Improved surrogates in inertial confinement fusion with manifold and cycle consistencies

Rushil Anirudh[a,1], Jayaraman J. Thiagarajan[a], Peer-Timo Bremer[a,b], and Brian K. Spears[c]

[a]Center for Applied Scientific Computing (CASC), Lawrence Livermore National Laboratory, Livermore, CA 94550; [b]Center for Extreme Data Management Analysis and Visualization (CEDMAV), University of Utah, Salt Lake City, UT 84112; and [c]Design Physics Division, Lawrence Livermore National Laboratory, Livermore, CA 94550

Edited by David A. Weitz, Harvard University, Cambridge, MA, and approved March 16, 2020 (received for review September 25, 2019)

Neural networks have become the method of choice in surrogate modeling because of their ability to characterize arbitrary, high-dimensional functions in a data-driven fashion. This paper advocates for the training of surrogates that are 1) consistent with the physical manifold, resulting in physically meaningful predictions, and 2) cyclically consistent with a jointly trained inverse model; i.e., backmapping predictions through the inverse results in the original input parameters. We find that these two consistencies lead to surrogates that are superior in terms of predictive performance, are more resilient to sampling artifacts, and tend to be more data efficient. Using inertial confinement fusion (ICF) as a test-bed problem, we model a one-dimensional semianalytic numerical simulator and demonstrate the effectiveness of our approach.

inertial confinement fusion | surrogate modeling | machine learning

Across scientific disciplines, researchers commonly design and evaluate experiments by comparing empirical observations with simulated predictions from numerical models. Simulations can provide insights into the underlying phenomena and are often instrumental to effective experiment design. Unfortunately, the most reliable, high-fidelity simulators are often too expensive to allow extensive calibration or parameter estimation. Hence, it is common to use ensembles of simulations to train a surrogate model that approximates the simulator over a large range of inputs, thereby enabling parameter studies as well as sensitivity analysis (1). Furthermore, one often fits a second—inverse—model to guide adaptive sampling and to identify parameters that drive the surrogate model into consistency with experiment.

Until recently, surrogate modeling has largely been restricted to one or at most a handful of scalar outputs. Consequently, scientists have been forced to distill their rich observational and simulated data into simple summary indicators or hand-engineered features such as the integral of an image, the peak of a time history, or the width of a spectral line. Such feature engineering severely limits the effectiveness of the entire analysis chain as most information from both experiments and simulations is either highly compressed or entirely ignored. Unsurprisingly, surrogate models designed to predict these features are often underconstrained, ill-conditioned, and not very informative.

Neural networks (NNs) have become a popular option to address this challenge due to their ability to handle more complex, multivariate datatypes, such as images, time series, or energy spectra. In a number of different application areas ranging from particle physics (1) to porous media flows (2) and many other scientific problems (2), NNs are able to effectively capture correlations across high-dimensional data signatures and produce high-quality surrogates, predictors, or classifiers. Inverse problems tend to be ill-posed, yet deep neural networks have shown remarkable progress in addressing challenging problems (3). Some notable examples are in imaging (4) and more recently leveraging novel regularizers such as structural priors (5, 6) or generative models (7, 8) for traditionally challenging inverse problems.

As a result there has been renewed interest in building better surrogates using neural networks for scientific problems. These include incorporating known scientific constraints into the training process (9, 10) or reducing dimensionality for better uncertainty quantification (11). However, surrogate forward models are often constructed in isolation such that they are inconsistent with an inverse model, leading to an implausible overall system in which the intuitive cycle of mapping inputs to outputs and back to inputs produces wildly varying results. Not only can an inverse prediction from the surrogate output be far away from the initial input, but even univariate sensitivities, i.e., inferring changes in predictions with respect to a single input parameter, are often unintuitive.

To address these issues, this paper advocates for the training of manifold and cyclically consistent (MaCC) surrogates using a multimodal and self-consistent neural network that outperforms the current state of the art on a wide range of metrics. Using a semianalytic model of inertial confinement fusion (ICF) (12, 13) as a test-bed problem, we propose a MaCC surrogate, containing two distinct components: 1) an autoencoding network to approximate the low-dimensional latent manifold and to accurately capture the correlations between multimodal outputs of a simulator, i.e., multiple images and a set of scalar quantities, and 2) an inverse (or pseudoinverse because of the ill-posed nature) neural network that trains alongside the surrogate network. Cyclical consistency has emerged as a powerful regularization technique in unsupervised problems in the past few years (14–16), improving the state of the art in a variety of applications including image-to-image translation (14), domain adaptation (17), visual

Significance

Neural networks have demonstrated remarkable success in predictive modeling. However, when applied to surrogate modeling, they 1) are often nonrobust, 2) require large amounts of data, and 3) are inadequate for estimating the inversion process; i.e., they do not capture parameter sensitivities well. We propose a different form of self-consistency regularization by incorporating an inverse surrogate into the learning process and show that it leads to highly robust, self-consistent surrogate models for complex scientific applications.

