Transfer Learning to Model Inertial Confinement Fusion Experiments

K. D. Humbird, J. L. Peterson, B. K. Spears, and R. G. McClarren

***Abstract*— Inertial confinement fusion (ICF) experiments are designed using computer simulations that are approximations of reality and therefore must be calibrated to accurately predict experimental observations. In this article, we propose a novel technique for calibrating from simulations to experiments, or from low fidelity simulations to high fidelity simulations, via "transfer learning" (TL). TL is a commonly used technique in the machine learning community, in which models trained on one task are partially retrained to solve a separate, but related task, for which there is a limited quantity of data. We introduce the idea of hierarchical TL, in which neural networks trained on low fidelity models are calibrated to high fidelity models, then to experimental data. This technique essentially bootstraps the calibration process, enabling the creation of models which predict high fidelity simulations or experiments with minimal computational cost. We apply this technique to a database of ICF simulations and experiments carried out at the Omega laser facility. TL with deep neural networks enables the creation of models that are more predictive of Omega experiments than simulations alone. The calibrated models accurately predict future Omega experiments, and are used to search for new, optimal implosion designs.**

***Index Terms*— Inertial confinement, neural networks.**

I. Introduction

MANY physical systems and experiments are designed using models—analytical theories or computer simulations that attempt to take into account the various components of the system to determine the efficiency, performance, and reliability of a design. In applications where the physics of the system is well known, the models can be accurate depictions of reality, however, when dealing with systems at extreme conditions (such as extremely high temperatures, pressures, and densities), the physics is not so well understood and the models are not always validated with experimental data. This is often the case in inertial confinement fusion (ICF), in which lasers are used to compress small fuel capsules filled with deuterium and tritium to high density, temperature, and pressure in order to create conditions that are favorable for nuclear fusion reactions [1], [2]. The computer simulations that model ICF experiments are complex and involve a wide variety of physics models: radiation hydrodynamics, atomic physics, nuclear burn physics, laser and plasma physics, magnetic field effects, and more [3]. These codes are validated in certain regimes, but acquiring data at the extreme conditions reached in ICF experiments is challenging and expensive, thus the accuracy of the models away from the validation data is not well known. Furthermore, fully integrated simulations which model everything from the laser beam propagation to the particle transport within the fuel capsule are extremely expensive to run in 3-D with high resolution. Researchers often need to make many approximations, such as running the simulation in 2-D with axi-symmetric constraints, in order to efficiently search the parameter space for promising experimental designs. The use of surrogate models trained on large databases of simulations have enabled rapid exploration of design spaces [4], [5], but these models are only as good of a representation of reality as the simulations upon which they are trained.

When the simulations used to design new ICF experiments contain simplifying assumptions, it is expected that there will be discrepancies between the simulation prediction and what is observed in the experiment. A common statistical approach to correcting an inaccurate simulator is model calibration—using experimental data to "calibrate" an inaccurate model to produce one which is more consistent with reality. Model calibration is a broadly researched topic [6]–[9], with one of the most popular techniques developed by Kennedy and O'Hagan [10]. In this approach, the true model is assumed to be an additive combination of a simulator, Gaussian-distributed error due to measurement uncertainty, and an unknown discrepancy term which is learned using experimental data. The form of this discrepancy term is often specified by the user (for example, the user might choose a Gaussian process discrepancy with a particular kernel function) and the complexity is limited by the amount of experimental data that is available. Researchers have explored Bayesian calibration with discrepancy terms for ICF data [11]; in this article, we propose an alternative calibration technique borrowed from the machine learning community, referred to as "transfer learning" (TL), for calibrating ICF models.

Manuscript received March 8, 2019; revised June 17, 2019 and September 25, 2019; accepted November 15, 2019. Date of publication December 13, 2019; date of current version January 20, 2020. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Released as LLNL-CONF-764063. The review of this article was arranged by Senior Editor S. J. Gitomer. *(Corresponding author: K. D. Humbird.)*

K. D. Humbird is with the Lawrence Livermore National Laboratory, Livermore, CA 94550 USA, and also with the Department of Nuclear Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: humbird1@llnl.gov).

J. L. Peterson and B. K. Spears are with the Lawrence Livermore National Laboratory, Livermore, CA 94550 USA.

R. G. McClarren is with the Department of Aerospace and Mechanical Engineering, University of Notre Dame, Notre Dame, IN 46556 USA.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPS.2019.2955098

Traditional machine learning models gain knowledge by observing large quantities of labeled data. For example, if the task is to classify photographs of animals, a model will need to be exposed to millions of labeled images of all the animals it is expected to classify, in a variety of scenarios, colors, perspectives, etc., in order to classify the animals correctly. Supervised learning tasks are straightforward to solve when large quantities of data are available, however, many of these techniques fail when limited to small sets of labeled data.

TL is an alternative learning technique that can help overcome the challenge of training on small data sets. TL is a method for using knowledge gained while solving one task, and applying it to a different, but related task. This approach is most commonly used for image classification [12]–[14], for which there are many large databases of labeled images [15]–[17] and several pretrained neural network models that are available for download, such as AlexNet [18] and Inception [19]. These neural networks have been studied extensively, and they appear to learn how to recognize images in a logical series of steps as you traverse hidden layers in the models. First, the networks often search for edges in the images, then for simple geometric patterns, and eventually begin to recognize specific characteristics, such as eyes, arms, ears, etc. [19]. In general, the deeper in the network you go, the finer the details the network appears to focus on. It therefore seems logical to expect that for a neural network trained on any image data set, the first several hidden layers are essentially the same—they learn about features common to all images. One can thus take a pretrained neural network, freeze the first several layers of weights, and focus on retraining only the last few layers on a new image data set to learn how to appropriately classify this new set of images. The old frozen layers are where the network learns to "see"; the last few are where the network learns to "recognize" specific types of images. TL is the process of taking a network trained on a large data set, freezing several layers of the network, then retraining the unfrozen layers on a different, often smaller data set. As an example relevant to ICF, researchers at the National Ignition Facility (NIF) [20] have used TL to classify images of different types of damage that occur on the optics at the NIF. There are not enough labeled optics images to train a network from scratch, but TL with a network pretrained on ImageNet [15] produces models that classify optics damage with over 98% accuracy. This methodology has enabled the group to automate their damage inspection by letting the network process the images of optical components, rather than having an optics expert inspect each image manually [21].

