Title
Nuclear data evaluation augmented by machine learning

Authors
Vicente-Valdez, Pedro
Bernstein, Lee
Fratoni, Massimiliano

Publication Date
2021-12-01

DOI
10.1016/j.anucene.2021.108596

Peer reviewed

Annals of Nuclear Energy 163 (2021) 108596

Contents lists available at ScienceDirect

Annals of Nuclear Energy

journal homepage: www.elsevier.com/locate/anucene

Nuclear data evaluation augmented by machine learning

Pedro Vicente-Valdez *, Lee Bernstein, Massimiliano Fratoni

Nuclear Engineering Department, University of California, Berkeley, United States

ARTICLE INFO

Article history:
Received 23 April 2021
Received in revised form 15 July 2021
Accepted 21 July 2021

Keywords:
Machine learning
EXFOR
Uranium benchmark
Cross section evaluation

ABSTRACT

The accuracy of neutrons modeling and simulation tools strongly depends on the quality of the nuclear data. Data libraries are generated by evaluators combining physics-based model codes and experimental data. There are many instances where experimental data are not available, are not reported rigorously or are discordant. In such cases, the evaluators need to make an expert judgment exposing the generated data to human bias and large uncertainties. This work proposes to support the evaluators' complex tasks by leveraging Machine Learning (ML) and Artificial Intelligence (AI). Two proof-of-concept ML models, a Decision Tree and K-Nearest-Neighbor, were developed to fit nuclear data from the EXFOR database in order to infer neutron induce reaction cross sections. Both models were used to predict nuclear data for ^{233}U, a well-characterized isotope in literature, and ^{35}Cl, a less studied but important nuclide for some advanced nuclear reactors. The predicted values for ^{233}U were validated using the ^{233}U Jezebel benchmark in Serpent2 model. The predicted values for ^{35}Cl(n,p) cross section were compared against recent new measurement not available in EXFOR. The predicted ML/AI values matched more accurately the new measurements than any of the evaluated data libraries, which overestimate experimental results by up to a factor of five. In turn, the proof-of-concept models explored in this work, reliant on learning underlying patterns of cross section data from other radionuclides, demonstrate evidence that ML models can aid traditional physics-guided models and have a role to play in nuclear data evaluations. Furthermore, incorporating ML models in the nuclear data pipeline can allow evaluators to make faster bias-free decisions in areas of uncertainty as well as better inform future data measurement campaigns on areas of greatest sensitivity in EXFOR.

1. Introduction

Research and development heavily rely on modeling and simulation in order to make rapid progress, reduce costs, and ensure safety. In the field of nuclear science and engineering, the need to understand the behavior of nuclear particles further exacerbates the reliance on simulation software, therefore, major efforts are directed to advance modeling tools aiming to improve fidelity, increase accuracy, and decrease computing time. Nevertheless, it is well understood that the reliability of any radiation transport modeling tool strongly depends on the accuracy of the nuclear data it relies on. These data are provided in the form of libraries such as the Evaluated Nuclear Data File (ENDF) or the Joint Evaluated Fission and Fusion (JEFF) that contain recommended data for energy differential cross sections, fission product yields, decay data, covariance for neutron cross section, product energy-angle distributions, etc. (Chadwick and Oblo, 2006). These libraries are generated by evaluators using a variety of tools including physics-based model codes (i.e., TALYS and EMPIRE) where the calculations are guided by data from experimental databases like the Experimental Nuclear Reaction Data (EXFOR) database. Recommended values are used all over the world by researchers and industry in Monte Carlo and deterministic modeling codes to simulate a variety of systems and phenomena including nuclear reactors, radiation detectors, particle accelerators, medical treatments, etc. All such derived studies rely on the available data and benchmarks, as well as, the judgment of the evaluators, to provide the best set of recommended values. However, there are many instances where experimental data are not available for a specific isotope-reaction channel pair or limited to a specific energy range, where experimental data were not collected with adequate rigor, or where experimental data sets present large discrepancies. It is the job of the evaluators to assess the trustworthiness of the data, select the physics model to combine with the experimental data, and ultimately fill the gaps to recommend data over the wide energy range that is in general of interest. This process can lead to large uncertainties and is vulnerable to human bias, in particular in

* Corresponding author.
E-mail address: pedro.vicentevz@berkeley.edu (P. Vicente-Valdez).

https://doi.org/10.1016/j.anucene.2021.108596
0306-4549/

those cases where experimental data are scarce or nonexistent. This work proposes to support the evaluators' complex tasks by leveraging Machine Learning (ML) and Artificial Intelligence (AI) during the process of cross section evaluation. The large set of experimental data collected throughout the years presents a logical opportunity for ML applications that can inform cross sections in areas of uncertainty, iterate faster through the evaluation steps, and ultimately reduce, if not eliminate, the human bias from the process. Such ML-driven tools are not meant to replace the evaluator or the physics-guided tools, but to enhance their analytical power and extract meaningful physics that can serve for future evaluations.

This manuscript discusses the potential role of ML in supporting cross section evaluation and proposes an ML augmented nuclear data evaluation pipeline for neutron-induced reactions cross section. The potential impact of this new process for evaluating cross sections is illustrated using two examples: ^{233}U and ^{35}Cl cross sections. Section 2 reviews the current evaluation pipeline, Section 3 describe the proposed methodology based on ML, and Section 4 illustrates the two examples. Finally, conclusions and additional considerations are provided.

2. Background

The United States government and many other international private and public entities have spent considerable resources collecting experimental nuclear data throughout many decades to better understand the behavior of isotopes important for nuclear power and nuclear weapon development such as uranium and plutonium isotopes. Additionally, data for many of the structural materials used in the current fleet of light water reactors have been extensively measured. Nevertheless, significant gaps remain in the cross sections data of many elements, isotopes, reaction channels, or energy ranges. Such missing data often constitute a critical obstacle to the development of new technologies like an advanced nuclear reactor. Filling data gaps, in general, requires lengthy and costly experimental campaigns. ML models, instead, can leverage existing data from well explored (isotope, reaction-channel) pairs and learn patterns and behaviors that can be applied to other less measured isotopes including those envisioned to be used in advanced reactor concepts. More specifically, the EXFOR database contains more than four million datapoints of neutron-induced reactions. The amount of data needed for an ML algorithm varies depending on the challenge and type of algorithm. It is the general consensus that as the number of available samples increases both the confidence and accuracy of the model is expected to increase (Jain et al., 1982). This is especially true in data-hungry models such as neural networks. Depending on the area of application, such as the nuclear data field, more data may not be easily gathered and/or might be extremely expensive. The only way to truly assess if the available data will suffice is by attempting to fit, optimize, and assess several models that might be appropriate for the challenge. Having collected this amount of data in EXFOR makes the field a prime candidate for the application of ML methods.

Currently, modeling neutron-induced reaction cross sections is a non-trivial problem as there is no one-model fits-all approach and the physics are not fully understood. Traditional approaches include using physics-based tools guided by experimental data to create a set of recommended values. In these tools, several physical parameters need to be adjusted for the model to generate a good function that fits and explains the experimental cross section data accurately although in some cases, these parameters can give too much flexibility consequently overfitting the data at the expense of physical meaning.