Author contributions: R.A., J.J.T., P.-T.B., and B.K.S. designed research; R.A. performed research; R.A., J.J.T., and B.K.S. contributed new reagents/analytic tools; R.A. analyzed data; and R.A., J.J.T., P.-T.B., and B.K.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1]To whom correspondence may be addressed. Email: anirudh1@llnl.gov.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1916634117/-/DCSupplemental.

First published April 20, 2020.

question answering (18), and voice conversions (19). We propose a direct coupling between forward and inverse models to enforce cyclical consistency, which regularizes the training to produce higher-fidelity and more robust models.

Main Findings

We find that manifold consistency significantly improves the predictive capabilities, while the cycle consistency helps in smoothing the high-dimensional function space in the outputs, resulting in improved resilience to sampling artifacts and data scarcity. Surprisingly, we find that cyclical consistency generalizes even to other inverse models (from data bootstraps) not accessed during training, demonstrating a tight coupling between the input and output spaces.

Surrogate Design for ICF

In any surrogate-based technique, the challenge is to build a high-fidelity mapping from the process inputs, say target and laser settings for ICF, to process outputs, such as ICF implosion neutron yield and X-ray diagnostics. Developing surrogates in the ICF context is particularly challenging. The physics of ICF fusion ignition are predicated on interactions between multiple strongly nonlinear physics mechanisms that have multivariate dependence on a large number of controllable parameters. This presents the designer with a complicated response function that has sharp, nonlinear features in a high-dimensional input space. While this is challenging, deep neural network solutions have made building surrogates for scalar-valued outputs relatively routine (20). However, to take full advantage of the rich range of diagnostic techniques, we require surrogates that can also replicate a wide range of array-valued image data. In ICF, the images can be produced by different particles (X-rays, neutrons) at different energies (hyperspectral), at different times, and from different lines of sight. These complicated modalities are more difficult to ingest, and techniques for learning them can introduce large model capacity and an associated need for excessive amounts of data. Thus, our principal design task is to develop a neural network surrogate that can handle multiple data modalities, can produce predictions acceptable for precision physics, and can be trained without requiring unreasonably large amounts of data.

Predictive Surrogates with Neural Networks

Formally, the surrogate modeling problem is defined as follows: Given a set of input parameters, $X \subset \mathcal{X}$ (obtained with an experiment design of choice, e.g., Latin hypercube sample), and the corresponding observations or outputs from the simulator, $Y \subset \mathcal{Y}$, where Y denotes a collection of images (Y_{img}) and scalar quantities (Y_{sca}), the task is to determine a function $\mathcal{F} : X \mapsto Y$, such that a user-defined measure of predictive accuracy, i.e., mean squared error (MSE), is minimized. Here, $\mathcal{X}$ and $\mathcal{Y}$ refer to the space of inputs and outputs, respectively. We refer to $\mathcal{F}$ as the forward model and the reverse process, $\mathcal{G} : Y \mapsto X$, as the inverse model. In many scientific problems a functional inverse may not exist because of the ill-posed nature of the problem, and in such cases we refer to $\mathcal{G}$ as a pseudoinverse. In recent years, deep neural networks have emerged as the most powerful predictive modeling tool because of their ability to approximate nonlinear and high-dimensional functions. Neural networks are modeled as a series of weights and nonlinearities that take the input parameters while predicting the outputs. They are most commonly optimized using stochastic gradient descent (SGD) with a loss function such as MSE.

In this paper, we propose two consistency requirements to improve surrogate modeling: first, a manifold consistency that ensures the predictions are physically meaningful and, second, a notion of cyclical consistency (14, 15) between the forward and inverse models. For the former, we use an autoencoder to embed all output quantities into a low-dimensional manifold, $\mathcal{Z}$, and repose surrogate modeling as $\mathcal{F} : X \mapsto Z$, i.e., to predict into the latent space in lieu of Y. To enforce the cycle consistency, we propose to penalize predictions of the forward model that are "inconsistent" with the inverse model. In other words, a prediction from the forward model, when put through the inverse $\mathcal{G}$, must give back the initial set of parameters; i.e., $\mathcal{G}(\mathcal{F}(X)) \approx X$. In the context of unsupervised image–image translation, cycle consistency has been shown to be an effective regularization technique (14, 15). On the contrary, our inverse formulation uses paired examples, yet suffers from severe ill-posedness. Both consistencies are illustrated in Fig. 1 and described in detail in the next section.

Notations

Since we have several networks interacting with each other, we clarify our notation for the rest of this paper. We refer to the inputs corresponding to a set of samples by matrix X, while each sample is denoted as $\mathbf{x}$. Similarly, the collections of outputs and latent representations are denoted as Y and Z, while their individual realizations are $\mathbf{y}$ and $\mathbf{z}$, respectively. The predictions from the trained models $\mathcal{F}$ and $\mathcal{G}$ are referred to as $\hat{\mathbf{y}}$ and $\hat{\mathbf{x}}$. Finally, we denote a cyclical prediction, i.e., $\mathbf{x} \to \hat{\mathbf{y}} \to \hat{\hat{\mathbf{x}}}$, with a double hat indicating predictions from both the forward and the inverse.