In ICF, it is often the case that we do not have enough experimental measurements to train a machine learning model on the experimental data alone. Image classification is not the primary task of interest in ICF, so traditional TL using a pretrained open-source model is not appropriate. However, it is possible to create massive databases of ICF simulations for neural network training [5], [22]. The simulations are suspected to be a good reflection of reality but require tuning to be more consistent with experiments. We, therefore, propose the use of TL as a nonparametric approach for calibrating a simulation-based neural network to experiments. The general concept is illustrated in Fig. 1: Train a feed-forward neural network on simulation data to relate simulation inputs to observable outputs. Next, freeze many of the layers of the neural network but retrain a small number of layers of the network using the sparse set of experimental data for which the inputs and output observables are known.

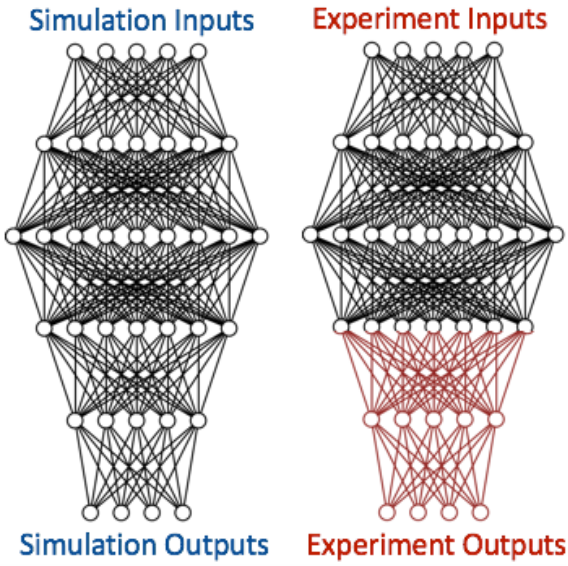

Fig. 1. To transfer learn from simulations to experiments, the first three layers of the simulation-based network are frozen, and the remaining two layers are available for retraining with the experimental data.

The feasibility of using TL to produce neural networks that accurately predict the outcomes of ICF experiments is tested on data generated at the Omega laser facility [23] at the Laboratory for Laser Energetics (LLE). In Section II, we introduce a proposed hierarchical approach to TL for numerical simulation and experimental data. In Section III, we compare standard and hierarchical TL for the Omega data set and use the transfer learned models to study the discrepancies between the Omega simulations and experiments in Section IV.

II. Hierarchical TL

Computer simulations of complex physical systems are often modeled at varying levels of fidelity. Computationally inexpensive, low fidelity models are used to explore vast design spaces for optimal settings, and expensive high fidelity models might be used in interesting regions of design space to compute predictions of planned experiments. The high fidelity simulations are often more accurate and reliable than the approximate models, but the expense of running the simulation often prevents their use in large parameter scans. It might be possible to create models that emulate high fidelity simulations with reduced computational cost with TL. For example, a model trained on a dense set of 1-D simulations could be calibrated to a sparse set of 2-D simulations that fill the same design space. Rather than running a dense set of 2-D simulations to train a 2-D surrogate model, an equally accurate surrogate might be obtained by TL from 1-D to 2-D with a relatively small number of 2-D simulations, saving

substantial computational resources. Furthermore, this 2-D-calibrated model can be subsequently calibrated to experimental data. If the 2-D model is a better reflection of reality than the 1-D model, the TL step between 2-D and the experiment should be easier than jumping from 1-D directly to the experimental data. We refer to this technique of TL from low to subsequently higher fidelity simulations to the experimental data as "hierarchical TL."

To demonstrate the utility of hierarchical TL, consider the simple function

$$f(a, x) = xe^{ax} \tag{1}$$

where x and a are random variables; x between $[-1, 1]$ and a between [0, 1]. This expression will be the "experiment" or true function. We also have a low and high fidelity approximation of the experiment

$$f_{\text{low}}(a, x) = x \tag{2}$$

$$f_{\text{high}}(a, x) = x + ax^2 \tag{3}$$

which are the first- (low fidelity) and second-(high fidelity) order Taylor expansions for the true function, respectively. This problem is used to study the benefits of hierarchical TL; specifically to see if stepping through the hierarchy results in better models than calibrating directly from low fidelity to experimental data.

This article uses fully connected neural networks designed with the algorithm "deep jointly-informed neural networks" (DJINN) [24]. DJINN is chosen for its efficiency in choosing an appropriate neural network architecture for arbitrary data sets, and eliminates the need for extensive hyper-parameter optimization. The algorithm uses a decision tree model trained on the data, and maps the decision tree structure into a neural network architecture with initialized weights. The networks have rectified linear unit activation functions [25], and minimize the L2 loss with weight regularization using the Adam optimizer [26] in Tensorflow [27]. DJINN performs particularly well for inertial confinement fusion data [28], [29], which is the focus of Sections III and IV. For further details on the algorithm, the reader is referred to the publication or publicly available software for DJINN [24].

For this comparison, DJINN models with three hidden layers are trained to map from (x, a) to $f(x, a)$ or one of the approximations in (2) and (3). First, the average explained variance score (averaged over 5 random training/testing data splits of 80%/20%) for DJINN models trained on experiments alone is computed; this is the baseline to which various TL techniques is compared, as the transfer learned models are not expected to exceed the accuracy of a DJINN model trained purely on experimental data. Next, TL from high fidelity simulations to experiments, then from low fidelity simulations to experiments, is tested. Finally, TL from low to high fidelity, and then to experiments is performed to evaluate the benefits of hierarchical TL. To compare the models, the mean and standard deviation of the explained variance score are computed. A Student's t-test [30] is performed to determine if the transfer learned models are statistically significantly different in performance than a model trained exclusively on

TABLE I

HYPER-PARAMETERS FOR ORIGINAL MODEL AND TL FOR THE TAYLOR EXPANSION EXAMPLE

Original Model Parameters	
Number of models	5
Hidden layer widths	4-8-14; 4-7-15, 4-8-14; 4-7-15; 4-7-14
Learning rate	0.004
Batch size	50
Epochs	300
Transfer Learning Parameters	
Retrained layers	Final 2
Learning rate	0.0001
Batch size	1
Epochs	300

TABLE II

COMPARISON OF HIERARCHICAL AND ONE-STEP TL TO DIRECT MODELING OF EXPERIMENTS. LOW/HIGH FI. INDICATES DATA PRODUCED WITH THE LOW/HIGH FIDELITY SIMULATIONS, RESPECTIVELY. EXP. INDICATES "EXPERIMENTAL" DATA PRODUCED WITH THE ANALYTIC EXPRESSION IN (1)

Model	Mean±SD Explained Variance	p-value
Train with 100 exp.	0.994 ± 0.004	-
Train with 100 high fi.; TL with 25 exp.	0.994 ± 0.006	0.957
Train with 100 low fi.; TL with 25 exp.	0.954 ± 0.041	0.016
Train with 100 low fi.; TL with 50 high fi. + TL with 25 exp.	0.981 ± 0.025	0.279

experiments. The explained variance scores and p-values from the Student's t-test are summarized in Table II. The neural network hyper-parameters are noted in Table I and are kept the same for all of the models; these parameters are manually tuned to optimize performance. The choice to retrain the final layers of the model is motivated by applications of TL to image recognition models; finer details are often the focus of the final layers, whereas the first layers in the networks focus on large scale features [12]–[14]. Similarly for regression tasks, modifying the final layers of the network allow for simple modifications to the network's predictions, whereas modifying early layers could lead to substantial nonlinear changes to the overall response surface. This is a design choice that is used throughout the remainder of the manuscript, however, it is not a necessary requirement; the layers chosen for retraining can also be optimized for a specific task.

