Data-driven methods for diffusivity prediction in nuclear fuels

Galen T. Craven,[1,*] Renai Chen,[1] Michael W. D. Cooper,[2] Christopher Matthews,[2] Jason Rizk,[2] Walter Malone,[3] Landon Johnson,[1] Tammie Gibson,[1] and David A. Andersson[2]

[1] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA*
[2] *Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA*
[3] *Department of Physics, Tuskegee University, Tuskegee, Alabama, USA*

The growth rate of structural defects in nuclear fuels under irradiation is intrinsically related to the diffusion rates of the defects in the fuel lattice. The generation and growth of atomistic structural defects can significantly alter the performance characteristics of the fuel. This alteration of functionality must be accurately captured to qualify a nuclear fuel for use in reactors. Predicting the diffusion coefficients of defects and how they impact macroscale properties such as swelling, gas release, and creep is therefore of significant importance in both the design of new nuclear fuels and the assessment of current fuel types. In this article, we apply data-driven methods focusing on machine learning (ML) to determine various diffusion properties of two nuclear fuels—uranium oxide and uranium nitride. We show that using ML can increase, often significantly, the accuracy of predicting diffusivity in nuclear fuels in comparison to current analytical models. We also illustrate how ML can be used to quickly develop fuel models with parameter dependencies that are more complex and robust than what is currently available in the literature. These results suggest there is potential for ML to accelerate the design, qualification, and implementation of nuclear fuels.

I. INTRODUCTION

Atomistic structural defects influence and alter the macroscopic properties of nuclear fuels and materials [1]. Macroscopic changes, such as volumetric swelling, gas release, and creep, can in turn give rise to alterations of the functionality of the fuel in a reactor. Therefore, developing theoretical methods to predict the growth rates of atomistic point defects and defect clusters is of significant importance in the design and qualification of fuels and materials that are used in reactors. The growth rates of defect clusters are governed by the diffusivity of their constituent point defects. This is because the rate at which defects move through the fuel lattice strongly influences how quickly those defects combine to form larger defect clusters. Predicting the diffusion properties of defects in nuclear fuels under reactor conditions is, therefore, an important research focus in reactor design and safety and surety analysis.

There are two primary theoretical approaches that are applied to determine diffusion properties in nuclear fuels: (1) deriving empirically-motivated analytical functional forms and fitting those forms to existing experimental data and (2) extracting diffusion coefficients from atomic scale calculations and simulations combined with rate theory approaches. One simulation method that is commonly applied to understand and predict defect growth in irradiated materials is cluster dynamics [2–8]. Cluster dynamics is a mean-field method that tracks the time evolution of concentrations of point defects and defect clusters [9]. Diffusivity predictions are generated from cluster dynamics simulation data by combining the predicted defect concentrations with mobility data. Empirical analytical models provide ease of use, transferablilty, interpretability, and are computationally simply to evaluate. However, they typically have a limited range of applicability with respect to variation of reactor conditions and often do not capture the salient features of diffusion processes at a quantitative level. Atomistic calculations combined with rate models for defect evolution, often under nonequilibrium conditions [10, 11], can be used to provide high-level predictions of diffusion coefficients [9, 12, 13] in irradiated materials. However the process to construct, parameterize, calibrate, and test atomistic and cluster dynamics models can be time-consuming.

In this article, we develop a data-driven workflow to predict diffusion properties of nuclear fuels. This workflow focuses on the application of machine learning (ML) methods to predict diffusion coefficients of various chemical species, defect types, and defect clusters. Machine learning is a broad term that typically defines a set of numerical methods that are applied to construct unknown model functions using existing data to train the ML process [14, 15]. ML methods have had a history of success in the fields of chemistry, physics, and material science [16–22]. In these fields, ML is often performed by training numerical models through the application of a specific learning algorithm to data from high-level electronic structure calculations or experimental data. Because ML uses existing data, it represents a powerful compliment to existing methods and not a replacement. ML methods can greatly reduce the amount of experimental data or simulation data needed to address a problem, allowing the prediction of properties of a material without running a computationally-expensive molecular simulation or performing an experiment. ML also allows for the development of models that capture more complex parameter dependencies than traditional models.

The nuclear engineering/physics community has started to adopt ML methodologies in the study of nu-

* galen.craven@gmail.com

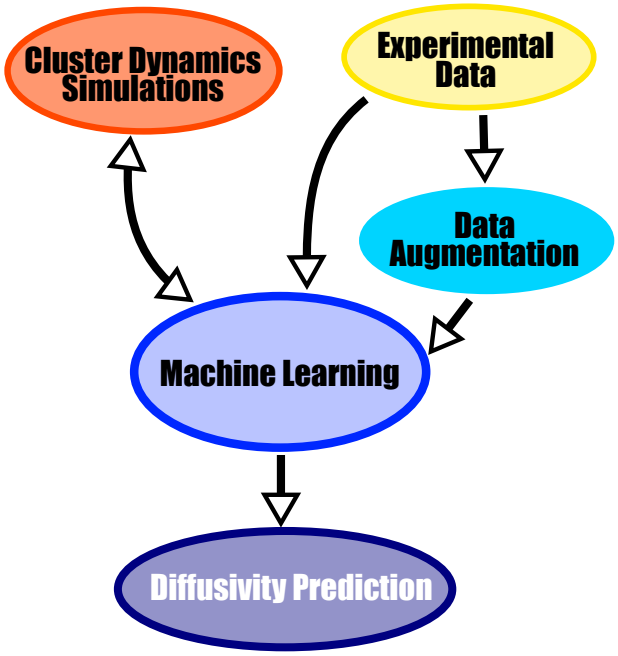

FIG. 1. Schematic representation of the data-driven workflow developed in this work.

clear reactors and the nuclear fuel cycle [23–28]. Historically, these are research areas in which both experimental and simulation data is scarce and laborious to generate. Some noteworthy examples of applications of ML in nuclear engineering include: the prediction of properties of light water reactors [24], nuclear criticality safety analysis [26], and thermal conductivity prediction [27]. Given the previous success of ML in chemistry, physics, material science—and now nuclear engineering— it is promising to expand the utility of ML techniques within the nuclear fuels discipline.

The two specific nuclear fuels examined in this work are uranium oxide UO_2 and uranium nitride UN—the former is an established nuclear fuel and the latter has been not as widely used, although it is a promising candidate for a variety of advanced reactors due to its high thermal conductivity and U density. [29–31]. The diffusion datasets used to train ML processes are obtained from three sources: experimental results taken from the literature, atomistically-informed cluster dynamics simulations, and data augmentation methods that are applied to expand small experimental datasets. There are a number of recent studies on the diffusion properties of UO_2 [9, 12, 32–34]. The data in these studies, and other experimental work [35–38], can be used to validate and calibrate the developed models. There is less available information and data for UN in comparison to UO_2. There are, however, several small datasets available for self-diffusion and fission gas diffusion in UN [29–31, 39–44]. We examine these fuels under irradiated and non-irritated conditions.

The primary goal in this paper is to demonstrate the application of ML to predict diffusivity behavior in nuclear fuels. The workflow for this project is shown in Fig. 1. Experimental diffusion data is used to train ML processes directly and is also fed into data augmentation algorithms to expand small datasets. The experimental data and augmented data are coupled with data generated using molecular simulations, specifically cluster dynamics simulations. The baseline parameters in the cluster dynamics model, informed by atomic scale calculations and simulations, are generated based on DFT and empirical potentials [45]. Those baseline parameters are then calibrated self-consistently by creating a feedback loop between the ML program, the molecular simulations and experimental data used for calibration. Note the range within which the parameters are allowed to change is based on the inherent uncertainty of the atomic scale simulations. The target of this process is to generate diffusivity data from cluster dynamics simulations that agree with experimental results. The overall output of this workflow is a set of calibrated diffusivity models built by merging data from simulations and experiments.

The remainder of this article is organized as follows: Section II contains an introduction to the methods used to examine and predict diffusion coefficients in nuclear fuels. An overview of the available experimental data is also given. In Section III, the results of the ML diffusion models are presented. Section III A focuses on fission gas diffusion and self-diffusion in uranium oxide and Section III B focuses on various diffusion properties of uranium nitride. Conclusions and future directions are discussed in Section IV.