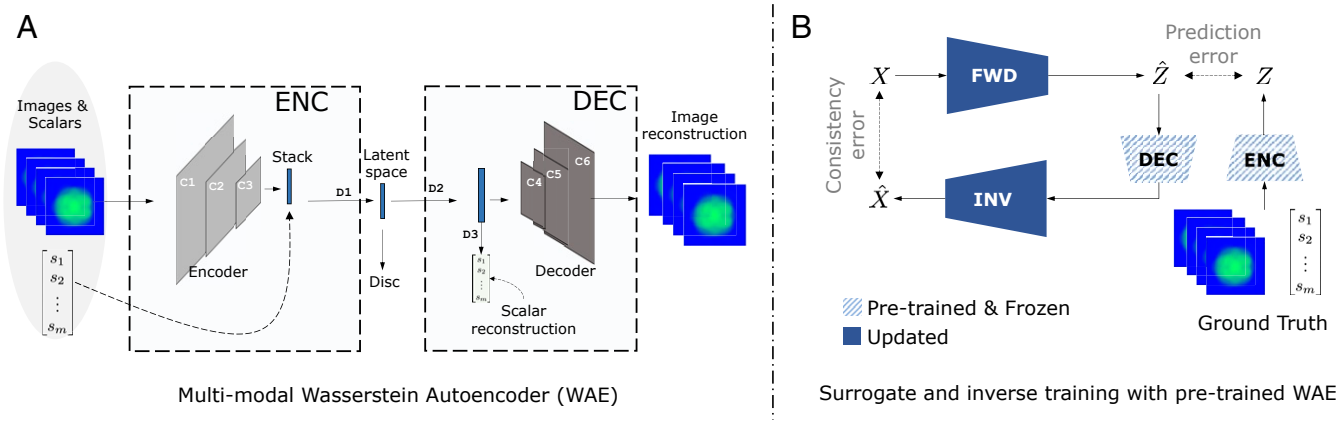

Fig. 1. MaCC surrogates. The proposed architecture uses a pretrained autoencoder (*A*) for ensuring manifold consistency and an inverse model (*B*) for cyclical consistency and robustness. ENC, encoder; DEC, decoder; FWD, forward; INV, inverse.

Methods

Multimodal Prediction Using an Autoencoder. Exploiting the correlation between multimodal outputs should lead to a better forward model because it disambiguates simulations that may otherwise appear similar in some aggregated response function. A straightforward multimodal forward model $\mathcal{F}: \mathrm{X} \rightarrow \mathrm{Y}$ has access to the correlation structure present in $\mathcal{Y}$, but the task of inferring the correlations from training data is combined with learning the forward model. Instead, MaCC splits both tasks by first designing an autoencoding neural network to capture the correlation and then explicitly utilizing this information to the forward model by predicting into the inferred latent space. We jointly infer an encoder $\mathcal{E}: \mathrm{Y} \mapsto \mathrm{Z}$ to map a multimodal observation onto the latent vector $\mathbf{z} \in \mathrm{Z}$ and a decoder $\mathcal{D}: \mathrm{Z} \mapsto \mathrm{Y}$ that reconstructs the multimodal outputs from the latent representation.

Design. As shown in Fig. 1*A*, the output space in our setup is composed of a set of images (treated as different channels) and diagnostic scalars $(s_1, \cdots s_m)$. The encoder is split into two branches: one that uses a convolution neural network to encode image features and another with fully connected layers to process the set of scalars. Both branches are then merged (by concatenation) using another fully connected layer to capture the relationships between image features and scalars. The joint output layer forms the overall latent representation and serves as a compressed description of the output space. The decoder is built symmetrically to reconstruct the original outputs. In addition to aiming for a high-fidelity reconstruction at the decoder, we encourage the latent space to be approximately uniform by placing a statistical prior in the latent space. This is a form of a Wasserstein autoencoder (WAE) (21) which reduces statistical dependencies between latent factors and helps to regularize the autoencoder training. It also enables us to sample from the latent space efficiently after training. Mathematically, this is achieved by placing a uniform prior $p(\mathbf{z})$ in the latent space and ensuring that the discrepancy $\mathcal{H}(p(\mathbf{z}), q(\mathbf{z}|\mathbf{x}))$ is minimized, where $\mathcal{H}$ denotes a suitable divergence measure.

Since the exact parameterization of $q(\mathbf{z}|\mathbf{x})$ is unknown, we adopt an adversarial training strategy (two-sample test) that uses an additional discriminator network to ensure that one cannot distinguish between the generated latent representations and realizations from a uniform distribution. Formally, the training objective $\mathcal{L}_{ae}$ can be written as

$$\sum_{\mathbf{y} \in Y} \|\mathbf{y}_{img} - \hat{\mathbf{y}}_{img}\|_2^2 + \gamma_s \|\mathbf{y}_{sca} - \hat{\mathbf{y}}_{sca}\|_2^2 + \gamma_a \mathcal{L}_{adv}, \qquad [1]$$

$$\text{where } \mathbf{z} = \mathcal{E}(\mathbf{y}_{img}, \mathbf{y}_{sca}), \text{ and } \hat{\mathbf{y}}_{img}, \hat{\mathbf{y}}_{sca} = \mathcal{D}(\mathbf{z}),$$

and $\mathcal{L}_{adv}$ is the discriminator cross-entropy loss that attempts to classify the latent representation as arising from a fake distribution, while assuming the real distribution to be uniform random (21). γ_s is a weight chosen to adjust the bias toward images, and we fix it at $\gamma_s = 1 \times 10^2$, and $\gamma_a = 1 \times 10^{-3}$. Given a pretrained autoencoder, we encode all training data to form $(\mathbf{x}, \mathbf{z})$ pairs and reformulate the surrogate as learning $\mathcal{F}: X \mapsto Z$.