The current nuclear data evaluation pipeline can be summarized in four steps: compilation, evaluation, processing, and validation (Bernstein and Brown, 2019). In the compilation phase, nuclear data from references contained in the Nuclear Science References database are extracted and compiled into the EXFOR library in the case of nuclear reaction data (Pritychenko et al., 2011). In the evaluation phase, the relevant data are used to guide physics-based model calculations which will result in best estimates, dependent on data availability, of mean values including uncertainties and covariances. These values can then form part of one or more regional libraries (i.e., ENDF/B, JENDL). In the processing step, data are processed into a format compatible with a particular application. It is a nontrivial step that requires knowledge of the evaluation process and physics used. The last phase, validation, consists of using the processed evaluated data on codes and a set of defined problems to measure its performance. These problems more popularly consist of integral benchmarks data which are measured with a precision of several orders of magnitude (N. E. Agency, 2020). An example of these integral benchmarks are critical assemblies where the multiplication factor (k_{eff}), the ratio of neutron production to neutron loss, is known. The evaluated data must perform adequately or at least be an improvement relative to the preceding evaluation otherwise the new evaluation is rejected and the process needs to start over.

The fields of ML and AI have improved drastically since their inception in the 1980s in part thanks to the fast increase in computational power over the last decade. This includes better Core Processing Units, faster storage Input/Output speeds, and the development of better Graphical Processing Units. The latter, coupled with the wide availability of data and more efficient ML algorithms and optimization methods, allows for more accurate models with faster training times. Consequently, the prominence of ML/AI throughout business and technology sectors has only grown, translating into a significant impact in people's daily life. Despite the continuous growth of ML/AI throughout virtually every industry, its extension/application to Nuclear Physics and Engineering remains somewhat limited.

3. Methodology

Fig. 1 presents a potential evaluation pipeline augmented by ML. Its steps can be summarized as (1) dataset creation, (2) feature extraction and processing, (3) ML model training and evaluation, (4) hybrid library generation, and (5) validation which in turn informs model selection. Similar to the current nuclear data pipeline, it ends on the application side. The first step consists of collecting the relevant experimental data. For ML to effectively tackle this challenge, a representative dataset containing the physical properties and features that are believed to affect cross section behavior is needed. This mainly includes data from the EXFOR database, the Atomic Mass Evaluation, the Evaluated Nuclear Structure Data File, and any other appropriate experimental data source that may have a role/impact in cross section behavior. After collecting all relevant data feature extraction and processing are performed. This step consists mostly of cleaning the data and transform it into a form suitable for ML algorithms. This process is model-dependent and needs to be carefully performed. After having all features in an ML-friendly format, the chosen model is trained and its performance is first evaluated using an unseen dataset subset. The trained model can then be used to generate ML-derived recommended values (libraries) based purely on patterns and behaviors it learned from the training dataset. These cross sections are then validated using benchmark calculations. In this particular challenge, it is not possible to just rely on the validation subset performance for model selection. The error from the

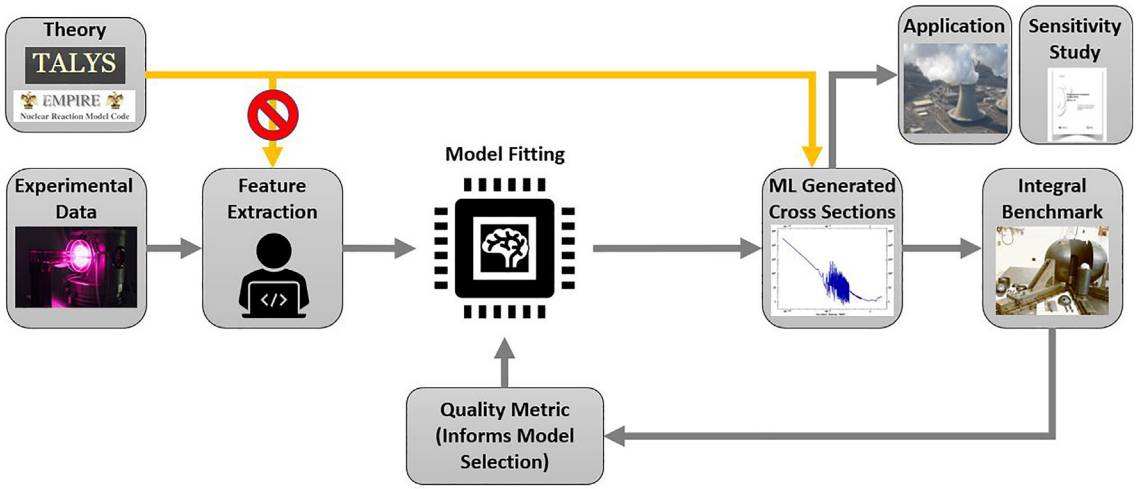

Fig. 1. Machine Learning Augmented Nuclear Data Evaluation Pipeline.

benchmark must have a higher weight on model selection since these are the closest representations to real-world deployment scenarios. However, the benchmark loss is currently not directly taken into account in model training but it provides essential metrics for further improvement. This process is iterative until an optimal set of model parameters is found. The selected models can then be used to make predictions to inform evaluators in areas of uncertainties where measurements have not been made or have been made but not accurately enough.

Several aspects require caution when implementing ML pipelines to this area in particular. Theoretically derived data including physical parameters or factors that scientists think can help the model learn behaviors in cross sections must not be included in the dataset. Including them will cause the ML model to have the same constraints as those other models used to derive said theoretical data. For example, filling missing values with ENDF data means the model will not only learn characteristics from the real physical phenomena but also from those models used to create the evaluated library (i.e., TALYS, EMPIRE). In addition, it inherits the bias in the creation of these values. Next, splitting must be performed in a stratified manner rather than randomly. It is important to fit the model on a representative diverse dataset, otherwise, it will have poor performance on data types that were not included in the training set. A clear example is elastic scattering of which there are thirty-seven thousand datapoints. Approximately two-thirds of those belong exclusively to ^{237}Np. Random splitting risks the majority of the elastic scattering points in the training subset belonging to ^{237}Np. This can translate into poor predicting capabilities on other non-seen isotopes. Another issue arises in the validation phase. Similar to the current nuclear data pipeline, integral benchmarks are part of the validation step to inform model selection. This means there is a risk of overfitting to a particular benchmark. The model selection process must be based on the average performance on all benchmarks. On another aspect, cleaning the data is a non-trivial process. This step is a hybrid realm between the data sciences and the nuclear data field. Although there might be good causes to discard one dataset versus another, this is outside the scope of the work presented here.

ML applied to cross section and nuclear evaluation presents very unique characteristics that are not usually present in common ML challenges due to the physical nature and meaning of the EXFOR database. EXFOR datapoints consist of experimental measurements of a variety of isotopes undertaken by researchers at different facilities. These carry uncertainty that is a function of the experimental settings, procedures, and more. Measurements are, therefore, only an approximation of an unknown true cross section value. The evaluation pipeline recognizes this and incorporates the validation phase by the use of benchmarks. For example, a critical assembly has a multiplication factor of one (critical). This is an actual true value. In other words, the assembly is critical or it is not (there is no in-between) and there exists a set of true cross section values that allows this assembly to exists in a criticality state given the composition and the geometry. These hidden sets of real cross section values would ideally be the labels in the datasets, however, only measurements are available that are regarded as being close enough to the unknown true values. Competing experimental campaigns and outliers exist in EXFOR and this is why traditionally the evaluator inspects each dataset for quality. To limit bias, the proposed pipeline does not discard datapoints manually and, therefore, an odd issue arises.

Most machine learning models try to minimize a given loss function which itself is a function of the predicted value and the given label. Knowing that the labels are not ideal or accurate enough and that the benchmarks are currently not part of the

Table 1
Dataset features.

