npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

# Machine learning predictions of irradiation embrittlement in reactor pressure vessel steels

Yu-chen Liu[1,2,3], Henry Wu[1], Tam Mayeshiba[1], Benjamin Afflerbach[1], Ryan Jacobs[1], Josh Perry[1], Jerit George[1], Josh Cordell[1], Jinyu Xia[1], Hao Yuan[1], Aren Lorenson[1], Haotian Wu[1], Matthew Parker[1], Fenil Doshi[1], Alexander Politowicz[1], Linda Xiao[1], Dane Morgan[1]✉, Peter Wells[4], Nathan Almirall[4], Takuya Yamamoto[4] and G. Robert Odette[4]

Irradiation increases the yield stress and embrittles light water reactor (LWR) pressure vessel steels. In this study, we demonstrate some of the potential benefits and risks of using machine learning models to predict irradiation hardening extrapolated to low flux, high fluence, extended life conditions. The machine learning training data included the Irradiation Variable for lower flux irradiations up to an intermediate fluence, plus the Belgian Reactor 2 and Advanced Test Reactor 1 for very high flux irradiations, up to very high fluence. Notably, the machine learning model predictions for the high fluence, intermediate flux Advanced Test Reactor 2 irradiations are superior to extrapolations of existing hardening models. The successful extrapolations showed that machine learning models are capable of capturing key intermediate flux effects at high fluence. Similar approaches, applied to expanded databases, could be used to predict hardening in LWRs under life-extension conditions.

*npj Computational Materials* (2022)8:85 ; https://doi.org/10.1038/s41524-022-00760-4

## INTRODUCTION

Nuclear power plants are an important source of electricity in the U.S., averaging 20% of U.S. electricity generation per year since 1990, which in 2018 amounted to about 1.1 TW[1]. This electricity generation comes from about 98 reactors at 60 power plants, with an average service age of about 38 years. The oldest plants began commercial operation in 1969. Over half of the plants began commercial operation between 1985 and 1996[1]. Commercial nuclear reactors had an initial license of 40 years, and most have now been licensed for 60 years[2]. Further lifetime extensions to 80 or even 100 years may be convenient and cost-effective, but must be based on reliably predicting reactor safety and integrity to such long times.

This paper addresses the issue of reactor pressure vessel (RPV) steel embrittlement, due to exposure to high-energy neutrons during service, manifested as decreased fracture toughness[3]. The RPV is the large 'outer shell' of a light water reactor (LWR) that enables the use of pressurized water for high-temperature operation as well as provides a barrier against the release of radiation in the event of a core damaging accident. Replacement of the RPV is considered economically unfeasible[4], although in recent years the Rosatom corporation announced it has extended the service lifetime by using thermal annealing on the VVER-1000's RPV[5]. Even with these thermal annealing measures, improved predictions of embrittlement and improved understanding of its causes are key components in determining the safety of LWRs for possible additional lifetime extensions.

An extensive recent review of modeling RPV embrittlement can be found in Odette, et al[3]. and a short introduction follows here. RPV embrittlement has been studied since the 1950s[6], producing the general understanding that radiation-enhanced diffusion accelerates the precipitation of copper-rich precipitates (CRPs), Mn-Ni-Si-rich precipitates (MNPs), and other solute-defect complex matrix features which impede dislocation motion and increase the steel yield stress $\sigma_y$ and flow stress[6]. The increase in the yield stress relative to an unirradiated reference sample, $\Delta\sigma_y$, results in an upward shift in the ductile to brittle transition temperature (DBTT), of Charpy impact tests energies, typically indexed at 41 J[7]. The transition temperature shift (TTS) increases under irradiation and is accompanied by a decrease in the Charpy upper-shelf energy ductile fracture toughness[8]. Current regulatory predictions of TTS are derived from the physics-based Eason-Odette-Nanstad-Yamamoto (EONY) model[6] and ASTM E900 standard practice[9]. Both of these models were evaluated based on actual RPV surveillance data[6,10]. These models have been extended to a higher effective fluence (EF) in the recently developed Odette-Wells-Almirall-Yamamoto (OWAY) model[3], which integrates highly physical models with empirical fitting to the data from Irradiation Variable (IVAR) and Advanced Test Reactor 1 (ATR1) experiments as well as very recent data from Advanced Test Reactor 2 (ATR2) experiments. The OWAY model, although preliminary, provides what is perhaps the most complete model presently available for high fluence RPV embrittlement. Specifically, OWAY is a reduced order fit for composition effects at one irradiation and flux (the ATR2 test) condition. The OWAY model is fully data driven and can be evaluated easily. It is physical in the sense that it is backed by and consistent with detailed physical models. While the OWAY model provides an excellent tool, it required significant human effort and physical insight to develop and has a very specific domain of applicability. There are therefore potentially still many advantages in speed of development, flexibility, and generality to establishing the efficacy of machine learning (ML) hardening models with few or no physical assumptions.