Cyclical Regularization in Surrogates. While the surrogate model introduced above performs well, it is important to recognize a number of implicit assumptions in the process and consider how they might affect the quality of the model. One of the most important and often disregarded assumptions is the choice of loss function used to construct $\mathcal{F}$. We formulate the training objective for the surrogate as

$$\min_{\mathcal{F}} \rho\left(\mathcal{F}(\mathbf{x}; \theta) - \mathbf{z}\right), \qquad [2]$$

where ρ denotes a measure of fidelity and $\mathcal{F}$ represents the parameterized surrogate model with parameters θ. Partially for convenience and partially due to a lack of prior knowledge on the residual structure, ρ is often chosen to be an ℓ_p norm. This implicitly assumes that the data manifold, i.e., the space of all outputs $\mathcal{F}(\mathbf{x})$ for $\mathbf{x} \in \mathrm{X}$, is Euclidean which is most certainly not the case. Furthermore, the choice of norm also assumes a distribution of discrepancies between the model and the ground truth. Specifically, if we express $\mathcal{F}(\mathbf{x}) = \mathcal{F}^*(\mathbf{x}) + \epsilon(\mathbf{x})$, where $\mathcal{F}^*$ is the ground-truth mapping, then choosing, for example, the ℓ_2 norm is implicitly assuming that ϵ follows a Gaussian distribution. In practice, neither the Euclidian space nor the Gaussian error assumptions are likely to be correct. However, designing a more appropriate and robust loss function in the latent space is difficult especially for the complex, multimodal data of interest here. Accordingly, we propose a regularization strategy based on self-consistency to produce more generalizable and robust forward models.

Conceptually, the challenge in using [2] to define $\mathcal{F}$ is twofold: First, since we cannot build a customized ρ and the space of θs is large, there likely exist many different $\mathcal{F}_i$s with an acceptable error that may represent physically better surrogates than the chosen $\mathcal{F}$. Second, the true error is unlikely to be isotropic, meaning some deviations from $\mathcal{F}^*$ are more plausible or less damaging than others. To choose among these $\mathcal{F}$s we impose a cycle consistency requirement defined as follows: We jointly train a pseudoinverse of $\mathcal{F}^*$, i.e., $\mathcal{G}: Y \mapsto X$, and introduce a regularization term $\delta(\mathcal{F}, \mathcal{G})$ computed as

$$\delta(\mathcal{F}, \mathcal{G}) = \sum_{\mathbf{x} \in X, \mathbf{z} \in Z} \|\mathbf{z} - \hat{\hat{\mathbf{z}}}\|_2^2 + \|\mathbf{x} - \hat{\hat{\mathbf{x}}}\|_2^2, \qquad [3]$$

where $\hat{\hat{\mathbf{z}}} = \mathcal{F}(\mathcal{G}(\mathcal{D}(\mathbf{z})))$ and $\hat{\hat{\mathbf{x}}} = \mathcal{G}(\mathcal{D}(\mathcal{F}(\mathbf{x})))$ are the cyclical predictions for $\mathbf{z}$ and $\mathbf{x}$, respectively. Note that different from $\mathcal{F}$, the pseudoinverse takes the decoded outputs Y instead of Z. The use of the ℓ_2 norm in Eq. **3** still makes the Euclidean assumption, but is more appropriate in the latent space Z, which is trained to be close to a full-dimensional, Euclidean space [although this cannot be guaranteed (22)]. We also expect the cyclical regularization to account for some of the nonisotropic error behavior. The cycle regularization directly in the data (or pixel) space can be unstable when the mapping between the two domains is not isomorphic, as is likely the case in a surrogate problem. Although this problem still persists, it is mitigated to a large extent by including cycle regularization in the latent space instead (similar observations have been reported by ref. 23 for image translation tasks). We explore this further in *Experiments and Results*.

Consequently, the optimization objective for MaCC surrogates can be expressed as

$$\min_{\mathcal{F}, \mathcal{G}} \rho\left(\mathcal{F}(\mathbf{x}; \theta) - \mathbf{z}\right) + \lambda_{cyc} \delta(\mathcal{F}, \mathcal{G}). \qquad [4]$$

Note that in general $\mathcal{G}$ cannot be a true inverse since $\mathcal{F}^*$ might not be bijective. In this case constructing $\mathcal{G}$ as a function, i.e., a neural network, induces a mode collapse in the estimated posterior $p(\mathbf{x}|\mathbf{z})$. However, we see that even a pseudoinverse $\mathcal{G}$ encodes a better local residual structure than $\mathcal{F}$ alone.

