The individual models had different assumptions regarding the noise level in the collected data, the correlations between variables in the tomographic reconstruction grid, and the individual variances of those same variables. The method then consisted in training a convolutional neural network to perform the bayesian model selection (marginalization) procedure, and bypass the need to perform that task analytically. Our results show that the neural network achieved good classification results when compared to the analytical bayesian marginalization step, with top-5 accuracy (out of 27 possible classes) reaching a value of 0.976 (out of a maximum of 1). Furthermore, while the marginalization procedure across the entire dataset (of 127528 tomographic projections), through analytical methods, took approximately 48h, the same computation, performed by the neural network, took only 43s. Thus, the neural network approach can be particularly useful for high-dimensional data settings such as ours, as well as problems where the number of models under consideration is large, which would otherwise render the model comparison problem too slow through analytical methods. This can be particularly useful for settings where not only real-time inversion of tomographic profiles, but also real-time comparison of different models for certain physical distributions is a necessity.

ACKNOWLEDGEMENTS

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014-2018 and 2019-2020 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

DATA AVAILABILITY

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

REFERENCES

[1]A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging* (IEEE Press, 1988).

[2]L. Ingesson, B. Alper, B. Peterson, and J.-C. Vallet, "Chapter 7: Tomography diagnostics: bolometry and soft-x-ray detection," Fusion science and technology **53**, 528–576 (2008).

[3]J. Mlynar, V. Weinzettl, G. Bonheure, A. Murari, and JET-EFDA contributors, "Inversion techniques in the soft-x-ray tomography of fusion plasmas: toward real-time applications," Fusion Science and Technology **58**, 733–741 (2010).

[4]T. Odstrčil, T. Pütterich, M. Odstrčil, A. Gude, V. Igochine, U. Stroth, and the ASDEX Upgrade Team, "Optimized tomography methods for plasma emissivity reconstruction at the asdex upgrade tokamak," Review of Scientific Instruments **87**, 123505 (2016).

[5]J. Mlynar, T. Craciunescu, D. R. Ferreira, P. Carvalho, O. Ficker, O. Grover, M. Imrisek, J. Svoboda, and Jet Contributors, "Current research into applications of tomography for fusion diagnostics," Journal of Fusion Energy **38**, 458–466 (2019).

[6]A. N. Tikhonov, "Regularization of incorrectly posed problems," in *Dokl. Akad. Nauk. SSSR*, Vol. 153 (1963) p. 49.

[7]A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," in *Dokl. Akad. Nauk. SSSR*, Vol. 151 (1963) p. 501.

[8]V. Loffelmann, J. Mlynar, M. Imrisek, D. Mazon, A. Jardin, V. Weinzettl, and M. Hron, "Minimum fisher tikhonov regularization adapted to real-time tomography," Fusion Science and Technology **69**, 505–513 (2016).

[9]D. R. Ferreira, P. J. Carvalho, and H. Fernandes, "Deep learning for plasma tomography and disruption prediction from bolometer data," IEEE Transactions on Plasma Science **48**, 36–45 (2019).

[10]F. A. Matos, D. R. Ferreira, P. J. Carvalho, and JET Contributors, "Deep learning for plasma tomography using the bolometer system at jet," Fusion Engineering and Design **114**, 18–25 (2017).

[11]A. Jardin, J. Bielecki, D. Mazon, J. Dankowski, K. Król, Y. Peysson, and M. Scholz, "Neural networks: from image recognition to tokamak plasma tomography," Laser and Particle Beams **37**, 171–175 (2019).

[12]J. Svensson, "Non-parametric tomography using gaussian processes," Tech. Rep. (EFDA-JET-PR(11)24, 2011).

[13]C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proceedings of the 32nd International Conference on Machine Learning*, Proceedings of Machine

Learning Research, Vol. 37, edited by F. Bach and D. Blei (PMLR, Lille, France, 2015) pp. 1613–1622.

[14] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," (PMLR, New York, New York, USA, 2016) pp. 1050–1059.

[15] A. Pavone, J. Svensson, A. Langenberg, N. Pablant, U. Hoefel, S. Kwak, R. Wolf, and the Wendelstein 7-X Team, "Bayesian uncertainty calculation in neural network inference of ion and electron temperature profiles at w7-x," Review of Scientific Instruments **89**, 10K102 (2018).