At the time of writing the ATR2 database was still being completed and was not fully available for use in this work. Therefore, the actual ATR2 experimental data has not been used in the present study. Instead, we used the chemical factor (CF) OWAY

[1]Materials Science and Engineering Department, University of Wisconsin-Madison, Madison, WI 53706, USA. [2]Hierarchical Green-Energy Materials (Hi-GEM) Research Center, National Cheng Kung University, Tainan 70101, Taiwan. [3]Materials Science and Engineering Department, National Cheng Kung University, Tainan 70101, Taiwan. [4]Mechanical Engineering Department, University of California, Santa Barbara, CA 93106, USA. ✉email: ddmorgan@wisc.edu

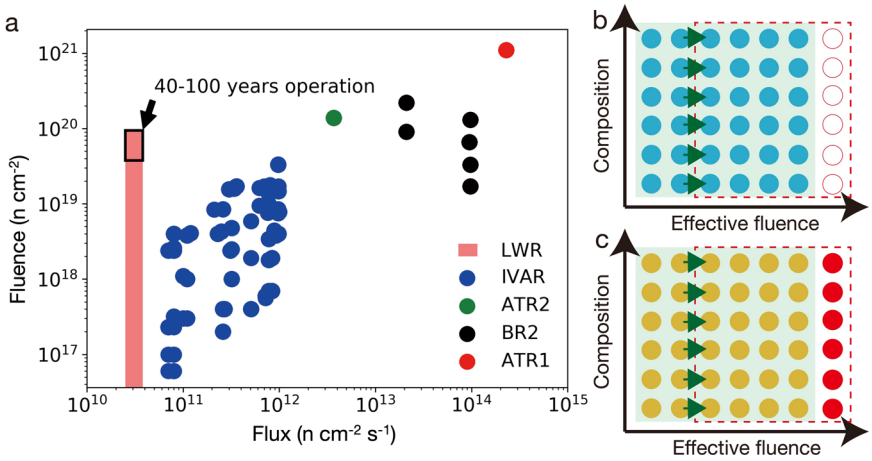


**Fig. 1 Dataset and model assessment of the present study. a** Fluence and flux regions of the University of California Santa Barbara (UCSB) Irradiation Variable (IVAR), Belgian Reactor 2 (BR2), Advanced Test Reactor 1 (ATR1) data and Advanced Test Reactor 2 (ATR2) data, as well as the light water reaction (LWR) condition region. IVAR, BR2 and ATR1 data were combined in this work and referred to collectively as 'IVAR+'. The black rectangle represents 40–100 years operation at the LWR conditions. **b**–**c** This figure also shows schematic diagrams of the model assessments of Test 2. The shaded green area, the red dotted rectangular and the green arrows are the training data set, the validation or testing data set and the targets the model is predicting, respectively. The blue solid circles shown in (**b**) are the IVAR+ data set, while the yellow solid circles shown in (**c**) are the CD-IVAR+ data set. The red circles shown in (**c**) are the testing data set at the LWR conditions of high effective fluence regimes for the alloys in IVAR+ data set. The hollow and solid red circles shown in (**b**) and (**c**) represent that we don't have such data experimentally and we do have such data simulated by cluster dynamics, respectively.

model-predicted hardening data (i.e., CF OWAY ATR2) for our model assessment[3]. Recent work stated that the standard deviation of the CF OWAY model in predicting ATR2 conditions was 18.9 MPa when comparing with experimental data[3], which is likely close to the experimental uncertainty in the experimental data itself. We therefore consider the CF OWAY data to be an adequate approximation to radiation response under ATR2 conditions and suitable for assessing the present ML models. The OWAY model was used in this work only to test the assumption free ML results at flux and fluence conditions that were not in the training data set.