For this simple example, there is not a statistically significant difference between a model trained exclusively on a large data set of experiments, models that are trained on high fidelity simulations and calibrated to experiments, and models that are hierarchically calibrated. However, these three models are statistically significantly better than the model which is calibrated directly from low fidelity to the experiments, illustrating that there is an advantage of informing the model of high fidelity data prior to experimental calibration. In order to make an accurate emulator of the experiments, the cost of each of these routes should be the determining factor in which method to choose; however, it is expected in most cases that the hierarchical method, which requires the least number

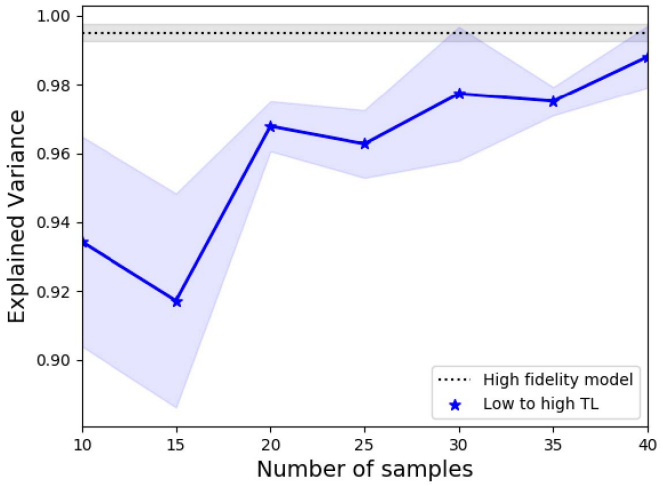

Fig. 2. "High fidelity" prediction quality as the number of high fidelity data points used for TL from low fidelity data is increased. Transfer learned models require approximately 30 high fidelity data points to achieve a model accuracy that is comparable to a model trained exclusively on 100 high fidelity simulations.

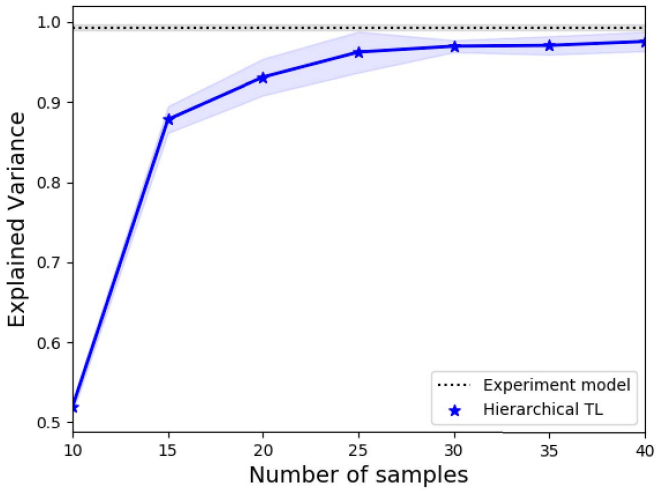

Fig. 3. Experimental prediction accuracy as the number of experimental data points is increased in hierarchical TL. Models are first trained on 100 low fidelity simulations, calibrated to high fidelity with 30 high fidelity data points and then subsequently calibrated to the experiments. The model quality converges with about 25 experiments and is comparable to the performance of a model trained on 100 experiments alone.

of experiments and/or high fidelity simulations, is the least expensive approach. For this simple example, the computational cost is negligible, but for applications in which complex multiphysics systems are being modeled, the computational cost difference between low and high fidelity simulations can be substantial; for such systems, experiments are often costly and limited in number.

In Table II only a single size data set is considered for the hierarchical model. To determine the minimum number of high fidelity simulations and experiments that is needed to create a model that is not significantly different from the baseline (experiment only) surrogate, the mean explained variance score is computed as the data set sizes are varied. First, the minimum number of high-fidelity points that are required to produce a transfer-learned model that is of similar performance to a model trained exclusively on 100 high fidelity simulations is determined. Next, the necessary number of experiments needed to recalibrate this model to be on par with the experiment-only baseline model is determined. As in the models from Table II, the number of points in the data set includes training and testing data, which are split into 80%/20% sets. The TL parameters are held constant as recorded in Table I, and each model starts with the same low fidelity surrogate trained on 80 data points and validated on the remaining 20 points.

Fig. 2 illustrates how the TL quality improves as the number of high fidelity data points is increased; the error bars reflect the variation in performance when the testing/training data sets are shuffled. TL with about 30 high fidelity samples produces models that perform similar to models trained exclusively on high fidelity simulations. Thus, the first step in the hierarchical calibration is performed with 30 high fidelity simulations. Next, the minimum number of experimental data points needed to calibrate this model to the experiments is determined. Fig. 3 illustrates the quality of the transfer learned model predictions as the number of experiments is varied. For this example, performance comparable to the baseline is achieved with 25 experiments; beyond this the improvement in model quality is minimal.

Whether it is beneficial to perform hierarchical TL depends on the relative expense of the varying levels of fidelity in simulations and the experiments. If 100 low fidelity and 30 high fidelity simulations are less expensive than 100 high fidelity simulations, it is worth taking the hierarchical approach for subsequently calibrating to 25 experiments; however, there may be situations in which running a high fidelity database is preferred. This example suggests that hierarchical TL does offer improvements over TL from low fidelity data directly to the experiments. The hierarchical calibration approach could also be improved by optimizing where the high fidelity simulations and experiments are placed in the design space; future work will explore sampling strategies for efficient hierarchical TL.

Previous results are for a very simple, low-dimensional function that is fast to evaluate. In Section III-A, the same technique is applied to a real-world application: ICF experiments from the Omega laser facility.