Feature Name	Values (min, max)
Incident energy (eV)[a]	5.7630e-10, 1.0170e + 11
Number of protons	1, 99
Number of neutrons	0, 156
Atomic mass number	1, 255
Reaction number (MT) [b]	1–4, 16–18, 22, 24, 28–29 32–33, 37, 41, 51, 101–108 111–113, 152–153, 155 158–161, 203, 1003, 1108 2103, 9000, 9001
Center of mass flag [b]	Lab, Center of mass
Target type [b]	Isotope, Natural
Atomic Mass (micro-amu)	1.0070e+06, 2.5509e+08
Target Atomic Radius (fm)	1.25–7.92
Neutron/Target Atomic Radius Ratio (fm)	1.0092e-01, 6.4000e-01
Mass Excess (keV)	-9.1652e+04, 8.4089e+04
Binding Energy (keV)	0.0000e+00, 8.7945e+03
β^- Decay Energy (keV)	−2.2898e+04, 1.8244e+04
S(2n) Energy (keV)	2.0254e+03, 3.7512e+04
S(2p) Energy (keV)	7.7180e+03, 3.6635e+04
S(n) Energy (keV)	1.0969e+03, 2.0577e+04
S(p) Energy (keV)	0.0000e+00, 2.0831e+04

[a]85% and 17% of the energy and cross section values respectively either did not report uncertainties or were not readily accessible, therefore, uncertainty for these features was not included. [b]Treated as categorical features.

training loss function directly, it may not be the best approach to completely minimize the loss with respect to the EXFOR dataset if this means increasing the loss with respect to the integral benchmarks. Given any model selection process (i.e., cross validation, train-val-test), an optimal model with the lowest mean absolute error (MAE) and no overfitting will in some cases perform worse in a benchmark than an over-fitted model. This is due to the "approximation" nature of the dataset, the high uncertainties in some measurements, and the frequent presence of outliers. In other words, the EXFOR database converged model might not be the best model. Given these arguments, one must understand the limitations and carefully go about the training process on these types of experimental datasets.

For this work, a new python library (NucML) was developed that allows to quickly navigate through the proposed ML evaluation pipeline in an automatized manner. NucML provides utilities for custom dataset creation and manipulation, model building and testing capabilities, ML-based library generation, benchmark testing, and visualization tools.

3.1. Data sources and processing

This section provides a brief overview of the data used for training and comparison of the ML models: the EXFOR database and the ENDF/B-VIII.0 library. The dataset features [1] are listed in Table 1.

3.1.1. EXFOR database

The nuclear reaction compilation database, known as the EXFOR library was collected from the International Atomic Energy Agency (IAEA) (Otuka and Dupont, 2014). By means of international collaboration between the Nuclear Reaction Data Centres and supervised by the IAEA Nuclear Data Section, a compilation of experimental data takes place by identifying the literature for publications where neutron-induced reaction measurements have been made. Once identified, the National Nuclear Data Center, the NRDC institution for data measured in the United States and Canada, creates an EXFOR entry number and data is revised between institutions for incorporation into the EXFOR Master File. In addition to reaction cross sections, the EXFOR database contains resonance integrals, fission yields, polarization data, etc. For this work, only neutron-induced reaction cross sections (MF = 3) were extracted which corresponds to approximately 4.5 million datapoints. Cross section measurements dependent as ratios or other types of derived cross section data were not used. These data types can be transformed into useful data for this methodology or be used for post-training model evaluation but are out of the scope of this work.

The EXFOR library, being a platform for experimental results, contains a variety of conflicting experimental campaigns. For example, Fig. 2 shows the ^{35}Cl(n,p)^{35}S cross section experimental campaigns. In the 10^1 to 10^6eV energy region, the campaigns performed by Koehler and Popov differ significantly in almost every region including the $1/v$ region and the magnitude of the first two resonance peaks (Koehler, 1991; Popov, 1961). Another example of experimental variety is the ^{233}U(n,f) X cross section depicted in Fig. 3. This well-measured isotope is representative of the abundance and variety of different experimental campaigns encounter in other well-measured isotopes. Fortunately, resistance to outliers is one of the strengths of many ML pipelines with the added benefit of providing bias-free calculations. Because of this, the dataset was only checked for consistency and converted into a clean numerical format appropriate for the ML algorithm of choice. Conflicting experimental campaigns were left as appeared.

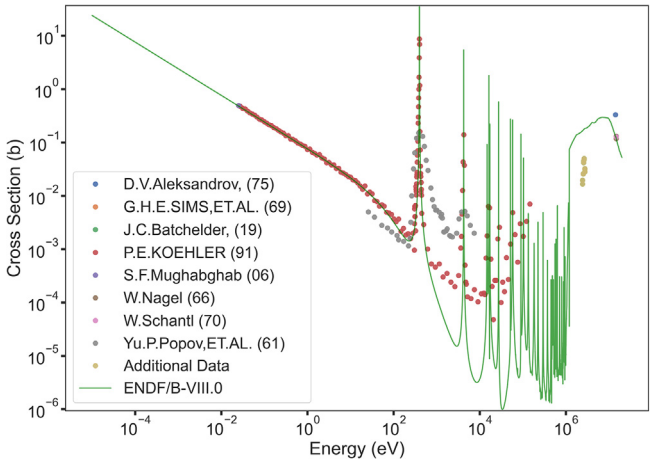

Fig. 2. ^{35}Cl(n,p)^{35}S reaction channel experimental datapoints in EXFOR vs. the ENDF/B-VIII.0 evaluation.

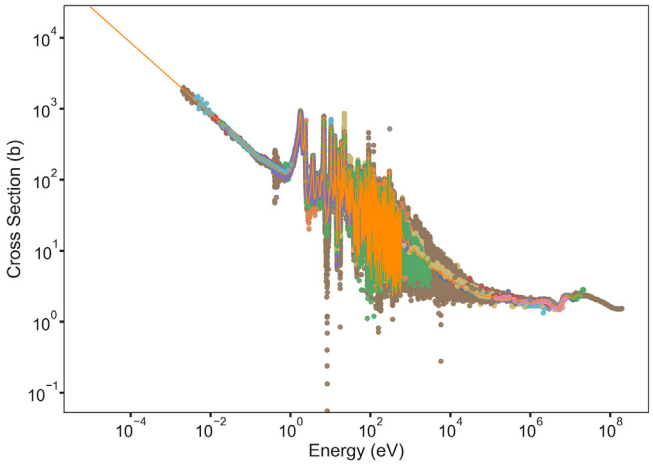

Fig. 3. ^{233}U(n,f) X reaction channel experimental datapoints in EXFOR vs. the ENDF/B-VIII.0 evaluation.

Although it is of interest to calculate the uncertainty in the ML model predictions, it is not possible to do so accurately given the current state of EXFOR. Of the four and a half million extracted data points, approximately four million energy points and eight hundred thousand cross section points uncertainty values were either not reported or were documented in another format not easily accessible. This makes it harder for the model to accurately infer or propagate uncertainties since it does not have enough accurate information to train on. Consequently, neither the uncertainty in energy nor in cross section was used to train the models presented here.