II. METHODOLOGIES AND DATA

A. Machine Learning

ML methods are typically applied to construct unknown model functions using existing data to train the ML process [15]. The general objective in *supervised* ML, which is the method used here, is to take a collection of input data (sometimes called features) and corresponding output data (sometimes called labels) and develop a function f that accurately maps the input data to the output data. The term *supervised* ML means that for every set of features $\vec{x}$ there is a corresponding label y. Described in mathematical terms, given a set of k features and a labelled dataset of size n, the general goal of ML is to take the input data $\boldsymbol{X} = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n\}$ where $\vec{x}_1 = \{x_1^{(1)}, x_1^{(2)}, \ldots, x_1^{(k)}\}$ and corresponding output data $\boldsymbol{y} = \{y_1, y_2, \ldots, y_n\}$ and develop a function f that accurately maps the input data to the output data:

$$f : \mathbb{R}^k \rightarrow \mathbb{R}. \tag{1}$$

In the context of diffusion models developed here, the feature vector $\vec{x}$ will consist of properties such as tem-

perature, partial pressure, and fission rate and the output will be a diffusion coefficient for a specific defect or chemical species. The goal of training a ML process is to develop a mathematical model using available data so that when new data is used as input the model generates an accurate output. In most ML applications, the developed function f takes a highly complex functional form that does not correspond to a simple and sparse analytical expression.

ML is commonly performed by separating available data into two subsets—training data and testing data. The training data is used to build the model, i.e., to train the ML model, and the testing data is used to quantify the predictive accuracy of the model. The testing data is not used in the model construction and is only used to test and quantify the model. A typical split between training and testing data is 80:20 meaning that eighty percent of the data is used for training and twenty percent for testing. An 80:20 training/testing ratio is used in all the ML models presented in this work.

There are multiple ML methods that can be used to construct the function f. Two well-known approaches are kernel methods and neural networks [14]. Neural networks are data-hungry methods that perform best when the size of the dataset used for training is large [46]. In the context of diffusion in nuclear fuels, the available datasets are typically small and therefore kernel methods are expected to perform better for these cases. The specific ML package we use to perform the ML procedures is `Scikit-learn` [47]. In this work, we illustrate data-driven ML methods to determine diffusion of various defects and species in uranium oxide and uranium nitride. The data used to train the ML processes comes from two primary sources: experimental datasets extracted from the literature and molecular simulations.

B. Experimental Data

The experimental diffusion datasets used as training data for the ML processes are taken from literature sources. Table I is a list of the datasets. In most cases, the experimental data was given in tabular form in the original papers. If the numerical values of the data were not listed in a table in the original papers, and were only shown in graphical figures, the data was digitally extracted from the figures using the `WebPlotDigitizer` program. It is important to note that the diffusion data in Table I was collected over several decades and that the experimental techniques used to collect the data are varied, therefore the quality and accuracy of the datasets also vary.

C. Cluster Dynamics Simulations

Molecular simulation methods can be applied to generate diffusivity data for nuclear fuels [9, 12, 34]. Here, cluster dynamics simulations are used to generate diffusion coefficients for various defect types in the respective fuel. Implementing the cluster dynamics method consists of parameterizing and then solving a typically large system of nonlinear coupled ordinary differential equations, where each equation in the system describes the time evolution of a specific defect type.

The cluster dynamics code `CENTIPEDE` [9, 12] was applied to incorporate physical parameters for the nuclear fuels and to solve the defect evolution equations. Details about the `CENTIPEDE` code and the physics behind it can be found in Ref. 9. In a `CENTIPEDE` simulation, the concentration c_d of every defect type d is tracked in time through a differential equation of the form

$$\begin{aligned}\frac{dc_d}{dt} = \dot{\beta}_d + \sum_{d'} \dot{R}_{d,d'}(c_d, c_{d'}, D_d, D_{d'}, T, G) \\ - \sum_{s} \dot{S}_{d,s}(c_d, c_s, D_d, T, G),\end{aligned} \tag{2}$$

where $\dot{\beta}_d$ is the generation rate of defect d due to irradiation, $\dot{R}_{d,d'}$ is the reaction rate between defect types d and d', and $\dot{S}_{d,s}$ is the sink rate between defect type d and sink type s. The sums in Eq. (2) are taken over all defect types (self-inclusive) and all sink types. The reaction and sink rates depend on the free energy of the system G and temperature of the system T. The reaction rate between defect types d and d' also depends on the concentrations of each defect and the diffusion coefficients D_d and $D_{d'}$ of those defects. We are primarily interested in solving for the steady-state concentrations with constant source and sink strengths, which are found when the rate of change of the concentration vanishes ($\frac{dc_d}{dt} = 0$) up to some numerical precision for all defect types (See Refs [9, 12, 45] for more details on `CENTIPEDE` implementations of UO_2). Once the concentration of point defects and clusters have been determined, self-diffusion and Xe diffusion can be obtained as the sum over the product of the relative concentration of each defect contributing to diffusion of a species and its mobility.

III. RESULTS

A. Uranium Dioxide

The generation and subsequent diffusion of fission gas in UO_2 impacts fission gas release and swelling, which, in turn, impact fuel performance [12]. Turnbull *et al.* have reported measurements for Xe diffusion coefficients in UO_2 under irradiation over the approximate temperature range $500\text{K} - 1700\text{K}$ [35]. Fig. 2 shows the Xe diffusion coefficient in UO_2 as a function of temperature—the red circular markers represent the Turnbull data. Analytical models have been previously developed to predict the Xe diffusion coefficient at various temperatures and under various irradiation conditions. Two analytical models are the Matthews model [12] and the Forsberg

TABLE I. List of experimental diffusivity datasets used in this work

Dataset	Fuel	Species	T(K)	p(atm)	# of data points	Ref.
Turnbull *et al.*	UO_2	Xe	$\approx 500-1700$	p_{O_2}= NA	34	35
Miekeley and Felix	UO_2	Xe	$\approx 1200-2000$	p_{O_2}= NA	32	36
Davies and Long	UO_2	Xe	fit only	p_{O_2}= NA	fit only	37
Matzke	UN	N	$\approx 1500-2300$	p_{N_2}= NA	36	39, 40
Holt and Almassy	UN	N	$\approx 2050-2300$	$p_{N_2} \approx 0.01-0.8$	12	41
Sturiale and DeCrescente[48]	UN	N	$\approx 1800-2400$	$p_{N_2} \approx 0.13$	27	42
Reimann *et al.*	UN	U	$\approx 1875-2150$	$p_{N_2} \approx 10^{-4}-0.6$	30	43

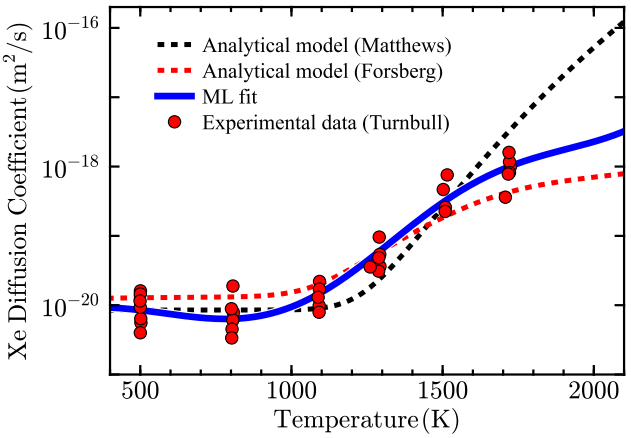

FIG. 2. Xe diffusion coefficient in UO_2 as a function of temperature. The red points are experimental results from Turnbull *et al.* [35]. The solid blue curve is the ML result. The dashed red and dashed black curves are, respectively, results of the analytical models of Matthews *et al.* [12] and Forsberg and Massih [49].

model [49], which is based on the original development by Turnbull *et al.* [35]. The details of these models are found in the Appendix. In Fig. 2, the Matthews model is shown as a dashed black curve and the Forsberg model is shown as a dashed red curve. Both analytical models are in agreement with the Turnbull data over most temperatures, but the Matthews model overestimates and the Forsberg model underestimates Xe diffusion at high temperatures. The parameters in the Matthews model are obtained by fitting to the results of cluster dynamics simulations performed using the CENTIPEDE code, not by directly fitting to the Turnbull data. Therefore, some discrepancy between the data and the model is expected.