III. TL for ICF Model Calibration

The performance of hierarchical TL is next tested on a real-world data set containing 23 experiments from the Omega laser facility that lie within a 9D design space. A database of 30k Latin hypercube sampled [31], [32] Lilac [33] simulations span the 9D space encompassing the experiments. The nine input parameters varied in the databases include laser pulse parameters: the average drive, drive rise time, energy on the target, the first picket power, foot power, foot width, and foot picket width, and capsule geometry parameters: the ice thickness and the outer radius of the capsule.

The 30k simulations are of low fidelity; they are 1-D and do not account for laser-plasma interactions (LPIs) such

TABLE III

HYPER-PARAMETERS FOR ORIGINAL MODEL AND TL FOR THE OMEGA DATA SET

Original Model Parameters	
Number of models	5
Hidden layer widths	11-13-22-16; 11-14-19-26; 11-14-24-16; 11-15-29-30; 11-14-22-29
Learning rate	0.004
Batch size	1500
Epochs	400
Transfer Learning Parameters	
Retrained layers	Final 2
Learning rate	0.0003
Batch size	1
Epochs	2300

TABLE IV

MEAN EXPLAINED VARIANCE SCORES ON THE TEST DATA SETS FOR FIVE DJINN MODELS TRAINED ON RANDOM 80% SUBSETS OF THE 30K LOW-FIDELITY LILAC SIMULATIONS

Observable: Mean Explained Variance			
Absorption Fraction	0.991	R0	0.962
Adiabat	0.886	RhoR	0.968
Burnwidth (BW)	0.869	Shock Mass	0.856
BangTime	0.990	RhoMaxBT	0.962
Convergence Inner	0.964	Tion	0.956
Convergence Outer	0.962	Tion_DD	0.959
In-flight aspect ratio (IFAR)	0.885	Vi	0.971
Peak Kinetic Energy	0.981	Yield	0.944
Pressure	0.958	Yield_DD	0.949
RhoNAve	0.967		

as cross-beam energy transfer (CBET) [34]–[36] and use flux-limited thermal diffusion models [37]. Each simulation requires about ten wall-clock minutes to run on a single CPU. Each of the 23 experiments is accompanied by a high fidelity simulation, which is a 1-D Lilac simulation with CBET, more accurate equations of state [38], and non-local transport; these simulations require approximately ten wall-clock hours to run. Both the low and high fidelity simulations produce a large number of scalar outputs; 19 of which are included in the following analysis. There are only five observables available for all 23 experiments that are common to the simulation database that will be used to test TL with experimental data.

The low fidelity simulation-based neural networks are trained using DJINN, these models are referred to as "low fidelity DJINN" models. In Section III-A, TL is used to calibrate from low fidelity to high fidelity simulations, and then to experiments.

A. Hierarchical TL With the Omega Data set

The hierarchical approach, in which the low fidelity model is calibrated first to high fidelity simulation data, then to the experimental data, is compared to TL directly from low fidelity simulations to experiments for the Omega database. The models transfer learned to high fidelity simulations are referred to as "high fidelity DJINN" models, and those that are subsequently transfer learned to experiments as "experiment DJINN" models. If the high fidelity simulations are more accurate depictions of reality than the low fidelity simulations, priming the DJINN model with high fidelity information prior to calibrating to the experiments could improve the ability of the model to adapt to the experimental data.

An ensemble of five low-fidelity DJINN models is trained on the database of 30k low fidelity simulations. The models predict all 19 observables simultaneously and are individually cross-validated. The variance between DJINN models, each of which has been trained on a different random 80% subset of the data and will provide uncertainty estimates on the model predictions. The hyper-parameters used to train the networks are summarized in Table III. The mean explained variance score of the low fidelity models for each output is given in Table IV.

The same ensemble of low fidelity DJINN models are used to compare hierarchical calibration to calibration directly from low fidelity simulations to experiments. The low fidelity DJINN models are calibrated independently, each on a different random subset of the high fidelity or experimental data. For each of the models, the first three layers of weights are frozen, and the remaining two layers are available for retraining, as shown in the cartoon in Fig. 1. Note that the architecture of the networks is not reflected in this cartoon; the true architectures are given in Table III for the ensemble of five DJINN models. The last two layers of weights are retrained to convergence for 2000 epochs with a batch size of one and a learning rate of 0.0003 in the Adam optimizer (these parameters are tuned to optimize predictive performance of the models). Each model is trained on a random 90% of the experimental data (20 points) and tested on the remaining 10% (3 points). The low fidelity and post-shot simulations have 19 outputs, but the experiments only have five available observables. To calibrate to the experiments, the cost function, which is the mean squared error (MSE) of the 19 scaled outputs, is modified such that the missing 14 outputs are not weighted. More explicitly, the cost becomes a weighted MSE where the weights are zero for outputs not measured in the experiment, and unity for those that are observed in the experiment. The predictions for the remaining 14 observables may change in nonphysically motivated ways, thus we focus only on the five available observables. An equivalent approach to the weighted cost function would be to train the low fidelity and high fidelity DJINN models with only the five outputs available in the experiment. We choose to retain all 19 outputs so the accuracy of the high fidelity calibration for all 19 observables can be evaluated. Note that all input and output data are scaled [0,1], using the parameter ranges set by the database of 30k simulations, prior to training. This prevents the cost function from being biased toward outputs that are larger in magnitude due to the choice of units.

First we consider TL from the low fidelity simulations to the high fidelity simulations. Fig. 4 illustrates the prediction error (calculated on training and testing data combined), computed as

$$\text{Error} = \frac{(\text{Prediction}) - (\text{High fidelity truth})}{(\text{High fidelity truth})} \tag{4}$$

for all nineteen available outputs for the uncalibrated (low fidelity) and calibrated (high fidelity) DJINN models. The error bars reflect the standard deviation in prediction errors from the