3.1.2. ENDF and JENDL library

Regional libraries often need to be processed further into a format suitable for application codes. For example, the ENDF library is often processed with tools like NJOY and AMPX into a format suitable for transport codes like MCNP, SERPENT2, and SCALE (MacFarlane et al., 2016; Kim et al., 2019; Werner et al., 2018; Leppänen and Pusa, 2015; Wieselquist et al., 2020). Once in the appropriate format, these libraries can be tested on the validation step using integral benchmarks (i.e., critical assemblies) from the

[1] In machine learning a feature is an individual measurable property or characteristic of a phenomenon being observed

International Criticality Safety Benchmark Evaluation Project (ICSBEP) Handbook, which contains criticality safety benchmark specifications from various experiments performed around the world (N. E. Agency, 2020). As a measure of current predicting power, the ENDF/B-VIII.0 library was used as a benchmark for the ML models to compare to. The error was measured with respect to the EXFOR datapoints. For the ^{35}Cl(n,p)^{35}S reaction channel the JENDL-4.0 evaluation (Shibata et al., 2011) was also employed.

3.2. Feature processing

This section provides an overview of the dataset transformations and standardization procedures which are important for model optimization and performance. First, all categorical features were one-hot encoded and were not subject to the transformations and normalization methods described here to preserve sparsity. One-hot encoding, also known as one-of-k, refers to the process of encoding categorical features as a one-hot numeric array meaning a binary column for each category is created. It is a typical step in any data processing pipeline for ML algorithms that require all data to be represented in numerical form. Afterward, the dataset was split into training, validation, and testing subsets in an 80–10-10 proportion. Although the performance on the validation subset was evaluated, the benchmarks are the main validation stage and the ultimate performance target.

3.2.1. Dataset transformations

Data ranges and skewness must be dealt with before any data transformation. Many of the data features are highly positively or negatively skewed including the incident energy. Highly skewed features can cause issues for many ML algorithms. The tail region of any skewed feature may act as an outlier for a statistical model, therefore, affecting the model's performance, especially in regression-based models (Han et al., 2012). Although there are models that are robust to outliers like Tree-based models, highly skewed features limit the possibility to try other models like K-Nearest-Neighbors and Neural Networks (John, 1995; SubbaNarasimha et al., 2000). Skewed data can be dealt with by transforming them into a Gaussian-like or Normal distribution by applying a Yeo-Johnson power transform (Eq. 1), a parametric-monotonic transform function, feature-wise for each instance x_i on the training set (Yeo and Johnson, 2000).

$$x_i^{(\lambda)} = \begin{cases} -\frac{[(-x_i+1)^{2-\lambda}-1]}{(2-\lambda)} & \text{if } \lambda \neq 2, x_i < 0, \\ \frac{[(x_i+1)^{\lambda}-1]}{\lambda} & \text{if } \lambda \neq 0, x_i \geqslant 0, \\ \ln(x_i+1) & \text{if } \lambda = 0, x_i \geqslant 0 \\ -\ln(-x_i+1) & \text{if } \lambda = 2, x_i < 0 \end{cases} \tag{1}$$

This method finds an optimal parameter (λ) that minimizes skewness through maximum likelihood estimation. Once the power transformation was fitted to the training data set, the same parameters were used to transform the testing set.

3.2.2. Standardization

All feature vectors (x) were standardized (Eq. 2) by removing the mean ($\bar{x}$) and scaling to unit variance using the feature standard deviation (σ).

$$x' = \frac{x - \bar{x}}{\sigma}. \tag{2}$$

This type of normalization is also known as Z-score normalization. It gives features the properties of a standard normal distribution with $\mu = 0$ and $\sigma = 1$ where μ is the mean and σ is the standard deviation from the mean. This is performed to prevent certain ML models to give more importance to higher magnitude features. Similarly to the power transformer, the mean and standard deviation for the training set was used to normalize the testing set. This step assumes that the training set is representative of the problem at hand. It helps also to identify outliers on the validation and testing set. In addition to the power transformer, a robust scaler was also tested to minimize the impact of outliers since it calculates the statistics based on the interquartile range (the range between the first and third quartile). These transformations were not applied to the dataset used to train the Decision Tree models.

3.3. Machine learning algorithms

There are a variety of regression algorithms including K-Nearest Neighbors (KNN), Linear and Logistic Regression, Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and Neural Networks (NN) (Goldberger et al., 2005; Han et al., 2012; Chang and Lin, 2011; Moisen and Service, 2008; Chen and Guestrin, 2016; Abadi et al., 2015). As the complexity of the model increases, more data for training is needed. The true goal of any ML model is generalization meaning adequate performance when deployed in real-world conditions with unseen data and, in this particular application, performance in benchmark calculations. For this proof-of-concept work, two basic but proven ML algorithms were chosen: KNN and DT. A brief description of KNN and DT algorithms is provided in this Section. Further details can be found in References (Goldberger et al., 2005; Moisen and Service, 2008). The Scikit-Learn implementations are employed for both models (Pedregosa et al., 2011).

3.3.1. K-nearest-neighbors

The KNN regression algorithm calculates a new data point's value by first querying the K nearest datapoints available in the training set and sorting them by increasing distance, hence the model's name. Here, K is a hyperparameter that has to be set before training. There is a variety of available distance metrics, including Euclidean and Manhattan, that can be used to measure the distance between points in the i dimensional space where i is the number of training features. In the case of Euclidean, the distance is calculated between points p and x by taking the square root of the squared difference between $(p_1, p_2, \ldots, p_n)$ and $(x_1, x_2, \ldots, x_n)$ (Eq. 3). The Manhattan distance in the other hand is calculated as the absolute difference between points x and p (Eq. 4). If the weights are set to uniform then the new data point's value is the mean of the K nearest points (Eq. 5). Alternatively, the K nearest points can be weighted according to distance, meaning closest points have a heavier influence on the new data point's value (they are weighted by the inverse of their distance).

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - x_i)^2}. \tag{3}$$

$$d(p,q) = \sum_{i=1}^{n}|p_i - x_i|. \tag{4}$$

$$\text{NewPoint} = \frac{1}{K}\sum_{i}^{K} k_i. \tag{5}$$

An issue with these types of algorithms is generalization beyond the training set features data ranges. For example, this means that the model will not perform well beyond the given energy ranges for a specific (isotope, reaction-channel) pair. Although a very simple algorithm, nearest neighbors have been successfully applied to a variety of challenges including satellite imagery classification and fraud detection, and it is worth exploring. Simple models also allow to study the characteristics of a dataset. Additionally, a major

advantage is the non-parametric nature of the algorithm. This makes it suitable for problems where decision boundaries are not well defined. The range of parameters tested is specified in Table 2.

3.3.2. Decision tree

DT models also belong to the family of non-parametric methods and can be applied to both classification and regression tasks. The model approaches both tasks by learning simple if-then-else decision rules from the available features. DT are white-box models meaning any prediction can be explained by boolean logic. The deeper the tree is allowed to be built, the more complex the rules become and consequently the more it is at risk of overfitting the dataset. The maximum depth is a hyperparameter that needs to be carefully selected before training and tuned in future iterations of the model. Additionally, DTs tend to be biased towards dominant classes or highly populated numerical regions in the training set. Achieving the right balance of classes (i.e. reaction channels) on the training data set is essential. An advantage of DT relative to KNN is the little data preparation needed. The best splits are found feature-wise independent of each other scale. There are a variety of DT implementations that are capable of creating trees with multiple branches, but the scikit-learn implementation is built around an optimized version of the CART algorithm meaning only binary trees are built through the training process.