In this context, the bidirectional consistency penalty in Eq. **3** encourages the surrogate $\mathcal{F}$ to be consistent with the pseudoinverse in different ways. The first term is not affected by the mode collapse in the inverse since it is entirely computed in the output space alone. As a result, it encourages the high-dimensional output function to be smoothly varying, while the second term constrains the forward model to make predictions closer to the data manifold.

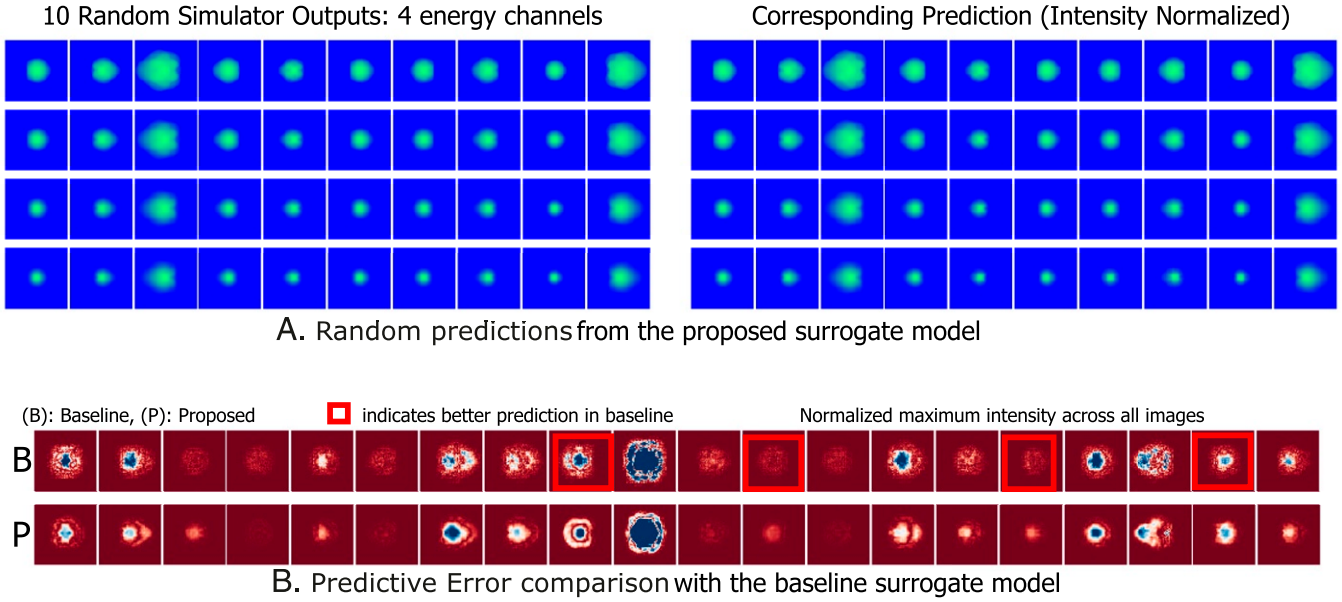

Fig. 2. (*A*) The proposed model is able to match the simulator's prediction very closely, across all of the four energy bands. Here we show a random sample comparing the simulator's outputs to predictions from a MaCC surrogate. (*B*) Residual images (absolute), with respect to the ground truth, for 16 examples (only one energy band shown). The intensities of images for both the baseline (B) and MaCC (P) are normalized to a global scale. Except for a small number of cases (highlighted with red border), MaCC produces improved quality predictions, when compared to the baseline.

We observe that due to the ill-conditioned nature of the inverse problem, a neural network takes significantly longer to converge than the forward network. To address this challenge, we first pretrain the inverse network; i.e., we train a standalone pseudoinverse neural network until convergence. We then load this pretrained model and resume training with the forward model which is trained from scratch using the cyclical consistency. This process is sometimes referred to as a "warm start." During cyclic training, the pseudoinverse continues to train with the loss

$$\min_{\mathcal{G}} \sum_{\mathbf{z}\in Z} \rho\left(\mathcal{G}(\mathcal{D}(\mathbf{z});\theta_I)-\mathbf{x}\right)+\lambda_{cyc}\|\mathbf{z}-\hat{\mathbf{z}}\|_2^2, \qquad [5]$$

where θ_I is the set of parameters of $\mathcal{G}$, and the other terms are the same as in Eq. **3**. Note that optimizing $\mathcal{F}$ according to Eq. **4** necessarily biases the model toward a particular pseudoinverse $\mathcal{G}$. However, as is discussed in more detail below, the resulting $\mathcal{F}$ is highly consistent with a diverse set of $\mathcal{G}$s, different from the one used during training, constructed by bootstrapping the training data. In other words, by including the consistency regularization, the surrogate $\mathcal{F}$ converges to a solution where the resulting residuals are better guided by the characteristics of $\mathcal{G}$. This achieves the same effect as explicitly constructing a specialized loss function ρ to better fit the data characteristics. As we show in our experiments, surrogates obtained using existing neural network solutions are inconsistent with the inverse model and result in nonsmooth, nonrobust models in practice.