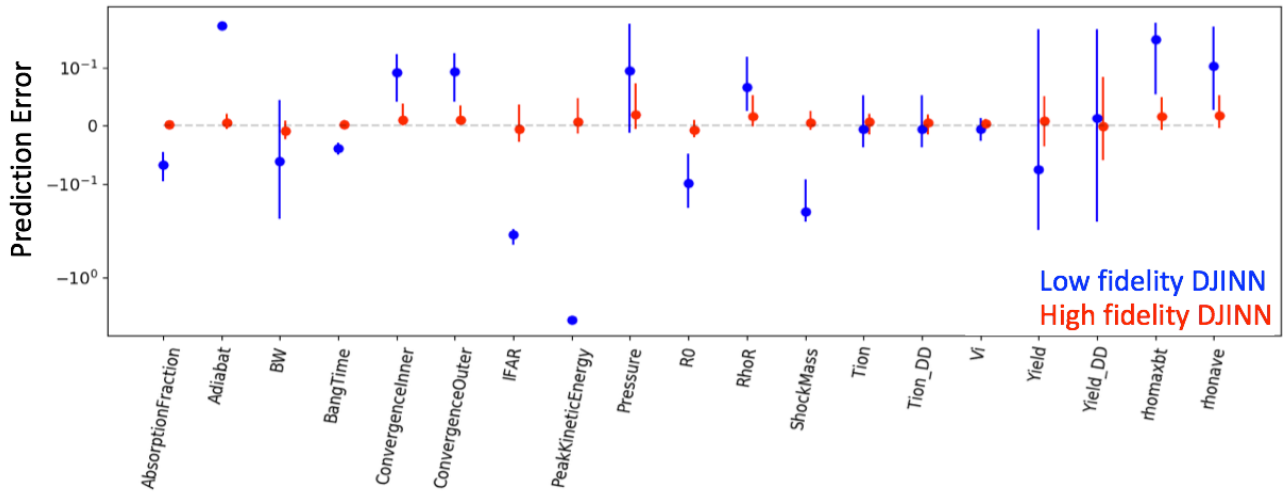

Fig. 4. Prediction error (with the high fidelity simulation as the ground truth) for calibrated and uncalibrated DJINN models. The low fidelity model predicts high fidelity observables with significant error, as the models contain different physics. The models calibrated to high fidelity data are able to predict all 19 high fidelity observables with low error.

ensemble of DJINN models; the points in Fig. 4 illustrate the mean error.

The low fidelity and high fidelity simulations differ significantly in their predictions of the nineteen observables, shown by the error in the blue points of Fig. 4. The low error in the red points indicates that the networks are able to successfully learn the high fidelity outputs via TL.

The high fidelity calibrated models are next calibrated to the experimental data, again by TL the last two layers of the network—the same layers that were modified to calibrate to the high fidelity data. The mean and standard deviation of final prediction error is now computed using the experiment as the ground truth

$$\text{Error} = \frac{(\text{Prediction}) - (\text{Experiment truth})}{(\text{Experiment truth})}. \tag{5}$$

The prediction quality is illustrated in Fig. 5 for the low fidelity, high fidelity, and experiment DJINN models.

Fig. 5 illustrates that the high fidelity simulations are not necessarily more predictive of reality than the low fidelity simulations, however, TL to experiments is still successful. The largest error for TL is in the yield, due to the model needing to adjust its predictions by over an order of magnitude for most experiments, however, the mean prediction error for all observables is close to 0.

Since the high fidelity simulations are not closer to the experiments than the low fidelity models, hierarchical modeling does not offer significant benefits for this data set; comparable results are achieved by TL directly from low fidelity simulations to the experiments. Table V records the average explained variance ratio for each the experimental observables for the hierarchical models, and those that are calibrated directly from low fidelity to experiments. The explained variance ratios are computed on the test data set, and are averaged for the five models.

The previous analyses involve randomly choosing the training and testing data for model calibration. To illustrate how these models can be used to predict the outcomes of future experiments, a model calibrated to the high fidelity simulations is calibrated to the experiments using only the oldest 19 experiments in the data set. The model is tested on the four most recent experiments; the predictions are shown in Fig. 6. Training on the old data and predicting the four most recent experiments requires the model to extrapolate in input space, away from the old experimental data. The model is able to successfully predict the outcomes of the newest four experiments, demonstrating it does have the ability to successfully extrapolate away from the experimental data, but within the bounds of the simulations. It should be noted that the same hyper-parameters given in Table III are used when training on the old data and predicting the most recent experiments. Since these hyper-parameters were chosen by doing random selections of training/testing data, the more recent experiments could have implicitly influenced the choice of the hyper-parameters. A more fair comparison would have held the four most recent experiments from the choice of hyper-parameters as well, although it should be noted that the performance of the models is robust to minor perturbations in the learning rate and number of training epochs and it is unlikely that optimization of these parameters without the four

TABLE V

EXPLAINED VARIANCE SCORES FOR MODELS CALIBRATED FROM LOW FIDELITY TO HIGH FIDELITY SIMULATIONS TO EXPERIMENTAL DATA AND LOW FIDELITY SIMULATIONS TO EXPERIMENTS. THE HIGH FIDELITY SIMULATIONS ARE NOT AN ACCURATE PICTURE OF REALITY, AND THUS THERE ARE NO SIGNIFICANT BENEFITS OF FIRST CALIBRATING TO THE HIGH FIDELITY DATA FOR THIS PARTICULAR DATA SET

	Mean±SD Explained Variance		
Observable	Hierarchical	Low fi. - Exp.	p-value
Burnwidth	0.975 ± 0.023	0.889 ± 0.079	0.139
Bangtime	0.987 ± 0.015	0.942 ± 0.092	0.364
ρR	0.874 ± 0.097	0.835 ± 0.179	0.712
T_{ion}	0.988 ± 0.009	0.924 ± 0.094	0.211
Yield	0.818 ± 0.143	0.956 ± 0.034	0.096

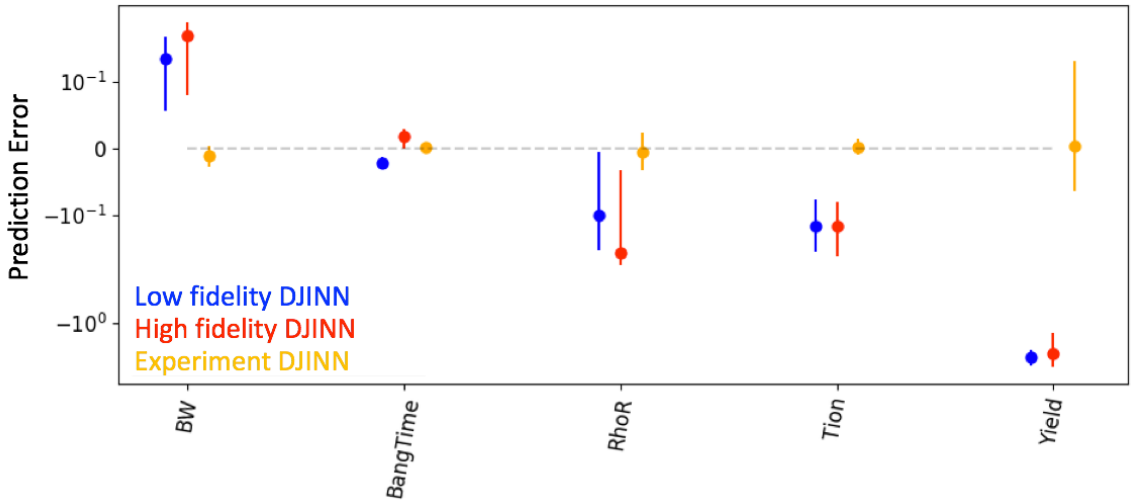

Fig. 5. Prediction error (with the experiment as the ground truth) for low fidelity DJINN models. DJINN models calibrated to the high fidelity data experimental data. The experimentally calibrated models predict five experimental observables with low error.