The domain of the standard EONY and ASTM E900 models is limited to intermediate fluences around $4 \times 10^{19}$ n cm$^{-2}$ (E > 1 MeV) that are well represented by the existing surveillance TTS database. Unfortunately, these models systematically and significantly underpredict TTS (and hardening) at higher fluence up to $10^{20}$ n cm$^{-2}$, or more, pertinent to extended life. Almost all of the higher fluence data are from accelerated test reactor irradiations over a wide range of higher flux. The largest and most comprehensive hardening database has been developed by researchers at the University of California at Santa Barbara (UCSB), as shown in the flux and fluence map in Fig. 1a. For example, the UCSB IVAR database includes a very large number of RPV steels, which include special chemically tailored split melt alloys, as well as model, surveillance and steels that had been irradiated in other programs. The split melt steel matrix included single and combined variable compositions of Cu, Ni, Mn, and P in order to characterize synergistic effects of these elements. For example, one alloy series nominally contained: (a) 0.0, 0.1, 0.2, 0.3 and 0.4 wt. % Cu at ≈ 0.8 wt. % Ni, 1.5 wt. % Mn and 0.005 wt. % P; and, (b) 0.0, 0.2, 0.8, 1.3 and 1.6 wt. % Ni at 0.0, 0.2 and 0.4 wt. % Cu. The IVAR irradiations covered a wide range of flux, fluence and irradiation temperature, and include single variable compositions differences for other elements like C. The other irradiation conditions shown in Fig. 1a generally involve a smaller number of the same alloys as those used in IVAR but extend up to high fluence at high flux. In this work we will train on what we call the IVAR+ database, which consists of the IVAR, Belgian Reactor 2 (BR2), and ATR1 databases. The IVAR+ UCSB database is a uniquely well-structured and precise, high-resolution description the single and combined (synergistic) effects on $\Delta\sigma_y$ of all embrittlement variables which are known to be important. The nominal average uncertainty in the experimentally measured $\Delta\sigma_y$ is ≈ ±20 MPa.

The major limitation of the IVAR+ database is that at lower fluxes, the fluence is limited to $<4 \times 10^{19}$ n cm$^{-2}$, while the higher fluence data is accompanied by high to ultra-high flux. Both flux and fluence are important variables and the major challenge for models being fit and/or validated on this data is to obtain accurate results when extrapolated to low flux and high fluence, service relevant, extended life conditions[6]. This work is intended to explore issues around the use of ML to predict irradiation-induced hardening in RPVs, but we stress that the focus is on how well the ML approaches can model and extrapolate from the highly structured IVAR+ database. The range of flux and fluence in the IVAR, BR2, ATR2 and ATR1 databases are given in Fig. 1a and details of both databases are described in detail in Odette, et al.[3], and Wells et al.[11]. EF for ATR1 and ATR2 condition is $1.84 \times 10^{20}$ and $5.27 \times 10^{19}$ n cm$^{-2}$, respectively. EF for BR2 condition ranges from $3.38 \times 10^{18}$ to $5.93 \times 10^{19}$ n cm$^{-2}$.