A New Self-Consistency Test for Surrogates. Given the limitations of commonly used error metrics in surrogate evaluation, we introduce a metric for surrogate fidelity that couples the performance of both the forward and inverse models. We create a test set by varying only a single input parameter using a linear scan of 100 steps (from min to max), while fixing all other parameters. These 100 samples are then passed through the forward model and subsequently through the inverse model before obtaining back input parameter predictions. We check whether the predictions are consistent with the "ground truth," i.e., the linear scan. This is conceptually similar to partial dependency tests in statistics and effectively captures sensitivities of the forward and inverse models.

Given the underdetermined nature of the inverse process, it is possible that the achieved self-consistency is biased by the specific solution of $\mathcal{G}$. Hence, we propose to evaluate the consistency with respect to different solutions from the space of possible pseudoinverse models. To this end, we use multiple random subsets of the original training set (bootstraps) and obtain independent estimates of $\mathcal{G}$. We find that the cyclical consistency remains valid for MaCC across all of these models, indicating that the self-consistency achieved is actually statistically meaningful. The consistency measure is given by

$$\mathcal{L}_c = \sum_{i=1}^{5} R^2(\mathbf{x}_{scan}, \mathcal{G}_i(\mathcal{D}(\mathcal{F}(\mathbf{x}_{scan})))). \qquad [6]$$

Here R^2 denotes the R-squared statistic and $\mathcal{G}_i$ corresponds to the inverse model inferred from the ith bootstrap.

Experiments and Results

Dataset. Our training dataset is composed of input parameter settings and the corresponding outputs from the semianalytical ICF simulator described in ref. 12, where each output is a

Table 1. Surrogates with MaCC show superior predictive performance as measured by mean squared error

Metric	Baseline (no MaCC)	Baseline + MaCC
Mean R^2 scalars	**0.9990**	**0.9974**
MSE image (band 0)	0.0476 ± 0.0449	**0.0351 ± 0.0296**
MSE image (band 1)	0.0458 ± 0.0446	**0.0374 ± 0.0371**
MSE image (band 2)	0.08745 ± 0.1355	**0.0736 ± 0.1236**
MSE image (band 3)	0.2035 ± 0.4441	**0.1742 ± 0.4010**

Here we use a cyclical weight $\lambda_{cyc} = 0.05$. Boldface indicates better performance.

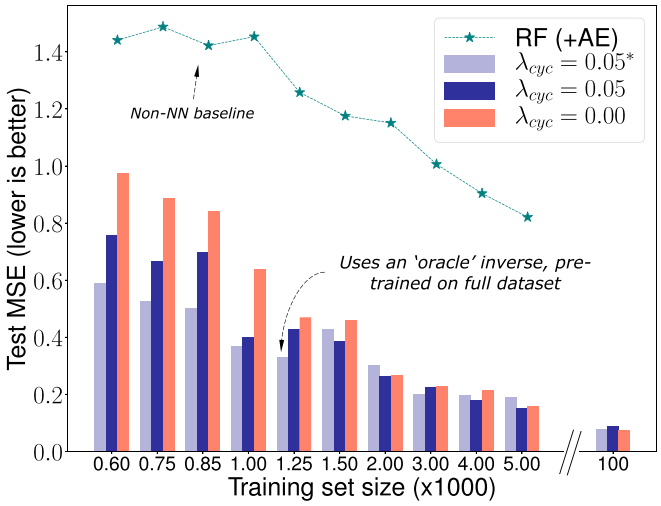

Fig. 3. Cycle consistency results in improved generalization with fewer training samples. RF (+AE) is a non-NN baseline that uses random forest with the auto-encoder.

collection of four multienergy images sized 64×64 and 15 diagnostic scalar quantities such as yield, ion temperature, pressure, etc. Our dataset was constructed as a random subset (100,000 samples) of a Latin hypercube experiment design containing 1 million samples in the five-dimensional input parameter space. All model evaluation is carried out using a held-out 10,000 validation set, which contains no overlap with the training set. Next, we describe the training strategies adopted for different components of a MaCC surrogate in our experiments. All models were trained using the Adam Optimizer (24), with the learning rate set at 1×10^{-4} and the minibatch size fixed at 128 samples. The architectures for all of the models are included in *SI Appendix*.

Experimental Details. First, we train the autoencoder with a 32-dimensional latent space until convergence requiring about 600 epochs. Additionally, we use a pretrained inverse that is trained for about 2,500 epochs. The architectural details for all networks are available in *SI Appendix*.

Baselines. We compare the performance of the surrogate across all of the proposed metrics with several baselines which we describe next: 1) For non-NN baseline, we train an extremely randomized tree model that predicts directly into the latent space, Z, coupled with the pretrained decoder. This is similar to recent work (20) in ICF where they use decision trees to initialize a surrogate that maps only to scalars. 2) For NN baseline, we consider an NN baseline (trained with and without cycle consistency) that takes in the inputs and predicts the images via two separate networks. We construct a baseline with similar architecture, with approximately the same number of parameters, the main difference being that it does not use the manifold consistency. In addition, we also create other baselines using ablation studies of the λ_{cyc} parameter, keeping the architecture exactly the same. More details about the baselines are in *SI Appendix*.

Results.