As proof of concept that ML can be used to improve the predictive accuracy of existing analytical models, we used the Turnbull data to train a ML process to predict the Xe diffusion coefficient. The specific ML method used was kernel ridge regression (KRR), a method which combines ridge regression with the so-called kernel trick[50]. We implemented KRR using the `Scikit-learn` [47] software package. Ridge regression is a method for approximating the coefficients of several multiple-regression models which works well if the independent variables are highly correlated [51, 52]. `Scikit-learn` specifically utilizes ridge regression with linear least squares with L_2-norm regularization. In some ML methods and applications, raw data must be transformed via a feature map into a feature vector representation. Kernel methods, through the use of kernel functions, can circumvent the need to directly calculate feature vectors, which can be computationally expensive. These methods achieve this by calculating the inner products between each image pair within the feature space [53].

The predictor function (the unknown function to be approximated) in KRR can be expressed as:

$$f(\vec{x}) = \sum_i \alpha_i \mathcal{K}(\vec{x}, \vec{x}_i), \tag{3}$$

where $\mathcal{K}(\vec{x}, \vec{x}_i)$ is the so-called kernel which can be chosen to take various functional forms depending on the properties of the data being analyzed; a polynomial kernel and radial basis function kernel [54] are used in this work. The elements α_i in Eq. 3 are taken from the matrix,

$$\boldsymbol{\alpha} = (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{y}, \tag{4}$$

where $\boldsymbol{K}$ is the kernel matrix with elements given by $K_{i,j} = \mathcal{K}(\vec{x}_i, \vec{x}_j)$, $\boldsymbol{I}$ is the identity matrix, λ denotes a regularization parameter in the ridge regression method that puts a penalty on the weights in order to reduce the variance of predictions, and $\boldsymbol{y}$ represents the target matrix.

The ML result for Xe diffusivity generated by training a KRR process is shown as a blue curve in Fig. 2. The ML result is in strong agreement with the experimental results over the entire temperature range of the Turnbull data. The dip in the ML model at $\approx$ 800K is a result of variance in the data used to train the ML model, not a change in the diffusion mechanism. This proof of concept example illustrates the power of ML to quickly generate models that accurately capture important trends in diffusion data. When a ML model is trained using the Turnbull data, there is large variance in the model depending on which points are randomly assigned as testing data and training data. To mitigate this variance, we developed a data augmentation approach to artificially expand the Turnbull dataset, and, therefore, reduce variance in the ML model. The results presented in Fig. 2 use this data augmentation method, which is described in detail below.

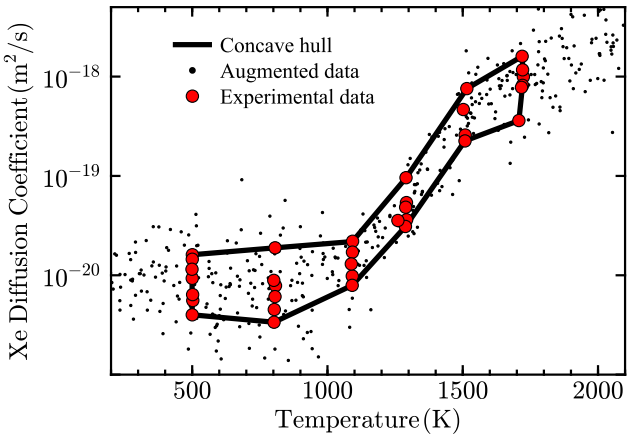

FIG. 3. The concave hull of the Turnbull *et al.* data is shown in solid black. The black dots are augmented data and the red points are the Turnbull experimental data.

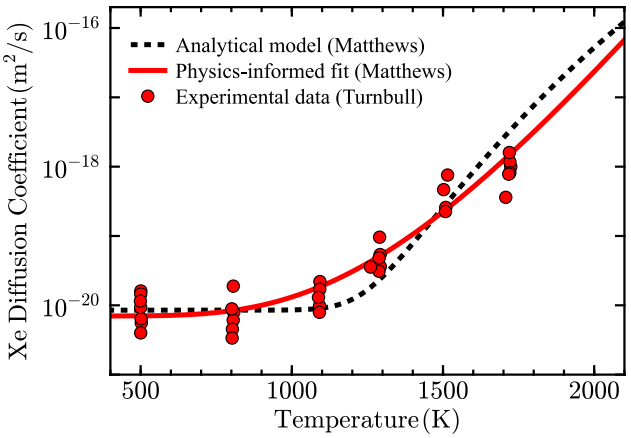

FIG. 4. Xe diffusion coefficient in UO_2. The red points are experimental results. The dashed black curve is the analytical model of Matthews *et al.* [12] and solid red curve is Matthews model parameterized using augmented data. Note that the Matthews model is not fit directly to the Tunbull data, so, here, the primary conclusion is the utility of PIML in the context of fitting to new data sources.

Large datasets are typically used to train ML models (e.g., neural network models). This highlights the traditional connection between big data and machine learning. However, in context of diffusion properties of nuclear fuels, there is often limited experimental data available. In the situations in which the available data is limited, performing ML can result in models with large variances and uncertainties. To mitigate these problems, a data augmentation technique has been developed to artificially expand the small available datasets. This approach shares similarities with density estimation methods [55] but is more specifically tailored to solve the problem of fission gas diffusivity in nuclear fuels. The core of this approach is to use the geometric size of the experimental dataset used as training data to estimate the variance of the data and the density of the experimental data to estimate the mean of the data, and then to generate augmented data by randomly sampling new datapoints from a distribution that uses the estimated variance and mean.

The first step in the augmentation process is to construct the concave hull (a boundary over the dataset) of the data. The concave hull of the Turnbull data is shown in Fig. 3. The Turnbull data consists of diffusivity data as a function of temperature—the concave hull of the data $\partial\mathcal{H}$ bounds the data. At a specific temperature T, augmented datapoints are generated by sampling from a normal distribution $\mathcal{N}(\mu(T), \mathcal{W}(T))$ where $\mu(T)$ is the distribution mean which is extracted from the hull density and $\mathcal{W}(T)$ is the width of the hull. The hull width is defined by drawing a vertical line at temperature T, and noting that the line intersects the hull boundary twice, once at a higher diffusion value $\partial\mathcal{H}_{\mathrm{H}}(T)$ and once at a lower diffusion value $\partial\mathcal{H}_{\mathrm{L}}(T)$. The width is $\mathcal{W}(T) = \partial\mathcal{H}_{\mathrm{H}}(T) - \partial\mathcal{H}_{\mathrm{L}}(T)$. The mean is constructed $\mu(T)$ from a spline interpolation across the data.

To generate an augmented datapoint, a random temperature is sampled from a uniform distribution over the desired temperature range, here $300\mathrm{K} - 2100\mathrm{K}$. At the sampled temperature, a sample is drawn from the normal distribution $\mathcal{N}(\mu(T), \mathcal{W}(T))$. That sample is an augmented diffusion value at the specified temperature. We then iterate over this process until the desired number of augmented datapoints are generated. After performing data augmentation and training the ML model on the augmented data, the variance in the model development is reduced. The black dots in Fig. 3 are augmented data. When the variance in the Turnbull data is small, the variance in the augmented data is also small. Similarly, when the variance in the data is large, the variance in the augmented data is also large. This illustrates how developed data augmentation method captures trends in the variance of the training data.

The kernel-based ML approach can be applied to generate accurate models for diffusion data, but it does not explicitly encode any fundamental physics in the fitting process. Explicitly encoding physics into ML models can be advantageous for situations in which the ML model is used to make predictions outside of the boundaries of the training data, i.e., when the ML is used for extrapolation. One way that physics principles can be encoded into an ML model is to use Physics-Informed Machine Leaning (PIML) methods [56]. Fig. 4 shows the result of a PIML model that is developed by reparameterizing the Matthews analytical model using augmented data generated from the Turnbull dataset. The PIML agrees well with the experimental data across all temperatures. One advantage of the PIML method over the kernel-based ML model is that it captures both the high and low temperature trends outside of the boundaries of the experimental data, similar to the Matthews analytical model [12].