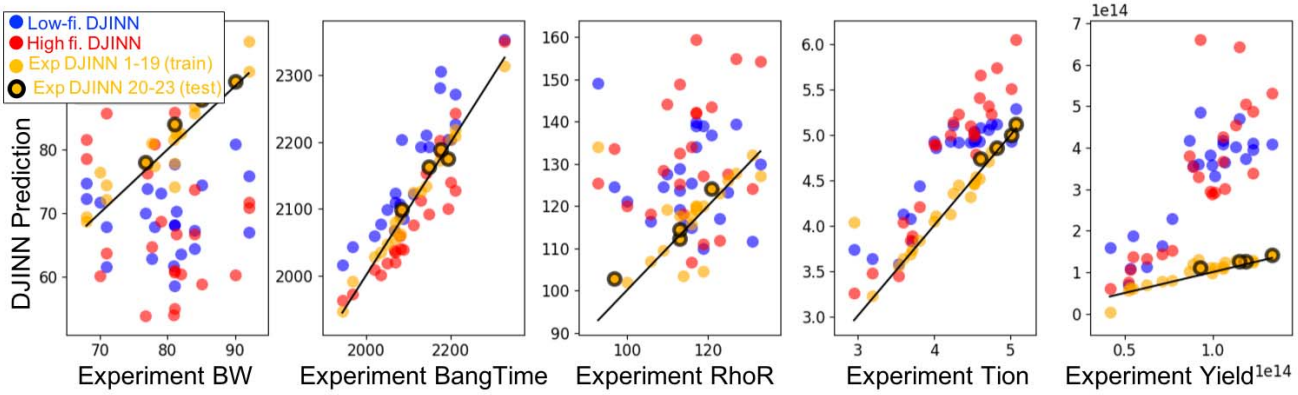

Fig. 6. Predictions of new experiments (yellow circled in bold black) after TL using previous experimental data. The model is able to accurately predict new experiments, which are small extrapolations in input space from the old experimental data. The yellow data points indicate the training data predictions and the points with black outlines are predictions for test data. The blue and red points indicate the low and high fidelity DJINN model predictions of the experiments, respectively.

most recent experiments would have resulted in significantly different model performance.

The low generalization error of the transfer learned models is a result of hyperparameter tuning that prevents over-fitting or "catastrophic forgetting" [39]. Over-fitting is when a machine learning model essentially memorizes the training data and cannot accurately predict data that are not included in the training set. Catastrophic forgetting is a related challenge that occurs in transfer learned models, in which the model is over-trained on the new task, and thus forgets all knowledge of the original task for which it was trained. In the case of the Omega data set, the neural network would "forget" all knowledge of the simulation data and would simply learn how to fit the experimental data, and thus its extrapolation properties would not be informed by the simulation data, as desired. The risk of over-training to the experimental data can be reduced by limiting the learning rate, the number of training epochs, and the number of retrained weights in the TL step. Limiting these hyper-parameters prevents the neural network from changing significantly, and thus it retains memory of the simulation data on which it was originally trained. High cross-validation scores and the ability to extrapolate away from the experimental data used in the training step indicate that the model is unlikely to be over-fitting to the experimental data. Although caution should still be taken when extrapolating away from the experiments, it appears that small extrapolations are tolerated, and the models can be updated after each experiment to enable further exploration away from the current data.

IV. Exploring Discrepancies Between Simulations and Omega Experiments With TL

A result of hierarchical TL is that it produces three sets of models which emulate low fidelity simulations, high fidelity simulations, and experiments. Studying the discrepancies between these models gives insight into the different physics included in the simulations and the experiment.

A primary use of ICF implosion simulations is to find optimal design settings for experiments. An interesting application of the three models is thus to search for "optimal" designs using each fidelity surrogate and determine if the

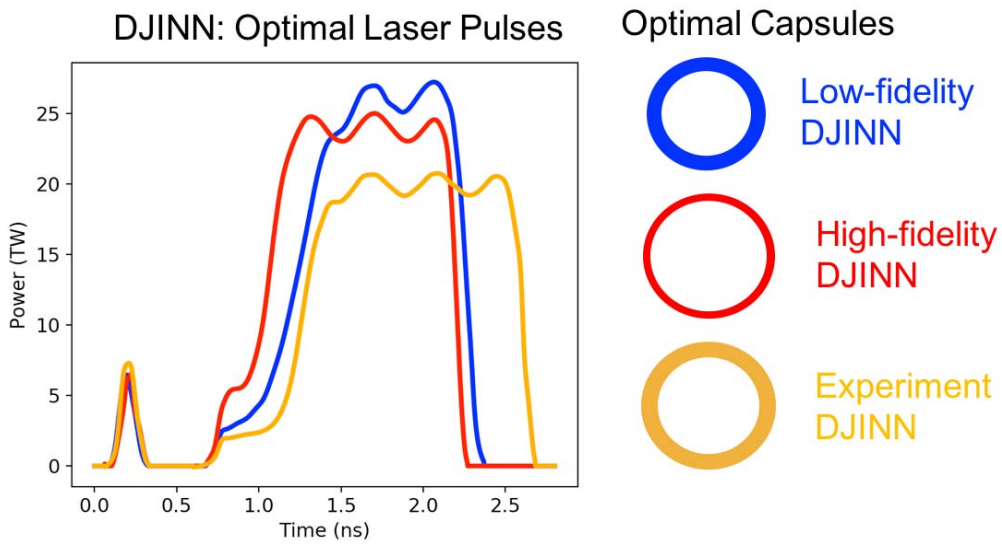

Fig. 7. Designs that optimize ITFX according to the low fidelity, post-shot, and experiment DJINN models. The three designs are distinct due to the lack of accurate physics models, asymmetries, and other experimental sources of performance degradation not included in the simulations.

simulation-based models suggest a similar "optimal" design as the experiment-informed model. For this exercise, we define an optimal design as one that maximizes the experimental ignition threshold factor (ITFX) [40]

$$\text{ITFX} \propto \text{Yield} \cdot (\rho \text{R})^2 \tag{6}$$

where ρR is the areal density. The resulting optimal designs are illustrated in Fig. 7.