The training process starts by feeding in the training vectors x_i and the label vector y. The decision tree will try to find an optimal point $\theta = (j, t_m)$ to split the data Q at node m by setting up a threshold t_m for each feature j in such a way that the impurity $G(Q, \theta)$ is minimized (Eq. 6).

$$G(Q,\theta) = \frac{n_{left}}{N_m} H(Q_{left}(\theta)) + \frac{n_{right}}{N_m} H(Q_{right}(\theta)) \tag{6}$$

where:

$$H(X_m) = \frac{1}{N_m} \sum_{i \in N_m} (y_i - \bar{y}_m)^2 \tag{7}$$

$$\bar{y}_m = \frac{1}{N_m} \sum_{i \in N_m} y_i \tag{8}$$

DTs have many other parameters to help reduce overfitting including the minimum samples required to make a split (MSS) and the minimum samples required for a node to become a leaf (MSL). The higher the number on both the more restricted the model will be resulting in less predictive power. A balance must be found between the max depth, the MSS, and the MSL. The range of tested model parameters is specified in Table 2. Contrary to KNN, DT does minimize a loss function given by the impurity (MSE). This makes it an ideal basic candidate to test the hypothesis that a lower MAE does not lead to a better performance in the benchmark in loss-minimizing algorithms.

4. Examples

The potential impact of ML to support cross section data generation for use in modeling and simulation is demonstrated in two examples. ^{233}U and ^{35}Cl cross sections were chosen as test cases for two opposite reasons. ^{233}U is a data-rich isotope and was chosen to validate performance and demonstrate the reliability of trained ML algorithms. Total, inelastic, fission, capture, and elastic cross sections of ^{233}U were generated, compiled into the ACE format, and tested by modeling the ^{233}U Jezebel benchmark with the Monte Carlo code Serpent 2 (Leppänen and Pusa, 2015). ^{35}Cl, instead, has far less experimental data and in particular, the lack of data for the (n,p) reaction in the energy range above 0.1 MeV has a substantial impact on the development of molten chloride fast spectrum reactors. In addition to reactor design, chlorine neutron absorption is thought to play an important role in other areas of nuclear criticality safety. For example, in storage and transportation of dual-purpose canisters, chlorine available in deep geological repository media can provide natural reactivity reduction which ensures sub-criticality of disposed material (Sobes et al., YYYY). Recently, a measurement performed at the University of California, Berkeley at about 2.5 MeV has shown that existing data libraries overestimate measured points by up to a factor of five (Batchelder and Chong, 2019). Therefore, ^{35}Cl was chosen to demonstrate the capability of the ML model to generate cross section data for data-starved reactions and its viability was evaluated against the new measurements whose results were not known to the algorithm at any stage.

Table 2
Tested K-Nearest-Neighbor and Decision Tree Model Parameters.

K-Nearest-Neighbor	
Number of Neighbors (K)	1–20
Weight Function	Distance, Uniform
Algorithm	Brute
Distance Metric	Manhattan, Euclidean
Decision Tree	
Criterion	Mean Squared Error
Splitter	Best
Max Depth	10–400
Minimum Samples for Split	2–15
Minimum Samples for Leaf	2–15

4.1. U-233 Jezebel benchmark

A SERPENT2 model was created for the ^{233}U Jezebel critical assembly based on the information contain in the ICSBEP Handbook, one of the three Jezebel assemblies that were fabricated and operated at Los Alamos Scientific Laboratory (R. Douglas O'Dell et al., 2020; N. E. Agency, 2020). This assembly consists of a bare ^{233}U metallic sphere (5.9838cm radius and 18.424g/cm^3 density) at room temperature of composition as specified on Table 3. The ^{233}U Jezebel benchmark was selected due to the simple, almost mono-isotopic nature. The generated cross sections included ^{233}U (n,tot), (n,γ), (n,inelastic), (n,elastic), and (n,f). For every channel, the ML algorithm was used to generate cross section values for the entire energy region. In line with the expected hybrid implementation, the 1/v region for all ML-generated cross sections (values up to the first resonance peak) was adjusted in post-processing due to the instabilities presented by both models using data from ENDF. In other words, the solution presented here consists of a machine learning and traditional tools hybrid modeling technique. All other uranium isotopes were left unchanged. During the compilation process, the cross sections were run through standard checks (i.e. making sure the appropriate reaction channels sum up to the total cross section).

4.1.1. KNN model selection and performance

While the size of the EXFOR dataset is large, brute-force calculations for the KNN algorithm are feasible. Even though the training and inferences time will be large, the challenge's objective does not carry a time constraint. This makes the training scenarios simpler by limiting the number of hyperparameters to the number of neighbors (K), the distance metric, and the weight function. Before training the models the impact of different scalers and normalizers in generalization performance tested, the robust scaler achieving overall better results. Due to the fact that cross section is a function of the energy, the "distance" weight function and the euclidean distance were chosen. The top graph in Fig. 4 represents the MAE for the train and validation sets as a function of the

Table 3
^{233}UJezebel Critical Assembly (R. Douglas O'Dell et al., 2020).

Isotope	Density (atoms/barn-cm)
^{233}U	4.6712E-02
^{234}U	5.9026E-04
^{235}U	1.4281E-05
^{238}U	2.8561E-04

parameter K (number of nearest neighbors). The model with $k = 8$ served as the traditionally selected model since it is at this point that the validation MAE starts to increase and deviate from a downwards trajectory while the train MAE kept decreasing indicating overfitting. Using this model, the Jezebel benchmark was run using the ^{233}U KNN-generated cross sections. The results for this model are listed in Table 4. The performance is adequate with an error of 0.55%.

To find if the traditionally selected model was indeed the best model in terms of benchmark performance, ^{233}U cross sections were generated using all KNN models. The bottom graph in Fig. 4 depicts the validation MAE and the multiplication factor as a function of K. As expected, the traditional model selection technique does not seem to correlate with benchmark performance. The train-based selected model ($k = 20$) provides better benchmark results than the traditionally selected model ($k = 9$) by around 75%. On the other side, handpicking the model with the best benchmark performance ($k = 18$) leads to a big improvement with an error of 0.011%. This is a big difference and already outperforms the ENDF-based benchmark results. One drawback of performing this kind of analysis in KNN models is that there is no loss function being optimized. The performance comes strictly from the dataset completeness. Still, it provides important insight into the dataset issues causing deviation from the expected validation MAE/multiplication factor correlation.

The deviations could be explained by a variety of factors. Using the best benchmark model ($k = 10$) for example. Fig. 5 shows the fission and capture cross section results along with the absolute differences relative to the ENDF cross sections. As expected, the resonance region is where most of the differences arise. Overall, the generated cross sections for ^{233}U appear to be similar to those in ENDF with the maximum difference being up to 130 barns for all reaction channels except the (n,elastic) cross section (Fig. 7). These differences should not be interpreted as an error since it is not the objective to minimize the differences between the current evaluations and the ML algorithms. As seen in Fig. 6, the experimental datapoints available are scarce. This limits the model's performance in this particular (isotope, reaction-channel) pair. Because of the high availability of other reaction channels, the MT 2 cross section was calculated as the difference between the ML generated (n,tot) and (n,nonelastic) cross sections. The resulting values show some erratic behavior at low energies. In intermediate energies, the resonances seem to have a higher magnitude in both directions relative to ENDF. This is a clear example of the limitations of using this type of model and performance metrics when working with databases like EXFOR. There are also differences in the transition between the resonance and fast energy regions. This is to be expected since resonances exist in the fast energy region but the available capabilities/resolution do not allow to measure them and therefore an average is taken in ENDF. In a real ML-hybrid modeling scenario, the evaluator would need to make a decision based on benchmark performance.