The accuracy of ML models can be compared with the analytical models to quantify any improvement in predic-

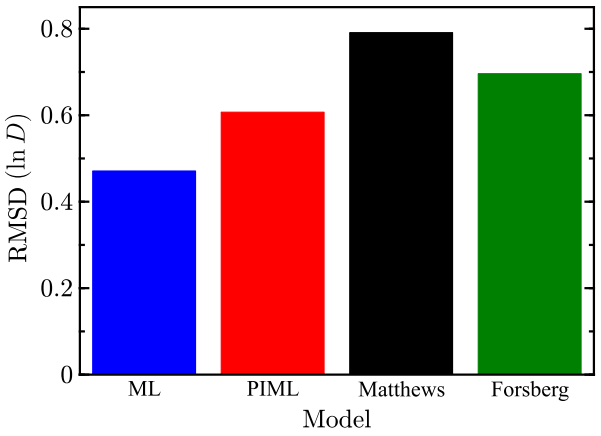

FIG. 5. Root mean square deviation of various models for the Xe diffusion coefficient in UO_2. The experimental data used to calculate the RMSD is from Turnbull *et al.* [35].

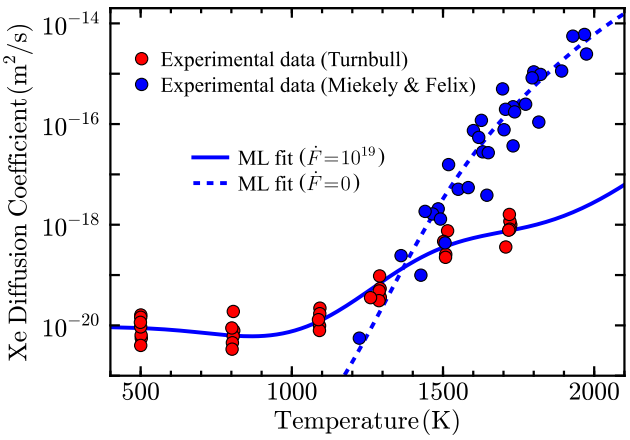

FIG. 6. Xe diffusion coefficient in UO_2 as a function of temperature T and fission rate $\dot{F}$ for the ML model. The solid curve is the result of the ML model for the irradiated case ($\dot{F} = 10^{19}$ fissions/m^3 s) and the dashed curve is the result for the nonirradiated case ($\dot{F} = 0$). The red points are experimental results from Turnbull *et al.* [35] and the blue points are experimental results from Miekeley and Felix [36].

tive accuracy that can be gained using data-driven methods. The root mean square deviation (RMSD) for the four methods/models described previously (taken with respect to the Turnbull data) is shown in Fig. 5. The ML result gives the lowest error and the PIML gives the second lowest error. Both data-driven approaches increase the predictive accuracy for Xe diffusion in comparison to analytical models. Compared to the Matthews model and the Forsberg model, the kernel-based ML approach reduces the error by approximate factors 1.7 and 1.5, respectively. The PIML method reduces the error by a factor 1.3 in comparison to the Matthews model and a factor 1.2 in comparison to the Forsberg model. Note that all the models including the analytical models perform well for Xe diffusion in UO_2. This can be observed qualitatively in Figs. 2 and 4. Therefore, while the data-driven models do provide improvements in predictive accuracy for well-studied fuels such as UO_2, we anticipate the primary use of these methods will be when modeling lesser-studied fuels.

1. *Multi-Dimensional Diffusion Models*

ML can also be applied to construct multidimensional diffusion models of the form $D(T, \dot{F})$ which capture the dependence of Xe diffusion on the fission rate $\dot{F}$ in the material in addition to capturing trends in the temperature dependence. Fig. 6 shows the results of a KRR ML process in comparison to the irradiated Turnbull data and the nonirradiated ($\dot{F} = 0$) data of Miekeley and Felix [36]. The ML model is trained on augmented data generated from these datasets. The ML prediction is in strong agreement with the experimental datasets over all examined temperatures and accurately captures the irradiation behavior. This illustrates the ability of ML to accurately capture trends in multidimensional diffusion data. We have confirmed that the ML model smoothly interpolates between the fission rate of the Turnbull data ($\dot{F} = 10^{19}$ fissions/m^3 s) and the nonirradiated ($\dot{F} = 0$) limit. However, because experimental data has not to our knowledge been generated at intermediate values between these limits, the ML model may not scale accurately in the intermediate regime due to lack of training data. Calibration of the ML model at intermediate fission rates could be accomplished using data generated from cluster dynamics calculations, for example, by using `CENTIPEDE`.

Other experimental datasets for Xe diffusion in UO_2 besides the Miekeley and Felix data can be found in the literature. For example, the analytical fit of Davies and Long is generally considered to more accurately capture Xe diffusivity at thermal equilibrium. Note that the primary goal in this work is to illustrate the utility of ML in nuclear fuel model development, not to assess the accuracy and validity of the datasets that are available in the literature. So, in the context of this work, the datasets are primarily tools to benchmark the developed ML methods. Shown in Fig. 7(a) is a comparison between the results of Matthews model (which was developed to agree with `CENTIPEDE` predictions that are close but not identical to the Davies and Long and Turnbull data sets), the Turnbull data, and the data generated using the analytical fit of Davies and Long. The Matthews model is in agreement with both datasets, but overestimates the diffusivity of the nonirradiated data. For comparison, the ML result shown in Fig. 7 (b) is in excellent agreement with both data sets over all irradiation and temperature conditions. Note that the different experimental data sets used in this subsection have different partial oxygen pressures as well as different irradiation conditions.

Using the results of `CENTIPEDE` cluster dynamics simu-

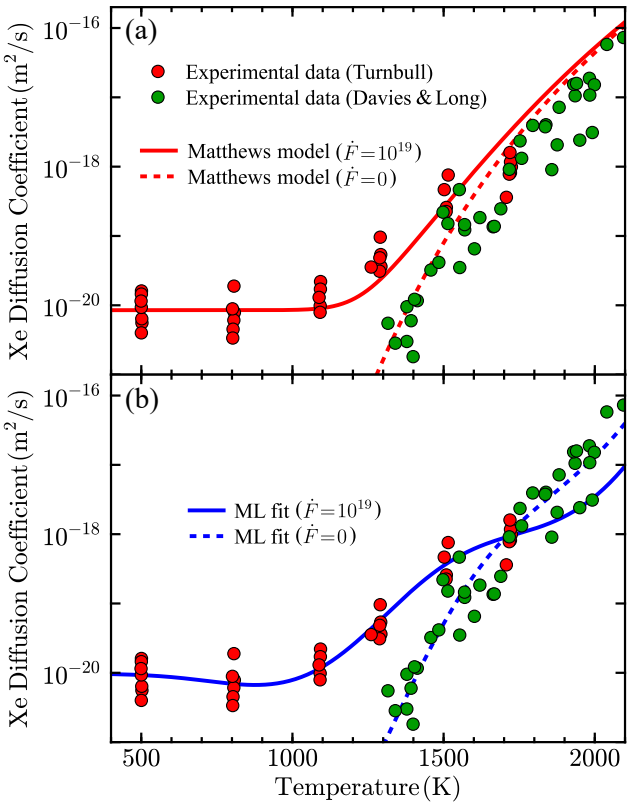

FIG. 7. Xe diffusion coefficient in UO_2 as a function of temperature T and fission rate $\dot{F}$ for (a) the Matthews model (b) a ML model. The solid curve in each panel is result of the respective model for the irradiated case ($\dot{F} = 10^{19}$ fissions/m^3 s) and the dashed curve is the result for the nonirradiated case ($\dot{F} = 0$). The red points are experimental results from Turnbull *et al.* [35] and the green points are experimental results from Davies and Long [37] generated through random sampling inside the error bounds of the fit.

lations, multi-dimensional diffusion models $D(T, \dot{F}, p_{O_2})$ can be developed using ML that include the dependence on the partial pressure of oxygen p_{O_2} as well as the fission rate and temperature. The training data for the ML process was generated by performing `CENTIPEDE` simulations at 1000 datapoints on a grid in the $\{T, \dot{F}, p_{O_2}\}$ parameter space. An 80:20 split between training and testing data was used. All features and labels were log-scaled before training and testing except the temperature. To predict the Xe diffusion and U diffusion, we again utilized KRR [50] with the addition of the nearest neighbors (NN) approach [57, 58]. The procedure is performed by taking an input state point $p = \{T, \dot{F}, p_{O_2}\}$ (the point where the diffusion coefficient would like to be predicted) and determining the $N_{\text{neigh}} = 8$ nearest neighbor points in the training data to the input point. The metric used to determine the nearest neighbors to the input point p was the weighted Euclidean distance with weights obtained using a grid search hyperparameter optimization. After the nearest neighbors are determined, the KRR was performed using only the NN points. The output of the KRR procedure at the target datapoint p in state space is the predicted diffusion value.