Qualitative evaluation. Fig. 2*A* shows random samples from the simulator and their corresponding predictions obtained using our surrogate, demonstrating that MaCC captures details very accurately, across the four energy channels. Next, Fig. 2*B* illustrates the residual error images for 20 randomly chosen examples (only one energy band shown) obtained using predictions from the baseline and MaCC. All images are intensity normalized by the same maximum intensity value. In most cases, MaCC predicts higher-quality outputs, where smaller residuals indicate higher-fidelity predictions.

We evaluate the quantitative performance of the surrogates using widely adopted metrics, namely MSE and R^2. More specifically, we report the following quantities: 1) for mean R^2 scalars, average coefficient of determination (R^2 statistic) across the 15 scalar outputs, and 2) for MSE image (band), mean squared error of prediction for the entire 10,000 test set, in each of the energy bands. The results are shown in Table 1, where we include the performance of the baseline approach and MaCC with $\lambda_{cyc} = 0.05$. From the results for image prediction, it is evident that MaCC outperforms the baseline neural network solution. In contrast, it is fairly straightforward to predict the scalar diagnostic outputs, with both models achieving an R^2 score of $\sim$0.99. Comparisons across more baselines and ablation studies are shown in Fig. 4*A*.

Cycle-consistency score. We show the results for one particular pseudoinverse trained with a random 50% of the training data. The results for other cases are reported in *SI Appendix*. In Fig. 4*A*, we show how cyclical regularization impacts the quality of the surrogate model, against its tendency to be self-consistent. We observe that a small λ_{cyc} does not adversely affect the quality of the surrogate model as measured by mean

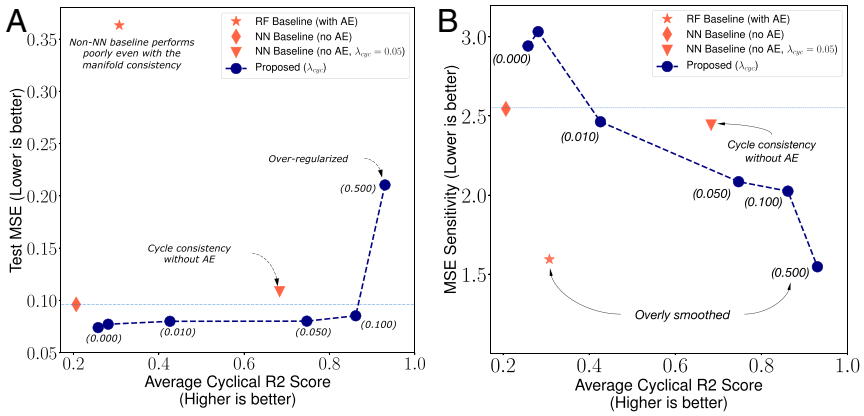

Fig. 4. (*A*) Ablation study of λ_{cyc} and mean squared error. A higher weight leads to more cyclically consistent predictions. Except for extreme cases, the training is fairly robust to values of λ_{cyc}, leading to a better performance than the baseline. (*B*) Cyclic consistency results in robustness to small local perturbations, as a result of smoothing the high-dimensional output prediction space. This also leads to better predictions in smaller data regimes as seen in Fig. 3.

squared error. As can be seen, until $\lambda_{cyc} = 0.10$ all of the models consistently perform better than the baseline. However, with a significant weight, the model tends to underfit, resulting in a higher MSE.

Benefits of Cyclical Consistency. Cyclical consistency acts as a regularization technique that helps in smoothing out the prediction space, and as a result we expect to see gains in predictive performance of the forward model when there are fewer training data available, as well as in improved robustness to perturbed inputs. We see both of these to be the case and discuss the results next.

Behavior in small data regimes. We observe improved predictive performance of the forward model when there are significantly fewer training samples, as shown in Fig. 3. We train different surrogates while providing access only to a fraction of the training set. It must be noted that the autoencoder is used in this experiment, which has been trained on the $100{,}000$ dataset, but it is unsupervised; i.e., it only approximates the physics manifold without any knowledge of the forward process. We evaluate the performance of all models on the same $10{,}000$ validation set as before to make them comparable Additionally, we show generalization when an "oracle" inverse is available, in which the inverse has access to the entire dataset as an upper bound. The benefit makes it clear that the inverse has useful gradients to improve the quality of the forward model, sometimes reducing prediction error by nearly $\sim 30\%$.

Robustness to sampling artifacts. At test time, we add a small amount of uniform random noise, $\hat{\mathbf{x}} = \mathbf{x} + \sigma * \mathcal{U}$ to the five input parameters, and measure how much the output has changed with regard to the ground-truth value at x. This is a measure of how smooth the predictions in the output (image) space are. Particularly of relevance to surrogates of scientific models, we expect the function value to change gradually in regions where there are few or no samples around a given test sample. This can be useful in scenarios with sampling artifacts or a poor design of experiments. We observe that cyclical consistency has a direct impact on the smoothness of the predictions as shown in Fig. 4*B*. On the *y* axis we show the sensitivity to local perturbations, i.e., the difference in MSE between $\mathcal{F}(\mathbf{x})$ and $\mathcal{F}(\hat{\mathbf{x}})$, with the consistency measure described in Eq. **6** on the x axis. We observe that the cyclical regularization results in significantly more robust models, while having very similar prediction errors on clean data, as seen in Fig. 4*A*. To ensure that the perturbations are not extreme, we pick $\sigma = 0.1$ for all samples. This was chosen by ensuring that the distance of the clean test set to the perturbed one is smaller than its distance of the nearest neighbor in the training set.