4.1.2. DT model selection and performance

Decision Trees on the other hand do optimize for MSE. As mentioned, a balance between the max depth, MSS, and MSL must be found to achieve good generalization and performance. Table 4 shows the results on the Jezebel benchmark for the validation-based and train-based selected models along with the final model parameters. Similar to the KNN results, the performance is adequate for the validation-based model with an error of around 0.28%. Having observed that traditional model selection techniques may not yield the best model in terms of benchmark performance, a parametric study was performed by generating the ^{233}U cross sections and running the benchmark using every trained DT model.

Fig. 8 depicts the validation MAE (left) and multiplication factor error (right) as a function of the max depth and the MSS. The train-based selected model achieved much better performance in the benchmark relative to the validation-based selected model. Similar to the outcome in the KNN based models, the figure shows that the lowest validation MAE does not yield the best benchmark performer. For both the validation and train MAE, there is a value where the optimal benchmark performer can be found. Handpicking the model based on benchmark performance leads to an improvement of huge improvement relative to the validation-based model with just an error of 0.0057%.

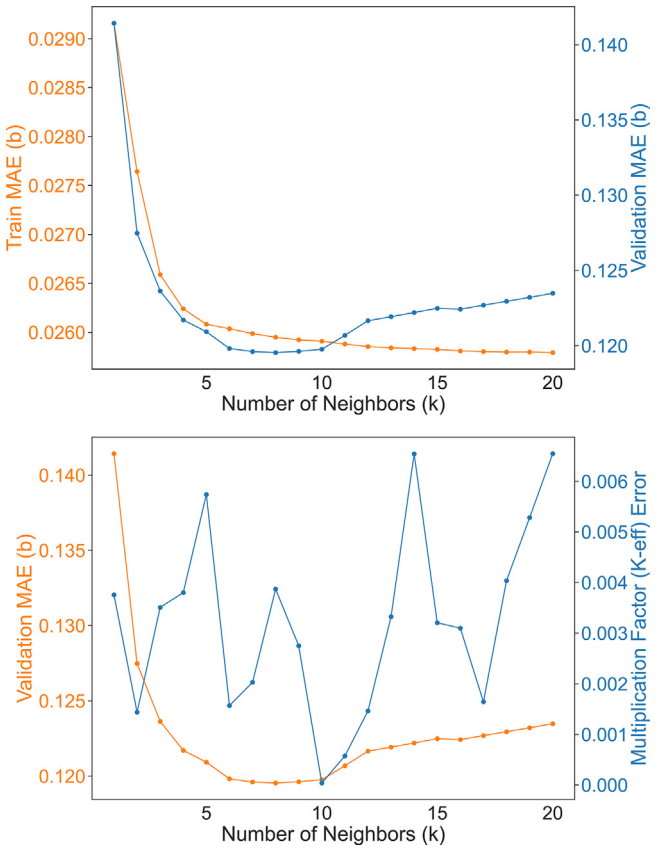

Fig. 4. Train and validation MAE as a function of the number of neighbors (k) (top). Validation MAE and Multiplication Factor (k_{eff}) error relative to 1 as a function of the number of neighbors (k) (bottom). Both the train and validation MAE are in a $\log_{10}$ scale.

4.2. Cl-35 cross section

The cross section for ^{35}Cl(n,p)^{35}S, a reaction of interest in the development of chlorine-based fast reactors, was inferred using both handpicked KNN and DT models. Recently, the joint Lawrence Berkeley Laboratory and UC Berkeley (LBNL/UCB) Nuclear Data Group measured the ^{35}Cl(n,p)^{35}S cross section for $2.42 < En < 2.74$ MeV neutron energies using a high flux neutron

Table 4
^{233}U Jezebel Benchmark Results.

Library	k_{eff}	Uncertainty	Error (%)	Train MAE	Validation MAE	Test MAE
ENDF	1.0002	+/- 0.0011	0.02	N/A	N/A	N/A
KNN[a] ($k = 9$)	1.0038	+/- 0.0004	0.38	0.025921	0.119010	0.118578
KNN[b] ($k = 20$)	0.9934	+/- 0.0004	0.66	0.025814	0.121110	0.120706
KNN[c] ($k = 18$)	1.0000	+/- 0.0004	0.00	0.025818	0.120711	0.120294
DT[a, d]	0.9971	+/- 0.0004	0.29	0.094443	0.118699	0.119142
DT[b, e]	1.0023	+/- 0.0004	0.23	0.025773	0.136140	0.135027
DT[c, f]	0.9999	+/- 0.0004	0.01	0.088061	0.120462	0.120684

[a]Validation-based selected model; [b]Train-based selected model; [c]Hand-picked model; [d]Max Depth = 70, MSS = 10, MSL = 7; [e]Max Depth = 400, MSS = 2, MSL = 1; [f] Max Depth = 80, MSS = 15, MSL = 3

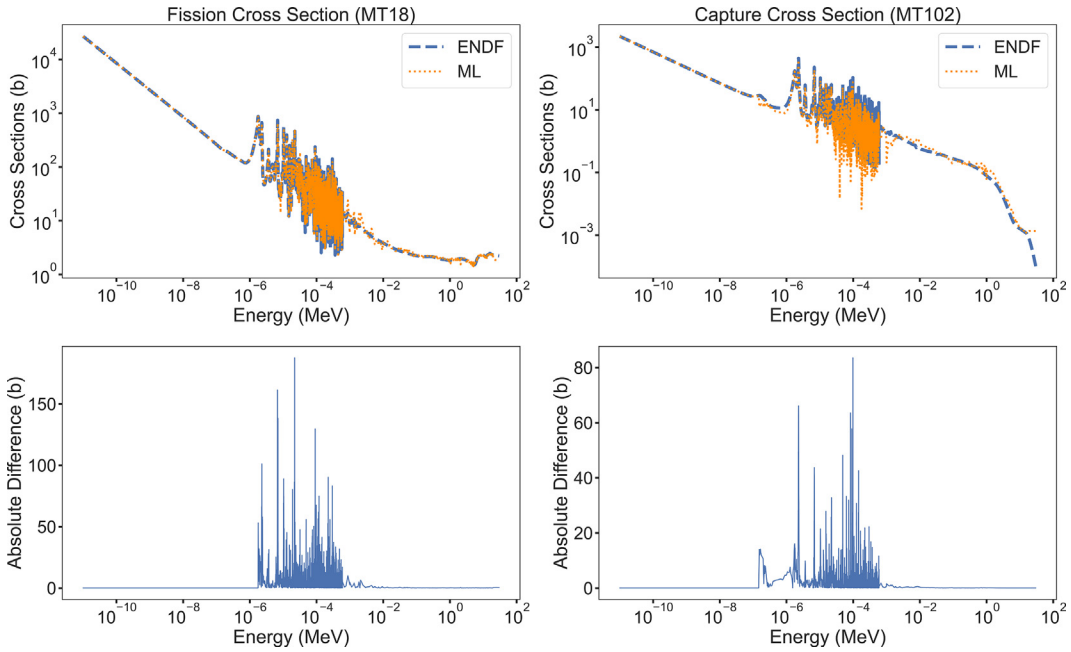

Fig. 5. Comparison of KNN generated and ENDF/B-VIII.0 cross sections for fission and radioactive capture.