Fig. 8(a) and (b) are plots of predicted values vs. true values for Xe diffusion and U diffusion, respectively. The average percent error of the testing data was approximately 10% for both U and Xe diffusion, illustrating excellent agreement between the ML model and the testing data. The different color points in each plot signify different temperatures. The data spans the temperature range 1200K to 2250K, and excellent agreement is observed between the ML model and the testing data over the entire temperature range. The data spans the fission rate range 10^{17} fissions/m^3 s to 10^{19} fissions/m^3 s.

We also used a similar ML procedure to develop a model for the partial pressure of oxygen $p_{O_2}(T, \dot{F}, D)$ that takes diffusivity as an input in addition to fission rate and the temperature. The result of this procedure is shown in Fig. 8(c). Excellent agreement is observed between the predicted values and the true values for the partial pressure. Experimental values for diffusivity are commonly reported at a specific temperature and fission rate. Because the developed partial pressure model takes typically reported quantities as inputs, it allows the determination of the thermodynamic state of a fuel or experiment (qualitatively described using the partial pressure, which controls the x in $UO_{2\pm x}$ linking to the defect concentration).

B. Uranium Nitride

The diffusion properties of UN are less studied and less understood than the diffusion properties of UO_2. This provides an opportunity to illustrate the utility and predictive power of data-driven methods in the context of developing diffusion models for emerging nuclear fuel candidates. Training data for diffusivity prediction in UN can be taken from existing experimental datasets, analytical models, new data generated using cluster dynamics simulations, or a combination of these data sources. Given that the experimental and analytical data are limited and uncertain, the use of cluster dynamics informed by atomic scale simulations is key to providing input for the ML models. A list of the UN experimental diffusivity datasets we use is shown in Table I. The analytical model we compare to our ML model is taken from a collection of analytical models for diffusivity in UN developed by Hayes [59]. The details of this model are found in the Appendix.

Fig. 9 is a plot of true vs. predicted values for U diffusion in UN calculated using three different models: the Hayes analytical model, a PIML model, and a kernel-based ML model. The true values are taken from the experimental measurements of Reimann *et al.* [43] which give U diffusion coefficients as a function of temperature T and the partial pressure of nitrogen p_{N_2}. All of the models we apply are multidimensional diffusion models

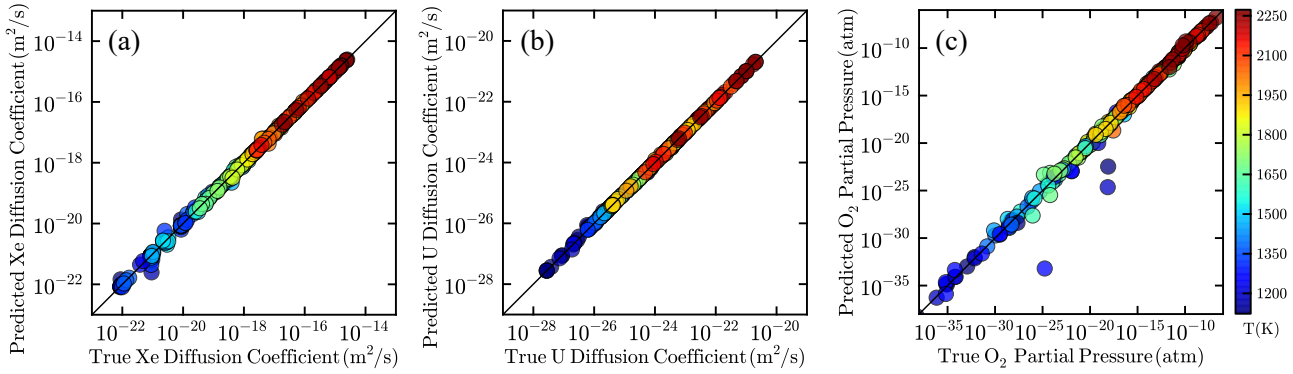

FIG. 8. Machine learning results for UO_2 showing predicted vs. true values for (a) Xe diffusion, (b) U diffusion, and (c) the partial pressure of O_2. The data used to train the models is obtained from *CENTIPEDE* cluster dynamics simulations. The diagonal line in each panel illustrates where the predicted value equals the true value. Different color markers correspond to different temperatures shown in the colorbar to the right.

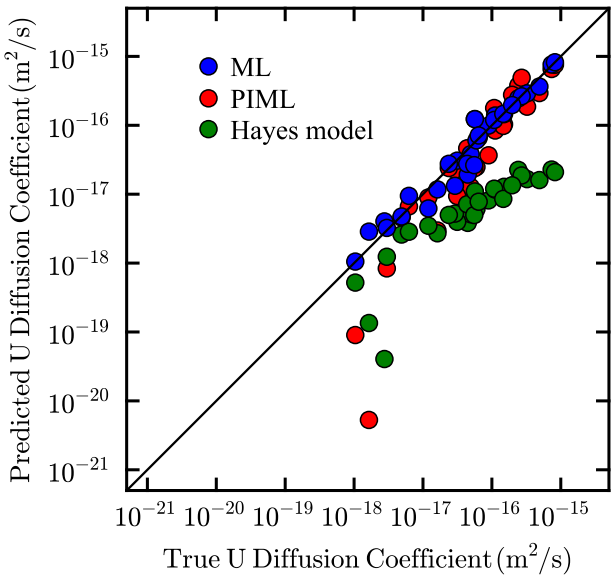

FIG. 9. Predicted vs. true values for uranium diffusion in UN. Results are shown for a ML model (blue), a PIML model (red), and the Hayes analytical model (green). The data used as true values are from Reimann *et al.* [43].

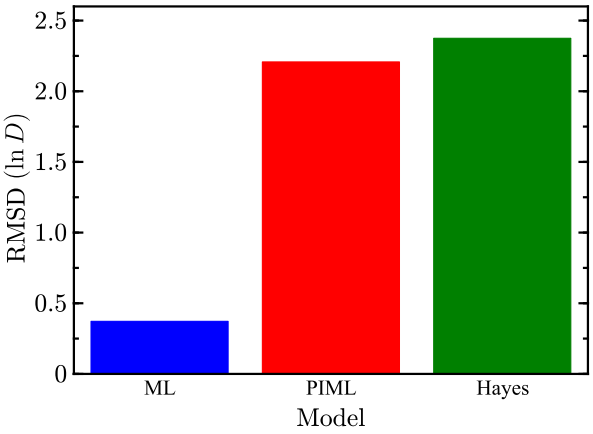

FIG. 10. Root mean square deviation of various models for the uranium diffusion coefficient in UN. The experimental data used to calculate the RMSD is from Reimann *et al.* [43].

of the general form $D(T, p_{N_2})$ that take temperature and pressure as inputs and return a predicted U diffusion coefficient as an output. The Hayes analytical model systematically underestimates the U diffusion values and also generates several points with significant error. A PIML method, developed by reparameterizing the Hayes model, improves on the Hayes fit and performs well for most datapoints, but there are some points (for lower coefficients end particularly) in which a significant difference exists between the true and predicted results because the reparameterization does not change the oversimplified construction of the empirical Hayes model. The ML results give the best results in terms of accuracy and variance.

Shown in Fig. 10 is a comparison between the error values for each method taken with respect to the experimental data from Reimann [43]. The error is quantified using the RMSD. The kernel-based ML method reduces the error by a factor of approximately 6 in comparison to both the PIML method and Hayes analytical model [59]. This significant error reduction highlights the utility of ML methods for quickly developing accurate diffusion models.