There are several important differences between the three optimal designs. First, consider the differences between the low fidelity (blue) and high fidelity (red) designs. The low fidelity design prefers high compression of a thick capsule and achieves this by driving the capsule with a very high power. This differs from the high fidelity design, which prefers a lower power and thinner shell in order to achieve a similar implosion velocity and yield. The high fidelity design includes CBET effects, so it is reasonable for this design to lower the peak power to avoid performance degradation from LPI [41].

Next consider the experimental design: unlike the high fidelity design, this design adjusts the picket and the foot of the pulse to guard against hydrodynamic instabilities that are not modeled in 1-D simulations. Lowering the foot of the pulse (which occurs around 0.75 ns) lowers the adiabat inside the shell, allowing for higher compression and therefore higher areal density. To mitigate instabilities associated with higher compression, the picket of the pulse (occurring at around 0.2 ns) is increased to increase the adiabat on the outer surface of the shell. The higher outer adiabat reduces hydrodynamic instabilities by increasing the ablation velocity [42]–[44]. The experimental capsule is thicker and is driven at an even lower power than the high fidelity design for a longer period of time; perhaps due to underestimation of the CBET effects by the high fidelity simulations.

It is interesting to compare the predictions for each fidelity of DJINN model for the three optimal designs, given in Table VI. The low fidelity model optimal design is expected to perform well according to simulations, but the experiment-calibrated model expects it to perform 24% worse than the experimental optimal. Thus, relying on simulations alone leads to incorrect conclusions about how to best maximize implosion performance.

The maximum ITFX design according to the experiment-calibrated model is consistent with other analyses of this database [45]–[47]. The researchers at Omega are training power law-based models to relate simulation outputs and experimental measurements; through this process they found that to optimize the yield, they should increase the thickness of the ice in the capsule; the experimentally calibrated DJINN model makes a similar suggestions. They confirm their model predictions with a series of experiments, each time making small extrapolations in shell thickness and updating the models with the new experimental data before predicting the outcome of the next experiment. After maximizing the yield, the researchers performed a set of experiments to independently optimize the areal density by modifying the picket and foot of the pulse. This approach is largely physics-guided, and treats the pulse and capsule independently to optimize yield and areal density; it does not explicitly account for interactions between the capsule geometry and the laser pulse. In contrast to the power law fits to the data, the neural network models can consider nonlinear interactions between the inputs and outputs, tuning the pulse and capsule simultaneously to maximize ITFX. However, the neural networks might be inaccurate far from the experimental data, thus caution should be taken to make small extrapolations from the data with this technique as well.

The fact that two distinct methods for creating data-driven models suggest similar design choices for optimizing performance is encouraging and illustrates the powerful role that TL can play in designing ICF experiments. Although a purely machine learning motivated ICF experiment has not yet been carried out, machine learning has been used to optimize performance of other large scale experiments with

TABLE VI
PREDICTIONS OF THE LOW FIDELITY (LOW FI.), HIGH FIDELITY (HIGH FI.), AND EXPERIMENT (EXP.) DJINN MODELS AT THE POINT OF OPTIMAL YIELD $(\rho R)^2$ ACCORDING TO EACH OF THE THREE MODELS

log10(Yield(ρR)^2) at:	Low Fi. Optimal	High Fi. Optimal	Exp. Optimal
Low Fi. DJINN	19.764	19.586	19.570
High Fi. DJINN	19.421	19.598	19.273
Exp. DJINN	18.951	19.019	19.070

promising results [48], [49]. In the future work, we intend to apply machine learning to guide the optimization of ICF experiments, however, we feel it is still necessary to analyze such designs to ensure they do not encounter sources of performance degradation or physical limitations not included in the training data. For example, facility limits or safety requirements might be considered, or one might want to limit how far from previous experiments the model can extrapolate.

V. CONCLUSION

TL with deep jointly informed neural networks has enabled the creation of surrogate models which emulate more expensive simulations and experiments, without requiring massive quantities of expensive data. TL uses low fidelity simulations to learn the approximate responses surfaces, and then uses a sparse collection of expensive high fidelity simulations or experimental data to modify a limited number of weights in the network. The resulting networks accurately emulate the expensive data, without requiring a large database of high fidelity simulations or experiments to train the neural network from scratch. Hierarchical TL, the process of calibrating from low to high fidelity simulations to experiments (or between levels of fidelity of simulations) enables the creation of accurate emulators for the highest fidelity simulations or experiments, with lower computational cost than creating a neural network on the highest fidelity data alone. TL with DJINN enables the creation of neural network models that are predictive of direct drive ICF experiments at the Omega laser facility, which are used to design optimal implosions for future experiments.

ACKNOWLEDGMENT

The authors would like to thank R. Betti, V. Gopalaswamy, and J. P. Knauer at the LLE for providing the Lilac simulations and the Omega experimental databases and candid conversation. They would like to thank B. Kustowski for conversations on transfer learning.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Released as LLNL-CONF-764063. This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security LLC, and shall not be used for advertising or product endorsement purposes.

REFERENCES

[1] S. Atzeni and J. Meyer-ter-Vehn, *The Physics of Inertial Fusion: Beam Plasma Interaction, Hydrodynamics, Hot Dense Matter*. Oxford, U.K.: Oxford Univ. Press, 2009.

[2] J. D. Lindl, *Inertial Confinement Fusion: The Quest for Ignition and Energy Gain Using Indirect Drive*. College Park, MD, USA: AIP-Press, 1998.

[3] M. Marinak *et al.*, "Three-dimensional HYDRA simulations of National Ignition Facility targets," *Phys. Plasmas*, vol. 8, no. 5, pp. 2275–2280, 2001.

[4] R. Nora, J. L. Peterson, B. K. Spears, J. E. Field, and S. Brandon, "Ensemble simulations of inertial confinement fusion implosions," *Stat. Anal. Data Mining*, vol. 10, no. 4, pp. 230–237, 2016.

[5] J. L. Peterson *et al.*, "Zonal flow generation in inertial confinement fusion implosions," *Phys. Plasmas*, vol. 24, no. 3, p. 032702, 2017.

[6] M. Van Oijen, J. Rougier, and R. Smith, "Bayesian calibration of process-based forest models: Bridging the gap between models and data," *Tree Physiol.*, vol. 25, no. 7, pp. 915–927, 2005.

[7] H. F. Stripling, R. G. McClarren, C. C. Kuranz, M. J. Grosskopf, E. Rutter, and B. R. Torralva, "A calibration and data assimilation method using the Bayesian MARS emulator," *Ann. Nucl. Energy*, vol. 52, pp. 103–112, Feb. 2013.