Discussion. In this paper, we introduced MaCC surrogates, which contain two distinct elements: a pretrained autoencoder that enforces the surrogate to map input parameters to the latent space, i.e., $\mathrm{X} \mapsto \mathrm{Z}$ instead of the traditional $\mathrm{X} \mapsto \mathrm{Y}$, and a pseudoinverse trained alongside the surrogate with a cyclical consistency objective, which encourages the predictions from $\mathcal{G}(\mathcal{F}(x))$ to be close to the input x. These properties lead to robust, data-efficient, and interpretable surrogates, which are properties critical for surrogate models in scientific applications.

1. M. Paganini, L. de Oliveira, B. Nachman, Calogan: Simulating 3d high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks. *Phys. Rev. D* **97**, 014021 (2018).
2. Y. Zhu, N. Zabaras, Bayesian deep convolutional encoder–decoder networks for surrogate modeling and uncertainty quantification. *J. Comput. Phys.* **366**, 415–447 (2018).
3. L. Ardizzone, J. Kruse, C. Rother, U. Köthe, "Analyzing inverse problems with invertible neural networks" in *International Conference on Learning Representations* (2019). https://openreview.net/forum?id=rJed6j0cKX. Accessed 7 April 2020.
4. K. H. Jin, M. T. McCann, E. Froustey, M. Unser, Deep convolutional neural network for inverse problems in imaging. *IEEE Trans. Image Process.* **26**, 4509–4522 (2017).
5. D. Ulyanov, A. Vedaldi, V. Lempitsky, "Deep image prior" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 9446–9454.
6. A. Shocher, N. Cohen, M. Irani, "'Zero-shot' super-resolution using deep internal learning" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 3118–3126.
7. R. A. Yeh *et al.*, "Semantic image inpainting with deep generative models" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), pp. 5485–5493.
8. A. Bora, A. Jalal, E. Price, A. G. Dimakis, "Compressed sensing using generative models" in *Proceedings of the 34th International Conference on Machine Learning* (PMLR, 2017), vol. 70, pp. 537–546.
9. Y. Zhu, N. Zabaras, P.-S. Koutsourelakis, P. Perdikaris. Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data. *J. Comput. Phys.* **394**, 56–81 (2019).
10. M. Raissi, P. Perdikaris, G. E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707 (2019).
11. R. K. Tripathy, I. Bilionis, Deep UQ: Learning deep neural network surrogate models for high dimensional uncertainty quantification. *J. Comput. Phys.* **375**, 565–588 (2018).
12. J. Gaffney, P. Springer, G. Collins, "Thermodynamic modeling of uncertainties in NIF ICF implosions due to underlying microphysics models" in *APS Division of Plasma Physics Meeting* (APS, 2014) http://meetings.aps.org/link/BAPS.2014.DPP.PO5.11. Accessed 7 April 2020.
13. A. L. Kritcher *et al.*, Metrics for long wavelength asymmetries in inertial confinement fusion implosions on the national ignition facility. *Phys. Plasmas* **21**, 042708 (2014).
14. J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks" in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE, 2017), pp. 2223–2232.
15. Z. Yi, H. Zhang, P. Tan, M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation" in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE, 2017), pp. 2849–2857.
16. Y. Choi *et al.*, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 8789–8797.
17. J. Hoffman *et al.*, "Cycada: Cycle-consistent adversarial domain adaptation" in *International Conference on Machine Learning* (PMLR, 2018), **vol. 80**, pp. 1989–1998.
18. M. Shah, X. Chen, M. Rohrbach, D. Parikh, "Cycle-consistency for robust visual question answering" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2019), pp. 6649–6658.
19. H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks" in *2018 IEEE Spoken Language Technology Workshop* (*SLT*) (IEEE, 2018), pp. 266–273.
20. K. D. Humbird, J. L. Peterson, R. G. McClarren, Deep neural network initialization with decision trees *IEEE Trans. Neural Networks Learning Systems* **30**, 1286–1295 (2018).
21. I. Tolstikhin, O. Bousquet, S. Gelly, B. Schoelkopf, "Wasserstein auto-encoders" in International Conference on Learning Representations (ICLR, 2018). https://openreview.net/forum?id=HkL7n1-0b. Accessed 7 April 2020.
22. G. Arvanitidis, L. K. Hansen, S. Hauberg, "Latent space oddity: On the curvature of deep generative models" in *International Conference on Learning Representations* (2018). https://openreview.net/forum?id=SJzRZ-WCZ. Accessed 7 April 2020.
23. M. Binkowski, D. Hjelm, A. Courville, "Batch weight for domain adaptation with mass shift" in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE, 2019), pp. 1844–1853.
24. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv:1412.6980 (22 December 2014).