ML methods have recently been aggressively pursued as a powerful tool to predict many complex physical properties[12], including phase stabilities[13], diffusion barriers[14,15], effective charges[16], elastic constants[17], etc. Most relevantly, some recent studies have used ML to predict radiation effects in steels, with a series of studies on mechanical property changes in ferritic/martensitic (F/M) alloys for high-dose applications[18] and RPVs[19]. Note that here we do not include in the category of ML the widely used semiempirical regulatory models of RPV hardening, e.g., the EONY[6] and E900[9] models. These are based on relatively simple polynomial forms and often make use of significant physical insight, and therefore are quite different from ML approaches that use very complex functional forms and are primarily numerically driven. The body of work using ML on radiation effects in steels has recently been reviewed by Morgan et al.[20] so we will not review this literature in detail. However, a highly relevant previous RPV studies for the present work is from Matthew et al. who modeled both irradiation-induced hardening values $\Delta\sigma_y$ and TTS

($\Delta T_{41J}$), where the authors freely converted between them assuming the fairly accurate simple relationship $\Delta T_{41J} = 0.6\,°C/MPa \times \Delta\sigma_y$[19]. The features used to fit the model included elements likely to be most relevant for hardening (Cu, Ni, Mn, Si, P) and radiation conditions (flux, fluence, and temperature). Both flux and fluence were raised to the ½ power to capture expected qualitative dependence of defect production (and therefore precipitation and hardening) with flux and fluence. Like a number of the studies on radiation effects in steels, Matthew et al. observe that raising flux and fluence to the ½ power improves model performance. The ML was done with an ensemble of Bayesian neural networks and the data in both a combined surveillance and test reactor data (the U.S. NRC Embrittlement Data Base database[21], which Mathew, et al.[19] simply refer to as the NUClear REGulatory (NUREG) database) and test reactor data (part of the IVAR database used in this work). For validation data not used in their fitting the authors predicted a mean absolute error of 16 MPa and 31 MPa for IVAR and NUREG, respectively. These results are generally very encouraging, with the IVAR values being very similar to what we find in the present work for a randomly selected validation set. The reduced accuracy on NUREG data is perhaps not surprising given the less homogenous sampling in the input variables, as pointed out by the authors.

In the present work we focused on an extended IVAR database, so we treated alloys with similar processing conditions[22]. The processing differences the do occur are expected have only minor or second order effects, and therefore we did not include processing features in our feature list. As the processing conditions were treated as constants in the present study, we would not expect to accurately model alloys that underwent different processing, unless that processing happened to be unimportant for radiation response. We followed Mathew et al.[19] in using a limited set of compositions, including only on those elements that are likely to be most important. This targeted feature set should allow the most robust fitting, although is perhaps most effective within the limited domain of the model database we studied. We extended the general observation that flux and fluence should be modified to reflect their non-linear contributions (previously done using a ½ power) by using the concept of EF and find an optimized power to represent the present data. Unlike Matthew et al. we included all of the IVAR database as well as additional high fluence BR2 and ATR1 data (see below), which provided much higher fluence and flux data than all previous studies. In addition, we focused on the ability of the model to extrapolate to conditions it has not seen, exploring a wide range of leave out group tests, that leave out whole regions of the data (e.g., higher fluence values) and assessed how well the model could predict without such information. In particular, by fitting to high fluence hardening data and evaluating the prediction on ATR2 conditions not used in the fit, we were able to assess the model's capability of capturing flux effects at high fluence as a step toward modeling LWR conditions. This study is the first time such an assessment of ML's ability to capture high-fluence flux effects has been possible. Furthermore, we assessed the model using synthetic data based on highly physical simulations to gain further insight into the model's strengths and limitations. Finally, we did not use the NN based approaches taken by Matthew et al. and most previous RPV ML studies and instead use Gaussian kernel ridge regression (GKRR), a powerful model for interpolating data points that has fewer hyperparameters than NNs and can be fit more quickly.

The primary purpose of this work was to explore the use of a well-controlled database and ML models for predicting the irradiation-induced hardening effect (which can be readily related to the TTS[23]) at low-flux and high-fluence conditions to understand strengths and weaknesses of ML approaches for modeling LWR condition, especially focusing on the 60–100 years operation extension. We used the UCSB IVAR database of 49 RPV-type steels, combined with six BR2 and ATR1 alloys from a higher flux regime (i.e., higher than $10^{13}\,n\,cm^{-2}\,s^{-1}$) over a range of fluences, as shown in Fig. 1a. The data set was referred to collectively as 'IVAR+' in this study. The model generally performed quite well, accurately predicting simulated data at LWR conditions when trained on simulated data from IVAR+ conditions, and accurately predicting hardening of high-fluence intermediate-flux ATR2 irradiation conditions not used in the training data for alloys reasonably similar to the training data. However, the model does fail to capture some of the known reduction in flux effects at high fluence flux, likely due to limited sampling of these conditions. These results together suggest that the present ML approach is very promising for predicting behavior of RPV-type alloys under LWR life-extension conditions, although this is expected to require expansion of the training data to include more high-fluence and varying flux samples.