[8] A. T. Booth, R. Choudhary, and D. J. Spiegelhalter, "A hierarchical Bayesian framework for calibrating micro-level models with macro-level data," *J. Building Perform. Simul.*, vol. 6, no. 4, pp. 293–318, 2013.

[9] R. G. McClarren, *Uncertainty Quantification and Predictive Computational Science: A Foundation for Physical Scientists and Engineers*. New York, NY, USA: Springer, 2018.

[10] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models," *J. Roy. Stat. Soc., B (Stat. Methodol.)*, vol. 63, no. 3, pp. 425–464, 2001.

[11] J. A. Gaffney, "Making ICF models more predictive: Combining simulations, experiments and expert knowledge using machine learning and Bayesian statistics," in *Proc. APS Meeting Abstr.*, 2018.

[12] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, Feb. 2016.

[13] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.

[14] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR*, Jun. 2009, pp. 248–255.

[16] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2014.

[17] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2007-001, 2007.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[19] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[20] G. H. Miller, E. I. Moses, and C. R. Wuest, "The national ignition facility: Enabling fusion ignition for the 21st century," *Nucl. Fusion*, vol. 44, no. 12, p. S228, 2004.

[21] L. Kegelmeyer, "Image analysis and machine learning for NIF optics inspection," Lawrence Livermore Nat. Lab., Livermore, CA, USA, Tech. Rep. LLNL-CONF-746729, 2018.

[22] S. H. Langer, B. K. Spears, J. L. Peterson, J. E. Field, R. Nora, and S. Brandon, "A HYDRA UQ workflow for NIF ignition experiments," in *Proc. 2nd Workshop Situ Infrastruct. Enabling Extreme-Scale Anal. Vis. (ISAV)*, Nov. 2016, pp. 1–6.

[23] J. Kelly *et al.*, "OMEGA EP: High-energy petawatt capability for the OMEGA laser facility," *J. de Phys. IV*, vol. 133, pp. 75–80, Jun. 2006.

[24] K. D. Humbird, J. L. Peterson, and R. G. McClarren, "Deep neural network initialization with decision trees," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1286–1295, May 2019.

[25] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[27] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," Mar. 2016, *arXiv:1603.04467*. [Online]. Available: https://arxiv.org/abs/1603.04467

[28] K. D. Humbird, J. L. Peterson, and R. G. McClarren, "Parameter inference with deep jointly informed neural networks," *Stat. Anal. Data Mining, ASA Data Sci. J.*, 2019.

[29] B. K. Spears and *et al.*, "Deep learning: A guide for practitioners in the physical sciences," *Phys. Plasmas*, vol. 25, no. 8, 2018, Art. no. 080901.

[30] M. O'Mahony, *Sensory Evaluation of Food: Statistical Methods and Procedures*. Evanston, IL, USA: Routledge, 2017.

[31] M. D. McKay, R. J. Beckman, and W. J. Conover, "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, vol. 21, no. 2, pp. 239–245, 1979.

[32] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments*. New York, NY, USA: Springer, 2013.

[33] J. Delettrez, R. Epstein, M. C. Richardson, P. A. Jaanimagi, and B. L. Henke, "Effect of laser illumination nonuniformity on the analysis of time-resolved X-ray measurements in uv spherical transport experiments," *Phys. Rev. A, Gen. Phys.*, vol. 36, no. 8, p. 3926, 1987.

[34] I. Igumenshchev *et al.*, "Crossed-beam energy transfer in implosion experiments on OMEGA," *Phys. Plasmas*, vol. 17, no. 12, 2010, Art. no. 122708.

[35] P. Michel *et al.*, "Tuning the implosion symmetry of ICF targets via controlled crossed-beam energy transfer," *Phys. Rev. Lett.*, vol. 102, no. 2, 2009, Art. no. 025004.

[36] D. H. Froula *et al.*, "Mitigation of cross-beam energy transfer: Implication of two-state focal zooming on OMEGA," *Phys. Plasmas*, vol. 20, no. 8, 2013, Art. no. 082704.

[37] C. D. Levermore and G. C. Pomraning, "A flux-limited diffusion theory," *Astrophys. J.*, vol. 248, pp. 321–334, Aug. 1981.

[38] H. Risken, "Fokker-planck equation," in *The Fokker-Planck Equation*. New York, NY, USA: Springer, 1996, pp. 63–95.

[39] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio, "An empirical investigation of catastrophic forgetting in gradient-based neural networks," 2013, *arXiv:1312.6211*. [Online]. Available: https://arxiv.org/abs/1312.6211

[40] P. Y. Chang *et al.*, "Generalized measurable ignition criterion for inertial confinement fusion," *Phys. Rev. Lett.*, vol. 104, no. 13, 2010, Art. no. 135002.

[41] B. J. MacGowan *et al.*, "Laser–plasma interactions in ignition-scale hohlraum plasmas," *Phys. Plasmas*, vol. 3, no. 5, pp. 2029–2040, 1996.

[42] V. N. Goncharov *et al.*, "Improved performance of direct-drive inertial confinement fusion target designs with adiabat shaping using an intensity picket," *Phys. Plasmas*, vol. 10, no. 5, pp. 1906–1918, 2003.

[43] J. P. Knauer *et al.*, "Improved target stability using picket pulses to increase and shape the ablator adiabat," *Phys. Plasmas*, vol. 12, no. 5, 2005, Art. no. 056306.

[44] K. Anderson and R. Betti, "Laser-induced adiabat shaping by relaxation in inertial fusion implosions," *Phys. Plasmas*, vol. 11, no. 1, pp. 5–8, 2004.

[45] V. Gopalaswamy *et al.*, "Tripled yield in direct-drive laser fusion through statistical modelling," *Nature*, vol. 565, no. 7741, p. 581, 2019.

[46] V. Gopalaswamy, "Optimization of direct-drive inertial fusion implosions through predictive statistical modeling," in *Proc. Amer. Phys. Soc. Division Plasma Phys. Meeting*, 2018.

[47] R. Betti, "The one-dimensional cryogenic implosion campaign on OMEGA: Modeling, experiments, and a statistical approach to predict and understand direct-drive implosions," in *Proc. Amer. Phys. Soc. Division Plasma Phys. Meeting*, 2017.

[48] E. A. Baltz *et al.*, "Achievement of sustained net plasma heating in a fusion experiment with the optometrist algorithm," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 6425.

[49] M. J. V. Streeter *et al.*, "Temporal feedback control of high-intensity laser pulses to optimize ultrafast heating of atomic clusters," *Appl. Phys. Lett.*, vol. 112, no. 24, Art. no. 244101, 2018.