[16] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434 (2015).

[17] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in neural information processing systems* (2016) pp. 2234–2242.

[18] J. Radon, "Über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten," Berichte über die Verhandlungen der Königlich-Sächsischen Gesellschaft der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse **69**, 262 (1917).

[19] R. M. Lewitt, "Reconstruction algorithms: transform methods," Proceedings of the IEEE **71**, 390–408 (1983).

[20] F. Matos, *Deep Learning for Plasma Tomography*, Master's thesis, Técnico Lisboa (2016).

[21] A. Murari, E. Joffrin, R. Felton, D. Mazon, L. Zabeo, R. Albanese, P. Arena, G. Ambrosino, M. Ariola, O. Barana, *et al.*, "Development of real-time diagnostics and feedback algorithms for jet in view of the next step," Plasma physics and controlled fusion **47**, 395 (2005).

[22] V. Igochine, A. Gude, M. Maraschek, and the ASDEX Upgrade Team, "Hotlink based soft x-ray diagnostic on ASDEX upgrade," Tech. Rep. IPP 1/338 (Max-Planck-Institut für Plasmaphysik, 2010).

[23] A. Jardin, J. Bielecki, D. Mazon, J. Dankowski, K. Król, Y. Peysson, and M. Scholz, "Synthetic x-ray tomography diagnostics for tokamak plasmas," Journal of Fusion Energy , 1–11 (2020).

[24] L. Ingesson, P. Böcker, R. Reichle, M. Romanelli, and P. Smeulders, "Projection-space methods to take into account finite beam-width effects in two-dimensional tomography algorithms," JOSA A **16**, 17–27 (1999).

[25] M. Odstrcil, J. Mlynar, T. Odstrcil, B. Alper, A. Murari, and JET contributors, "Modern numerical methods for plasma tomography optimisation," Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **686**, 156–161 (2012).

[26] M. Anton, H. Weisen, M. Dutch, W. Von der Linden, F. Buhlmann, R. Chavan, B. Marletaz, P. Marmillod, and P. Paris, "X-ray tomography on the tcv tokamak," Plasma physics and controlled fusion **38**, 1849

(1996).

[27] D. Carvalho, D. Ferreira, P. Carvalho, M. Imrisek, J. Mlynar, H. Fernandes, and J. Contributors, "Deep neural networks for plasma tomography with applications to jet and compass," Journal of Instrumentation **14**, C09011 (2019).

[28] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016) http://www.deeplearningbook.org.

[29] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition* (O'Reilly Media, Incorporated, 2019).

[30] D. Li, J. Svensson, H. Thomsen, F. Medina, A. Werner, and R. Wolf, "Bayesian soft x-ray tomography using non-stationary gaussian processes," Review of Scientific Instruments **84**, 083506 (2013).

[31] D. J. MacKay, "Comparison of approximate methods for handling hyperparameters," Neural computation **11**, 1035–1068 (1999).

[32] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning* (MIT press, 2006).

[33] D. MacKay, *Bayesian Methods for Adaptive Models*, Ph.D. thesis, Caltech (1991).

[34] J. Svensson, A. Werner, and JET-EFDA Contributors, "Current tomography for axisymmetric plasmas," Plasma Physics and Controlled Fusion **50**, 085002 (2008).

[35] C. Lechte, G. Conway, T. Görler, C. Tröster-Schmid, and the ASDEX Upgrade Team, "X mode doppler reflectometry k-spectral measurements in asdex upgrade: experiments and simulations," Plasma Physics and Controlled Fusion **59**, 075006 (2017).

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations* (2015).

[37] F. Chollet, "Keras," https://github.com/fchollet/keras (2015).

[38] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 37, edited by F. Bach and D. Blei (PMLR, Lille, France, 2015) pp. 448–456.

[39] X. Li, S. Chen, X. Hu, and J. Yang, "Understanding the disharmony between dropout and batch normalization by variance shift," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019) pp. 2677–2685.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980 (2014).

[41] G. James, *An introduction to statistical learning: with applications in R* (Springer, New York, NY, 2013).