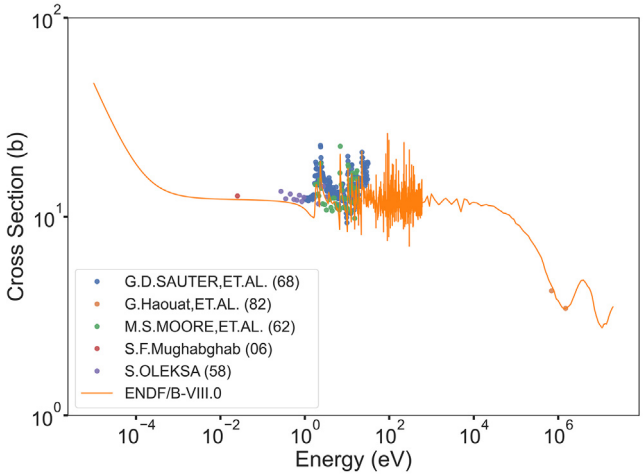

Fig. 6. ^{233}U(n,elastic) reaction channel experimental datapoints in EXFOR vs the ENDF/B-VIII.0 evaluation.

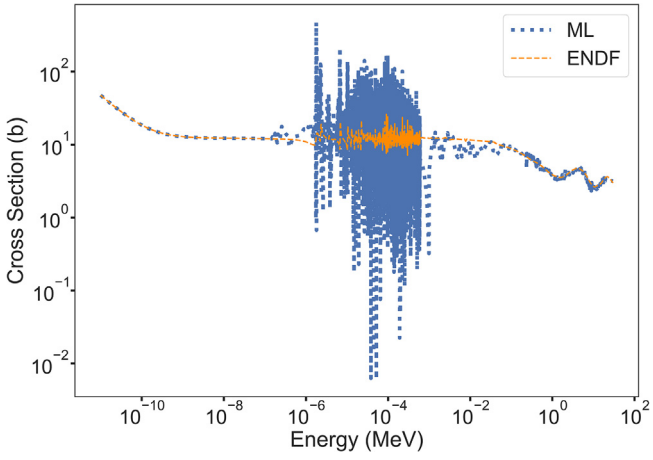

Fig. 7. KNN inferred cross section values for the ^{233}U elastic channel vs the ENDF/B-VIII.0 evaluation.

generator and decay gamma spectroscopy (Batchelder and Chong, 2019). Results ranged from 30–50 millibarn and were used in this work exclusively for validation of both the KNN and DT models. In other words, these datapoints were not part of EXFOR dataset. Fig. 9 shows the available EXFOR datapoints, the LBNL/UCB measured datapoints, the ENDF/B-VIII.0 evaluation, and the ML-generated cross sections. Table 5 shows the errors with respect to the LBNL/UCB measurements for the ML-generated values, the ENDF/B-VIII.0 library, and the JENDL-4.0 library.

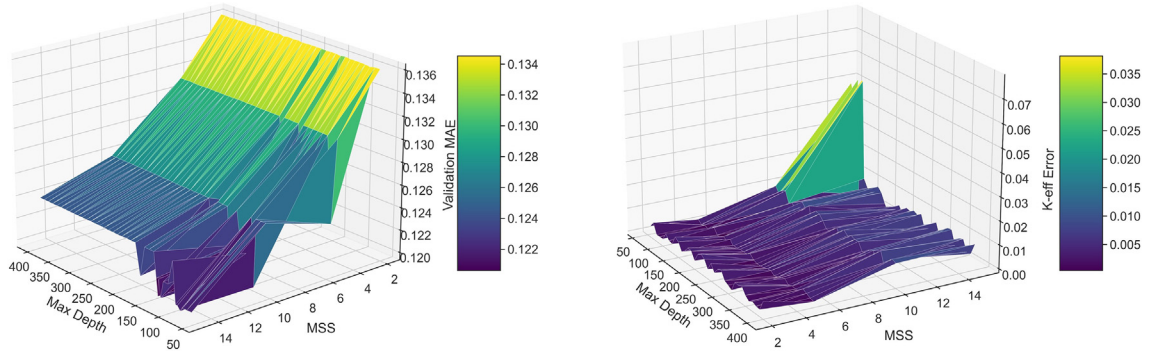

Fig. 8. Validation mean absolute error (MAE) and multiplication factor error relative to 1 as a function of the max depth and MSS. The validation MAE are in a $\log_{10}$ scale.

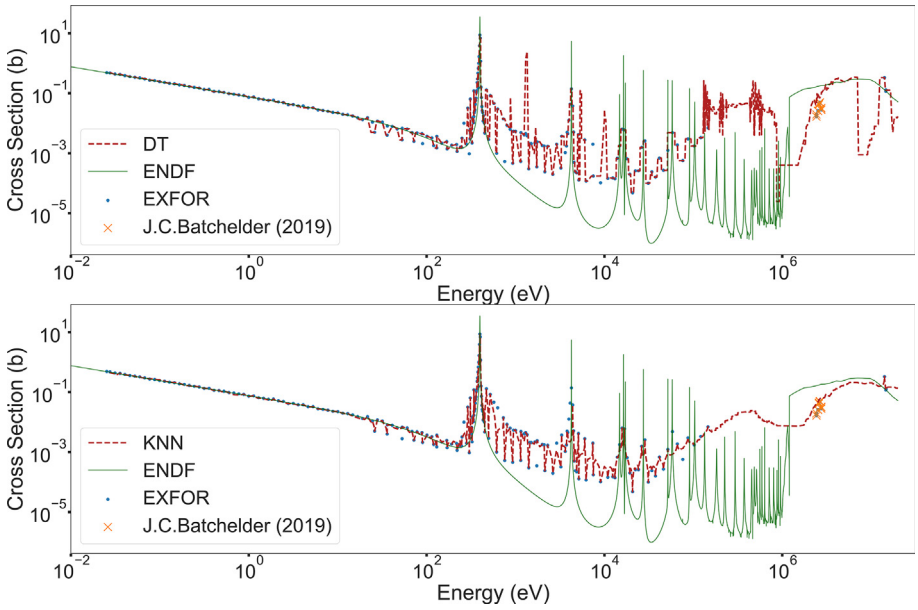

Fig. 9. KNN and DT generated cross section for the $^{35}Cl(n,p)^{35}S$ reaction vs the ENDF/B-VIII.0 library and the available EXFOR experimental datapoints.

Table 5
Evaluated Libraries and ML model predictions error on the LBNL New Measurements for $^{35}Cl(n,p)^{35}S$ Reaction

Energy (eV)	Cross Section (b)	Decision Tree (b)	KNN (b)	ENDF (b)	JENDL-4.0 (b)
2.42E6	0.0196	0.040300	0.040723	0.162609	0.165603
2.52E6	0.0261	0.091000	0.045894	0.162724	0.171788
2.58E6	0.0446	0.078307	0.054121	0.165948	0.173959
2.64E6	0.0500	0.070000	0.059396	0.168482	0.176129
2.74E6	0.0315	0.085276	0.064900	0.169618	0.178300
MAE		0.072977	0.053007	0.136624	0.145688

The DT algorithm was able to predict the cross section values better than any of the evaluated libraries by learning a set of if-then-else rules based on important features from not only other X(n,p) Y reactions but also other reaction channels. Not only was the magnitude of the newly measured datapoints more accurately described, but the expected trend after 10^6 eV seems similar to that of the ENDF evaluation even when no known datapoints were available in the entire dataset constraining the calculations. Similarly, the KNN also outperforms the ENDF and JENDL evaluations in these particular datapoints. There is certainly room for improvement. Due to the if-then-else nature, the DT predictions have a step-like behavior. Post-processing smoothing can be applied to make the $1/v$ region and the resonance peaks smoother. The same can be applied to the KNN model predictions. Both the DT and KNN seem to be overfitting some of the EXFOR datapoints in the $1/v$ region. More powerful algorithms can be applied that do not have the limitations that these two models possess. The next step in the process would be to test the newly generated cross sections in a benchmark. Although some critical configurations containing chlorine exists, most are thermal spectrum and will therefore provide limited information on the fast spectrum cross sections. A suitable benchmark was not identified.