1. *Sensitivity Analysis of UN Diffusion*

There is a limited amount of experimental diffusion data available for UN. Therefore, mechanistic cluster dynamics models can (and in general must) be used to augment the experimental data in order to make accurate predictions about diffusivity. In cluster dynamics models, understanding which defect types contribute

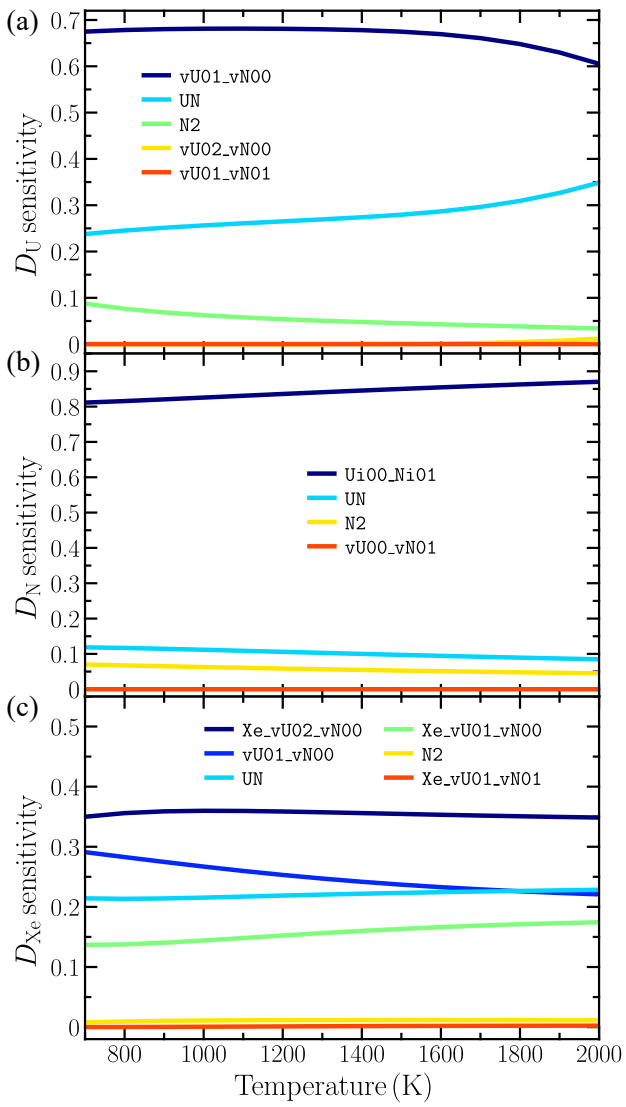

FIG. 11. Sensitivity indices for various defect types in UN. Results are shown for (a) U diffusion, (b) N diffusion, and (c) Xe diffusion.

the most to the overall diffusion mechanism of a species can be quantified and understood using Sensitivity Analysis (SA)—a collection of mathematical and statistical methodologies that are used to understand what inputs contribute the most to a model output. Sensitivity Analysis is a powerful mathematical tool for building and testing the complex computational models that play a significant role in almost all social and physical scientific disciplines. Some example uses of SA methods in the context of model construction are: (a) identifying the most influential inputs to a model output, (b) improving the understanding of the relations between the inputs and output of a model, (c) calibrating input errors, and (d) assessing the quality and confidence of a model [60].

In this article, we use Global Sensitivity Analysis (GSA) methods to investigate the impact of different parameters in the cluster dynamics models for diffusivity. Some of the advantages of GSA methods are that they are capable of handling and analyzing high-dimensional inputs in a computationally efficient manner and that they can be built to taken into account nonlinear effects in the model [61]. Specifically, we use the well-known Morris method [62, 63] to quantify what defects and parameters are most important for diffusivity in UN. The result of applying the Morris method to a model is a collection of sensitivity indices, one index for each input. The value of each index quantifies the relative importance of the corresponding input. A larger value for a sensitivity index implies a higher importance in the model output. The Morris method is computationally more efficient than other SA methods because it scales linearly with the number input parameters. Therefore, it is well-suited for analyzing cluster dynamics simulations of UN which involve a high (> 50) number of input variables. All the numerical algorithms used to perform SA are taken from the `SALib` package [64] and implemented using PYTHON scripts.

In our SA results, a sensitivity index is assigned to every defect type that is tracked in the cluster dynamics simulation. The value of each index quantifies the importance of the corresponding defect to the examined diffusion process. SA was performed for N, U, and Xe diffusivity in UN. The notation `vU`x`_vN`y denotes a vacancy cluster of x uranium vacancies and y nitrogen vacancies. A similar notation is used for interstitials. UN represents perfect UN without defects and N_2 nitrogen gas. Crystal `Xe_vU`x`_vN`y denotes that a Xe atom resides in the `vU`x`_vN`y cluster. The cluster dynamics model applied here differs from the model used in Ref. 45 because we do not include antisites.

Shown in Fig. 11 are the SA results for UN. The uranium vacancy `vU01_vN00` dominates the sensitivity for temperatures below 2000K as shown in Fig. 11(a). Interestingly, U diffusivity is not strongly dependent on the uranium interstitial `Ui01_Ni00`, however, we have found that in specific temperature, pressure, and irradiation regimes it is the dominant defect. The sensitivity indices for D_N and D_{Xe} are shown respectively in Figs. 11(b) and 11(c). For N diffusivity, the nitrogen interstitial `Ui00_Ni01` defect is almost 10 times more important than all the other defects across the temperature range sampled. This is consistent with interstitials dominating nitrogen diffusion under nitrogen rich conditions. Compare this to the results for Xe diffusion shown in Fig. 11(c) which show that a number of defects contribute over 10% to the model output. All of those defects are linked to Xe diffusing by a vacancy mechanism, which is also the mechanism observed to dominate for most temperatures. UN corresponds to the perfect lattice, which impacts all defect energies and consequently appears in all sensitivities. It is possible that it represents a constant shift in all defect energies. The sensitivity indices for U and N diffusivity do not vary strongly as the temperature is varied. Across all the examined models, only

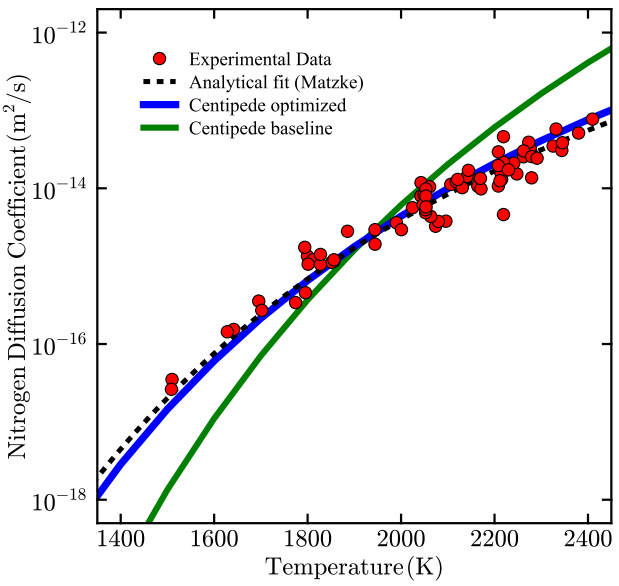

FIG. 12. Nitrogen diffusion coefficient in UN as a function of temperature. The red points are experimental results (see Table I). The dashed black curve is the analytical fit by Matzke. The green curve is the baseline result from CENTIPEDE and the blue curve is the CENTIPEDE result after performing the genetic optimization procedure.

a few defects make major contributions to the overall diffusivity in the diffusion of the three species (U,N,Xe) we have examined.

2. *Optimization of* CENTIPEDE *Parameters*

The parameters in the UN cluster dynamics model implemented using CENTIPEDE can be calibrated such that the quantities of interest agree with experimental results. A detailed discussion of the specific physical parameters that are used in our model can be found in Ref. 45. However, the process of calibrating the approximately 50-dimensional parameter space to experimental data by hand can be very time-consuming if not intractable. To accelerate the calibration process, we used a genetic optimization procedure to find the optimal set of parameters that produce CENTIPEDE results that best match the experimental data for N diffusion. As a proof of principle, we perform the optimization on N diffusion only, and do not include U and Xe diffusion values in the optimization. A multi-objective optimization that includes diffusion for all of the species is a target of future work.

The genetic optimization procedure was implemented using a baseline set of CENTIPEDE parameter values obtained from a combination of electronic structure calculations, and molecular dynamics simulations. An error bound for each parameter was also assigned. Each set of parameter values is a solution to the optimization problem. An initial set of solutions was generated by randomly sampling a value for each parameter using the assigned error bounds. This initial set of solutions is called a generation. Genetic optimization of the baseline parameters was performed by generating new sets of solutions (i.e., by generating new generations) using the best solutions from the previous generation. Each generation had 60 solutions and the optimization was performed for 5 generations. The procedure was performed across the temperature range $1300\text{K} - 2500\text{K}$.