This paper used controlled data from test reactors on primarily model (albeit highly representative) alloys, rather than surveillance data from actual reactor RPV steels that had been exposed to reactor conditions. Therefore, successful modeling of the data in this work cannot be considered a definitive demonstration of effectiveness for actual reactor RPVs. Nonetheless, successes of the present model generally suggests that the approaches taken here might be applied to databases including surveillance data (and potentially model alloy and/or test reactor data) and yield useful predictions for real RPVs under actual reactor conditions.

## RESULTS

### Model assessment

In this sub-section we describe the four different types of tests we used to assess the behavior of the model in interpolating or extrapolating to different flux and fluence conditions. It is often very unclear what is interpolation vs. extrapolation in high dimensional space. We use the term extrapolation when predicting data with features that have one or more values clearly outside the range of all the model training data. We also adopt the convention of referring to predictions of ATR2 conditions as interpolation and prediction of LWR life extension as extrapolation, as the latter is farther from training conditions and still needs to be verified.

Test 1 (Basic cross-validation (CV) tests for the model): Statistical assessments with a full fit and various CV test scores were performed as an overall assessment of model performance. A full fit refers to the case where all the data is used in training the model, but errors from such a fit are not good guides to the model accuracy for prediction as the model can easily overfit the data. CV tests separate the data into subsets used for training and validation. The model is then trained on just training sets and model accuracy is assessed on validation sets.

Test 2 (LWR prediction test on IVAR+ data): In order to show how well the model can predict at the LWR conditions, Test 2 used a training data set with all IVAR+ data and made predictions at IVAR+ compositions with the LWR flux, fluence and temperature, as shown schematically in Fig. 1b. Such predictions cannot be assessed quantitatively against experimental data as no such data is available, but can be examined for signs of qualitatively unphysical behavior—e.g., hardening increasing dramatically with almost no significant fluence or hardening decreasing with increasing fluence. We can also compare the results under LWR conditions to IVAR+ measurements at equivalent EF, which should be qualitatively similar. The goal of Test 2 was to assess if the algorithm has obvious limitations in extrapolating to LWR conditions. On the other hand, in order to quantitatively assess how well the model can predict at LWR flux conditions and at high fluence, we introduced simulated hardening data by using a cluster-dynamics (CD) model to simulate the hardening effect at

both the IVAR+ and LWR conditions[24], as shown schematically in Fig. 1c. We then performed additional parallel tests to this Test 2 on the CD data and the details are discussed below.

Test 3 (Prediction on ATR2 condition): Test 3 shows the hardening prediction for a set of surveillance and select IVAR+ alloys under ATR2 conditions when the high fluence ATR1 and other high fluence BR2 data is and is not included in the ML model. Details on these alloys can be found in the Supplementary Note 1 (surveillance alloys). The ATR2 irradiations are at a higher flux and fluence than IVAR, a lower flux than BR2 and ATR1, and an EF similar to that of LWR life extension. The ability to predict hardening for these alloys at ATR2 conditions without using data from these conditions in the training data is a direct (although incomplete) assessment of the potential of the model to predict the very low-flux and high-fluence LWR conditions. This test therefore gives insight on the capability of the present ML model to capture at least some flux effects at high fluence for modeling LWR conditions.

As the actual experimental hardening data for these surveillance alloy data is still being completed and was not fully available at the time of the present research, we used the CF OWAY model-predicted hardening data for our model assessment[3]. These predictions have a root-mean-square error (RMSE) of 18.9 MPa vs. the experimental data and are therefore similar enough to the true experimental results to serve as a surrogate for assessing our ML models. Details of how to compute the CF OWAY model prediction can be found in Supplementary Note 2.