5. Conclusions

Nuclear data evaluators have an important task at hand when an evaluation takes place and it is important they have the necessary tools to navigate this process. Physics-based reaction modeling codes (i.e., EMPIRE and TALYS), guided by experimental databases like EXFOR, are used to create an appropriate fit that explains most of the experimental data points. However, some evaluations are inevitably created with higher uncertainty and human bias, especially in reactions where experimental data is not available to constrain the model calculations.

This work proposes a framework for ML-augmented nuclear evaluations that would reduce human bias and accelerate the evaluation process. ML models were built based on the experimental data contained in the EXFOR database and atomic properties, and were tested in two example cases. In general, the success of any ML model not only depends on the characteristics of the chosen algorithm but also on the quality and quantity of the data obtained. The latter is often regarded as the most important aspect for the successful application of ML approaches. In the nuclear data field, data is expensive and sometimes wrongly reported. For example, a large percentage of the energy and cross section values in EXFOR have either missing uncertainties or were not logged appropriately. Despite this, the results obtained show the capability of ML algorithms to inform evaluations and provide useful information in areas where no experimental data is available. Both the K-Nearest-Neighbor and Decision Tree models performed adequately on the ^{233}U Jezebel Benchmark, in some cases providing results closer to the experimental values than data from the ENDF library. The fast evaluation aspect is also evident. The KNN model only took 2.4 h whereas the DT model took just a couple of minutes. Training the entire set of KNN and DT models took in total four days. Furthermore, the (n,p) cross section for ^{35}Cl, a less measured nuclide, was investigated. Both the KNN and DT model, reliant on learned patterns and behaviors of other X(n,p) Y reactions, predicted the latest LBNL/UCB measurements more accurately than any of the evaluated data libraries which overestimate experimental results by up to a factor of five. This also demonstrates the potential for ML models to aid traditional physics-guided models. Due to the issues presented by training an ML algorithm in this type of experimental dataset, the best KNN and DT models were handpicked based on the benchmark performance. However, the selection process should be automated potentially relying on the best average performance in a validation set of benchmarks.

Future work also includes using more powerful algorithms that can generalize well beyond the given data ranges, a current limitation of both the KNN and DT models. These include powerful algorithms like Gradient Boosting and advanced and complex models like Deep Neural Networks (DNN) which also allow for multi-output calculations, a feature useful for incorporating uncertainty calculations. For these capabilities to be accurate and reliable, the current state of the EXFOR database needs to be improved. Additionally, DNNs allow the incorporation of custom loss functions which can be written to enforce unitarity during training rather than post-training. New EXFOR specific loss functions can also provide robustness to outliers provided an optimal architecture is found. Including new relevant features like the number of valence neutrons and protons, and the promiscuity factor can also provide more information than complex new algorithms.

CRediT authorship contribution statement

Pedro Vicente-Valdez: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Lee Bernstein:** Conceptualization, Methodology, Resources, Writing - original draft, Writing - review & editing, Supervision, Funding acquisition, Project administration. **Massimiliano Fratoni:** Conceptualization, Methodology, Resources, Writing - original draft, Writing - review & editing, Supervision, Funding acquisition, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been carried out under the auspices of the U.S. Department of Energy by Lawrence Berkeley National Laboratory and the U.S. Nuclear Data Program under contract # DE-AC02-05CH11231.

References

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, (2015), software available from tensorflow. org..

Batchelder, J.C., Chong, S.A., et al., 2019. Phys. Rev. C 99, 1.

Bernstein, L.A., Brown, D.A., et al., 2019. Annu. Rev. Nucl. Part. Sci. 69, 109.

Chadwick, M.B., Oblo, P., et al., 2006. Nucl. Data Sheets 107, 2931.

Chang, C.-C., Lin, C.-J., 2011. ACM Trans. Intell. Syst. Technol., 2, p. 27:1. software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Chen, T., Guestrin, C., 2016. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 (ACM, New York, NY, USA, 2016) pp. 785–794..

Douglas O'Dell, R., Brewer, R.W., Atkinson, C.A., 2020. ICSBEP Handbook 2020 5.

Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R., 2005. Adv. Neural Inform. Processing Syst. 17, 513.

Han, J., Kamber, M., Pei, J., 2012. Data Mining: Concepts and Techniques, third edition ed., edited by J. Han, M. Kamber, and J. Pei, The Morgan Kaufmann Series in Data Management Systems (Morgan Kaufmann, Boston, 2012)..

Jain, A., Chandrasekaran, B., 1982. Classification Pattern Recognition and Reduction of Dimensionality. Handbook of Statistics, Vol. 2. Elsevier, pp. 835–855.

John, G.H., 1995. In Knowledge Discovery and Data Mining. AAAI Press, pp. 174–179.

Kim, K.S., Williams, M.L., Holcomb, A.M., Wiarda, D., Jeon, B.K., Yang, W.S., 2019. Ann. Nucl. Energy 132, 161.

Koehler, P.E., 1991. Phys. Rev. C 44, 1675.

Leppänen, J., Pusa, M., et al., 2015. Ann. Nucl. Energy 82, 142.

MacFarlane, R.E., Muir, D.W., Boicourt, R.M., et al., 2016. The NJOY Nuclear Data Processing System, Version 2016, Tech. Rep. (Los Alamos National Laboratory, 2016)..

Moisen G.G., Service, U.S.F., 2008. Ecological Informatics, 582..

N. E. Agency, (2020), https://doi.org/https://doi.org/10.1787/7e0ebc50-en..

Otuka, N., Dupont, E., et al., 2014. Nucl. Data Sheets 120, 272.

Pedregosa, F., Weiss, R., Brucher, M., 2011. J. Mach. Learn. Res. 12, 2825.

Popov, Y.P., S.F. L., J. Exp. Theor. Phys. 13, 1132 (1961)..

Pritychenko, B., Běták, E., Kellett, M., Singh, B., Totans, J., 2011. Nuclear Instruments and Methods in Physics Research Section A: Accelerators. Spectrometers, Detectors Associated Equipment 640, 213.

Shibata, K., Iwamoto, O., et al., 2011. J. Nucl. Sci. Technol. 48 (1).

Sobes, V., Scaglione, J.M., Wagner, J.C., Dunn, M.E., YYYY. Validation study for crediting chlorine in criticality analyses for spent nuclear fuel disposition..

SubbaNarasimha, P., Arinze, B., Anandarajan, M., 2000. Expert Syst. Appl. 19, 117.

Werner, C.J., Bull, J.S., Solomon, C.J., Brown, F.B., McKinney, G.W., Rising, M.E., Dixon, D.A., Martz, R.L., Hughes, H.G., Cox, L.J., Zukaitis, A.J., Armstrong, J.C., Forster, R. A., 2018. L. Casswell. https://doi.org/10.2172/1419730.

Wieselquist, W., Lefebvre, R.A., Jessee, M., 2020, 10.2172/1616812..

Yeo, I.-K., Johnson, R., 2000. Biometrika 87. https://doi.org/10.1093/biomet/87.4.954.