The results of the genetic optimization procedure are shown in Fig. 12. The diffusion values generated using the CENTIPEDE parameters chosen as the baseline do not agree with the experimental results. However, after performing optimization in the high-dimensional parameter space, the CENTIPEDE result is in strong agreement with the experimental data. The power of this result is that while the analytical result of Matzke (shown as dashed black curve) accurately captures the temperature dependence of N diffusion in the equilibrium regime, it cannot be used predict the diffusion behavior under different irradiation conditions and at different partial pressures. Compare this to the calibrated CENTIPEDE result which can be used to predict diffusion coefficients at different system conditions. The calibrated cluster dynamics model is therefore more robust than simple analytical models for modeling diffusion under reactor conditions. Overall, the defect parameters that change the most during the optimization procedure are kinetic parameters such as attempt frequencies and migration activation energies, while the enthalpic and entropic properties related to defect formation, in general, exhibit the least difference between the calibrated and baseline sets. Moreover, the Centipede model predicts diffusion of not only N, which is relatively easy to measure experimentally, but also U and Xe, which are much harder to measure experimentally, especially under irradiation conditions. By validating and optimizing the CENTIPEDE model for N diffusion data, the trust in the CENTIPEDE predictions of U and Xe diffusion is also increased. Future work will investigate how these relations can be quantified using uncertainty measures.

IV. CONCLUSIONS

A data-driven workflow has been developed to predict diffusion coefficients in nuclear fuels. The developed workflow has been shown to predict transport and thermodynamic properties of the nuclear fuels uranium oxide UO_2 and uranium nitride UN with increased accuracy in comparison to previous models and methods. We have specifically shown that using ML can reduce the predictive error in comparison to previously developed analytical models and to reduce the time it takes to develop diffusion models. We have also shown how data-driven methods can be used to calibrate complex mechanistic models for diffusion properties of nuclear fuels. Machine learning models trained using small experimen-

tal datasets were expanded by developing and applying a data augmentation method. This augmentation technique may be particularly useful in nuclear fuel development and qualification when large amounts of experimental results are not available or are difficult to obtain. Sensitivity analysis methods have been applied to determine the most important structural defects within the mechanistic cluster dynamics models under different reactor conditions. This analysis can be used to improve model development and fuel analysis by giving information about what defect types should be targeted for further study using experiments and/or electronic structure calculations.

In future work, data-driven methods will be applied to enhance the predictive capabilities of mechanistic models for use in nuclear fuel qualification and reactor modeling. Data-driven methods will also be used to generate complex multi-dimensional analytical functions with enhanced transferability and interpretability in comparison to black-box ML models developed here. Another important future focus will be to quantify the uncertainty in the developed ML models.

V. ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy through the Los Alamos National Laboratory. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy. This research was supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project number 20220053DR. The computing resources used to perform this research were provided by the LANL Institutional Computing Program.

Appendix A: Analytical Diffusion Models

Hayes Model:

The Hayes analytical model for U diffusion in UN in units of m^2/s is:

$$D_{\mathrm{U}}(T,P) = c_1 p^{c_2} e^{c_3/T} \tag{A1}$$

with $c_1 = 2.215 \times 10^{-15}$, $c_2 = 0.6414$, and $c_3 = -7989.3$, where p is the partial pressure of nitrogen (atm) and T is the temperature (K).

Matthews Model:

The Matthews analytical model for Xe diffusion in UO_2 in units of m^2/s is:

$$D_{\mathrm{Xe}}(T,\dot{F}) = \frac{c_1 e^{c_2/k_{\mathrm{B}}T}}{c_3 + c_4 e^{c_5/k_{\mathrm{B}}T}} + c_6 e^{c_7/k_{\mathrm{B}}T}\sqrt{\dot{F}} + c_8\dot{F} \tag{A2}$$

with $c_1 = 2.216 \times 10^{-7}$, $c_2 = -3.26$, $c_3 = 1.0$, $c_4 = 29.03$, $c_5 = -1.84 \times 10^{-4}$, $c_6 = 2.821 \times 10^{-22}$, $c_7 = -2.0$, and $c_8 = 8.5 \times 10^{-40}$ where $\dot{F} = 1.0 \times 10^{19}$ is the fission rate (fissions/m^3 s), T is the temperature (K), and k_{B} is the Boltzmann constant (eV/K).

Forsberg Model:

The Forsberg analytical model for Xe diffusion in UO_2 in units of m^2/s is:

$$D_{\mathrm{Xe}}(T,\dot{F}) = \frac{v_g(T,\dot{F}) D_{\mathrm{eff}}(T,\dot{F})}{v_g(T,\dot{F}) + g(T,\dot{F})} \tag{A3}$$

with

$$D_{\mathrm{eff}}(T,\dot{F}) = c_1 e^{c_2/T} + 4c_3 e^{c_4/T}\sqrt{\dot{F}} + 4c_5\dot{F}, \tag{A4}$$

$$v_g(T,\dot{F}) = c_6 \pi l \dot{F}(c_7 e^{c_8 T} + \delta)^2, \tag{A5}$$

$$g(T,\dot{F}) = 4\pi c_7 e^{c_8 T}(c_9/T - c_{10}) D_{\mathrm{eff}}(T,\dot{F}), \tag{A6}$$

where $\dot{F} = 1.72 \times 10^{19}$ is the fission rate (fissions/m^3 s) and T is the temperature (K). The parameters are $c_1 = 7.6 \times 10^{-10}$, $c_2 = -35247$, $c_3 = 1.41 \times 10^{-25}$, $c_4 = -13800$, $c_5 = 2.0 \times 10^{-40}$, $c_6 = 3.03$, $l = 6.0 \times 10^{-6}$, $c_7 = 1.453 \times 10^{-10}$, $c_8 = 1.023 \times 10^{-3}$, $\delta = 1.0 \times 10^{-9}$, $c_9 = 1.52 \times 10^{27}$, $c_{10} = 3.3 \times 10^{23}$.

[1] G. S. Was, *Fundamentals of Radiation Materials Science: Metals and Alloys* (Springer, 2016).
[2] M. F. Wehner and W. G. Wolfer, Philos. Mag. A **52**, 189 (1985), doi:10.1080/01418618508237618.
[3] S. I. Golubov, A. M. Ovcharenko, A. V. Barashev, and B. N. Singh, Philos. Mag. A **81**, 643 (2001), doi:10.1080/01418610108212164.
[4] C. J. Ortiz and M. J. Caturla, Phys. Rev. B **75**, 184101 (2007), doi:10.1103/PhysRevB.75.184101.
[5] M. P. Surh, J. B. Sturgeon, and W. G. Wolfer, J. Nucl. Mater. **378**, 86 (2008), doi:10.1016/j.jnucmat.2008.05.009.
[6] B. D. Wirth, X. Hu, A. Kohnert, and D. Xu, J. Mater. Res. **30**, 1440–1455 (2015), doi:10.1557/jmr.2015.25.
[7] J. A. Stewart, A. A. Kohnert, L. Capolungo, and R. Dingreville, Comput. Mater. Sci. **148**, 272 (2018), doi:10.1016/j.commatsci.2018.02.048.
[8] A. A. Kohnert, B. D. Wirth, and L. Capolungo, Comput. Mater. Sci. **149**, 442 (2018), doi:10.1016/j.commatsci.2018.02.049.
[9] C. Matthews, R. Perriot, M.W.D. Cooper, C. R. Stanek, and D. A. Andersson, J. Nucl. Mater. **527**, 151787 (2019), doi:10.1016/j.jnucmat.2019.151787.
[10] G. T. Craven and R. Hernandez, Phys. Rev. Lett. **115**,