Test 4 (EONY and E900 comparison): As E900 and EONY models are commonly used RPV hardening models, it is useful to assess our predictions at LWR condition with the ones predicted by the E900 and EONY models. Test 4 will show the comparison and assess what it suggests about the different models.

Some additional CV tests leaving out different groupings of the data to assess limitations of the model were also performed (namely Test S1–S7 and their details can be found in Supplementary Note 3), which included the following:

Test S1: the EF test. This test is to explore how well a ML model trained on the IVAR+ data can be extrapolated to the low-flux and high-fluence regions where the target LWR conditions reside, where no training data is available. Since flux varied in this study by a factor of almost 1400, these differences must be accounted for. While detailed recombination models are complex, it has been shown that EF ($\phi t_e$) can be approximately defined by a scaling power $p$, as $\phi t_e = \phi t(\phi_r/\phi)^p$, where $\phi t$ is the actual fluence, $\phi$ is the actual flux, and $\phi_r$ is a reference flux taken as $\phi_r = 3 \times 10^{10}$ n cm$^{-2}$ s$^{-1}$. It physically accounts for how flux affects the excess irradiation-induced vacancies, which result in radiation enhanced diffusion (RED). By using EF, the extent of extrapolation needed to explore the low flux and high fluence regions can be reduced. The ML analysis was asked to optimize the value of a global $p$ (=0.2) (see Supplementary Fig. 1). Independent analysis has shown $p$ value for EF typically varies between ≈0.15 and 0.35 over a wide range of higher flux (or equivalent displacement per atom rates for charged particle irradiations)[6]. The optimal $p$ of 0.2 is consistent with the independent analysis shown in the literature[6]. The test suggests that introducing EF as one of the descriptors in place of separate flux and fluence produces a better model.

Test S2: leave-out (LO) alloy CV test, which evaluated the extrapolative ability of the ML method to new alloy compositions. This test suggests that if a given new alloy is nearby the ones shown in the training data, the model has more chance to accurately predict its hardening behavior.

Test S3: LO one higher EF CV test, which explored the model's ability to predict higher EF from lower EF data. This test suggests that even if we don't have a higher EF data for a given alloy, the model is still able to learn higher EF information from the nearby alloys.

Test S4: LO all higher EF CV test, which explored model predictive ability if we left our all the higher EF data for a more demanding true extrapolation. This test suggests that if the goal is to predict the higher EF data, then the training data set must include the same level or near the same level of the EF data to avoid errors from excessive extrapolation.

Test S5: LO alloy LWR prediction test, which explored how well we can predict hardening for new alloy composition at LWR conditions. This test generally showed good ability to predict new alloys although had significant degradation vs. prediction of hardening for compositions in the training data.

Test S6: Higher EF data weighting test, which explored the effect of higher EF data weighting. This test suggests that rebalancing the higher EF data weighting did not significantly improve prediction.

Test S7: Parallel tests on CD-IVAR+ data set, which were explored because the experimental LWR conditions' data at high fluence is not available and cannot be obtained practically for use in model validation. In order to quantitatively assess how well the model can predict at LWR flux conditions and at high fluence, we introduced simulated hardening data by using CD model to simulate the hardening effect at both the IVAR+ and LWR conditions. Test S7 suggests that the CD-IVAR+ data set behaves similarly to the IVAR+ dataset, and that the model has the predictive ability except for a small set of four alloys of unconventional composition.

These tests show largely expected behavior and are discussed in detail in Supplementary Note 3. We believe that overall the tests in this paper provide both useful insight into valuable features (Tests S1), mimic the typical and extreme cases that might be faced in ML applications to RPV hardening (Tests 1, 2, S2–S6), provide comparison to other similar models (Test S7, 4), and assess the model's ability to interpolate to flux and fluence at higher EF on alloys under ATR2 conditions (Test 3). Therefore, success on these tests provides strong support that ML on RPVs can provide useful insight and predictions. The following sub-sections show the results and associated discussion for Tests 1–4 for assessing the prediction at LWR conditions. Each sub-section is numbered *Test-X (NAME)*, where *X* is 1–4 and *NAME* is the test name.