148301 (2015), doi:10.1103/PhysRevLett.115.148301.
[11] G. T. Craven and A. Nitzan, Proc. Natl. Acad. Sci. **113**, 9421 (2016), doi:10.1073/pnas.1609141113.
[12] C. Matthews, R. Perriot, M.W.D. Cooper, C. R. Stanek, and D. A. Andersson, J. Nucl. Mater. **540**, 152326 (2020), doi:10.1016/j.jnucmat.2020.152326.
[13] X.-Y. Zhou, J.-H. Zhu, and H.-H. Wu, Int. J. Hydrog. Energy **46**, 5842 (2021), doi:10.1016/j.ijhydene.2020.11.131.
[14] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning* (Springer, New York, 2001).
[15] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112 (Springer, 2013).
[16] M. Kulichenko, J. S. Smith, B. Nebgen, Y. W. Li, N. Fedik, A. I. Boldyrev, N. Lubbers, K. Barros, and S. Tretiak, J. Phys. Chem. Lett. **12**, 6227 (2021), doi:10.1021/acs.jpclett.1c01357.
[17] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, Nature **559**, 547 (2018), doi:10.1038/s41586-018-0337-2.
[18] J. Carrasquilla and R. G. Melko, Nature Phys. **13**, 431 (2017), doi:10.1038/nphys4035.
[19] G. Carleo and M. Troyer, Science **355**, 602 (2017), doi:10.1126/science.aag2302.
[20] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, Nature **549**, 195 (2017), doi:10.1038/nature23474.
[21] D.-L. Deng, X. Li, and S. Das Sarma, Phys. Rev. X **7**, 021021 (2017), doi:10.1103/PhysRevX.7.021021.
[22] Y. Liu, W. Hong, and B. Cao, Energy **188**, 116091 (2019), doi:10.1016/j.energy.2019.116091.
[23] D. Morgan, G. Pilania, A. Couet, B. P. Uberuaga, C. Sun, and J. Li, Curr. Opin. Solid State Mater. Sci. **26**, 100975 (2022), doi:10.1016/j.cossms.2021.100975.
[24] B. Ebiwonjumi, A. Cherezov, S. Dzianisau, and D. Lee, Nucl. Eng. Technol. (2021), doi:10.1016/j.net.2021.05.037.
[25] E. J. Kautz, A. R. Hagen, J. M. Johns, and D. E. Burkes, Compl. Matls. Sci. **161**, 107 (2019), doi:10.1016/j.commatsci.2019.01.044.
[26] P. Grechanuka, M. E. Risingb, and T. S. Palmera, J. Comput. Theor. Transp. **47**, 552–565 (2019), doi:10.1080/23324309.2019.1585877.
[27] M. G. Fernandez, A. Tokuhiro, K. Welter, and Q. Wu, Nucl. Eng. Des. **324**, 27 (2017), doi:/10.1016/j.nucengdes.2017.08.020.
[28] L. Cai, F. Xu, F. G. Di Lemma, J. J. Giglio, M. T. Benson, D. J. Murray, C. A. Adkins, J. J. Kane, M. Xian, L. Capriotti, et al., Materials Characterization **184**, 111657 (2022), doi:10.1016/j.matchar.2021.111657.
[29] R. Matthews, K. Chidester, C. Hoth, R. Mason, and R. Petty, J. Nucl. Mater. **151**, 345 (1988), doi:10.1016/0022-3115(88)90029-3.
[30] K. S. Chaudri, W. Tian, Y. Su, H. Zhao, D. Zhu, G. Su, and S. Qiu, Prog. Nucl. Energy **63**, 57 (2013), doi:10.1016/j.pnucene.2012.11.001.
[31] J. K. Watkins, A. Gonzales, A. R. Wagner, E. S. Sooby, and B. J. Jaques, J. Nucl. Mater. **553**, 153048 (2021), doi:10.1016/j.jnucmat.2021.153048.
[32] M.W.D. Cooper, G. Pastore, Y. Che, C. Matthews, A. Forslund, C. R. Stanek, K. Shirvan, T. Tverberg, K. A. Gamble, B. Mays, et al., Journal of Nuclear Materials **545**, 152590 (2021), doi:10.1016/j.jnucmat.2020.152590.
[33] J. Rest, M.W.D. Cooper, J. Spino, J. Turnbull, P. Van Uffelen, and C. Walker, J. Nucl. Mater. **513**, 310 (2019), doi:10.1016/j.jnucmat.2018.08.019.
[34] R. Perriot, C. Matthews, M.W.D. Cooper, B. P. Uberuaga, C. R. Stanek, and D. A. Andersson, J. Nucl. Mater. **520**, 96 (2019), doi:10.1016/j.jnucmat.2019.03.050.
[35] J. Turnbull, C. Friskney, J. Findlay, F. Johnson, and A. Walter, J. Nucl. Mater. **107**, 168 (1982), doi:10.1016/0022-3115(82)90419-6.
[36] W. Miekeley and F. Felix, J. Nucl. Mater. **42**, 297 (1972), doi:10.1016/0022-3115(72)90080-3.
[37] D. Davies and G. Long, Tech. Rep., United Kingdom Atomic Energy Authority (UK) (1963), www.osti.gov/biblio/4660737.
[38] A. Sabioni, W. Ferraz, and F. Millot, J. Nucl. Mater. **257**, 180 (1998), doi:10.1016/S0022-3115(98)00482-6.
[39] H. Matzke, in *Studies in Inorganic Chemistry*, edited by Øivind Johannesen and A. G. Andersen (Elsevier, 1989), vol. 9, pp. 353–384, doi:10.1016/B978-0-444-88534-0.50018-7.
[40] H. Matzke, J. Chem. Soc., Faraday Trans. **86**, 1243 (1990), doi:10.1039/FT9908601243.
[41] J. B. Holt and M. Y. Almassy, J. Am. Ceram. Soc. **52**, 631 (1969), doi:10.1111/j.1151-2916.1969.tb16064.x.
[42] M. DeCrescente, M. Freed, and S. Caplow, Tech. Rep., Pratt and Whitney Aircraft (USA) (1965).
[43] D. Reimann, D. Kroegbr, and T. Lundy, J. Nucl. Mater. **38**, 191 (1971), ISSN 0022-3115, doi:10.1016/0022-3115(71)90042-0.
[44] J. B. Melehan and J. E. Gates, Tech. Rep., Battelle Memorial Institute (USA) (1964), www.osti.gov/biblio/4674918.
[45] M.W.D. Cooper, J. Rizk, C. Matthews, V. Kocevski, G. T. Craven, T. Gibson, and D. A. Andersson, J. Nucl. Mater. (2023), doi:10.1016/j.jnucmat.2023.154685.
[46] A. T. Mohan and D. V. Gaitonde, arXiv **1804.09269** (2018), doi:10.48550/ARXIV.1804.09269.
[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., J. Mach. Learn. Res. **12**, 2825 (2011), http://jmlr.org/papers/v12/pedregosa11a.html.
[48] The Sturiale and DeCrescente (S&D) diffusion values are sometimes noted as being too high due to the specific experimental technique that was used to obtain them. However, a change in temperature scales in the reported data from Celsius to Kelvin makes these values well aligned with the diffusion values measured by other sources. We therefore change the temperature scale in the S&D data here.
[49] K. Forsberg and A. R. Massih, Model. Simul. Mater. Sci. Eng. **15**, 335 (2007), doi:10.1088/0965-0393/15/3/011.
[50] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning* (MIT press, 2018).
[51] A. E. Hoerl and R. W. Kennard, Technometrics **12**, 55 (1970), www.jstor.org/stable/1267351.
[52] A. E. Hoerl and R. W. Kennard, Technometrics **12**, 69 (1970), www.jstor.org/stable/1267352.
[53] K. Koutroumbas and S. Theodoridis, *Pattern Recognition* (Academic Press, 2008).
[54] K. P. Murphy, *Machine Learning: A Probabilistic Perspective* (MIT press, 2012).
[55] M. A. Tanner, *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*, vol. 67 (Springer Science & Business Media, 2012), doi:10.1137/1035020.

[56] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, Nat. Rev. Phys. **3**, 422 (2021), doi:10.1038/s42254-021-00314-5.

[57] G. T. Craven, N. Lubbers, K. Barros, and S. Tretiak, J. Phys. Chem. Lett. **11**, 4372–4378 (2020), doi:10.1021/acs.jpclett.0c00627.

[58] G. T. Craven, N. Lubbers, K. Barros, and S. Tretiak, J. Chem. Phys. **153**, 104502 (2020), doi:10.1063/5.0017894.

[59] S. Hayes, J. Thomas, and K. Peddicord, J. Nucl. Mater. **171**, 289 (1990), doi:10.1016/0022-3115(90)90376-X.

[60] B. Iooss and P. Lemaître, *A Review on Global Sensitivity Analysis Methods* (Springer US, Boston, MA, 2015), pp. 101–122.

[61] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, *Global Sensitivity Analysis: The Primer* (John Wiley & Sons, 2008).

[62] M. D. Morris, Technometrics **33**, 161 (1991).

[63] F. Campolongo, J. Cariboni, and A. Saltelli, Environmental Modelling & Software **22**, 1509 (2007), doi:10.1016/j.envsoft.2006.10.004.

[64] J. Herman and W. Usher, The Journal of Open Source Software **2** (2017), doi:10.21105/joss.00097.