### Test 1 (Basic CV tests for the model)

Test S1 found that, except for the LO alloy CV score, all the other CV scores with EF in place of flux and fluence were lower (see Supplementary Note 3). It concludes that introducing EF as one of the descriptors in place of separate flux and fluence produces a better model. With the EF feature and the other features described in Supplementary Note 1 we obtain a random fivefold CV score of 14.8 ± 0.2 MPa with $R^2$ of 0.96 (scores on other CV-test are given in Supplementary Note 3 and Supplementary Table 7). The parity plots for the full fit and fivefold CV test are shown in Supplementary Figs. 4 and 5, respectively. LO alloy group CV tests from Test S2 suggests that we are able to predict a new composition only for alloys having low predicted errors from the kernel density estimation. The cross-plot analysis shown in Supplementary Note 2 suggests that our model had some ability to capture the hardening effect across the given composition range.

### Test 2 (LWR prediction test on IVAR+ data)

In Test 2, we plotted the $\Delta\sigma_y$ vs. the EF diagram for measured and predicted hardenings for every alloy we had in the IVAR+ condition. As a representation example, consider Fig. 2a, where the blue line shows the prediction of hardening at the LWR condition as a function of EF (the LWR hardening curve), in this case for alloy LD (see selected other hardening curves in the Supplementary Fig. 9a–c, and Supplementary Fig. 16). The green triangles are the observed $\Delta\sigma_y$ in the IVAR+ database, the red

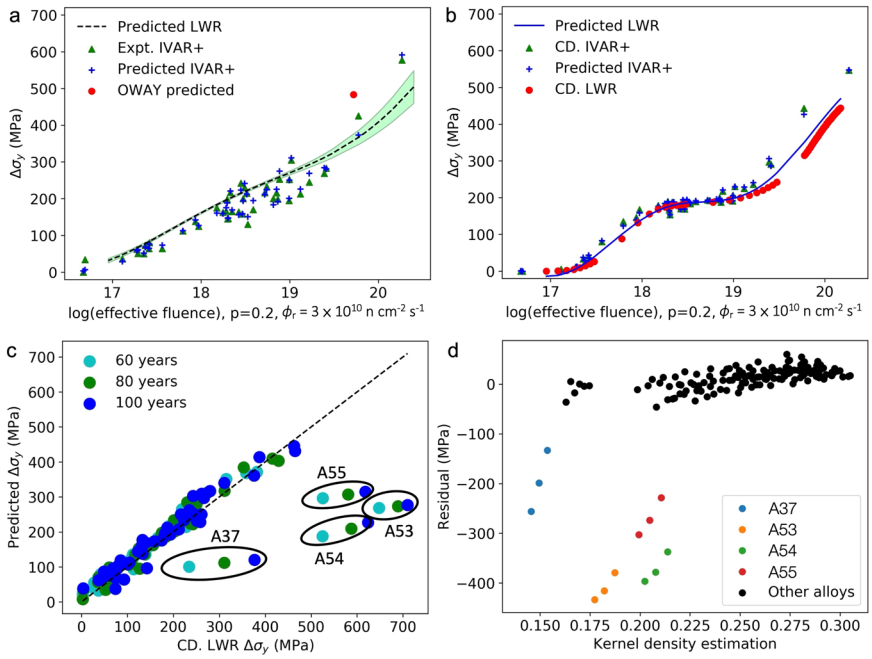


**Fig. 2 Extrapolation to LWR conditions. a** The predicted hardening curves (e.g., for alloy LD) for LWR conditions by the model fit to the IVAR+ data set. Selected additional hardening curves can be found in Supplementary Figs. 9a–c, and 16. The bands represent one standard deviation error bars predicted by the bootstrap ensemble method. **b** The predicted hardening curves (e.g., for alloy LD) by the model fit to CD-IVAR+ data set. All the hardening curves for the CD model and model predictions from fitting to CD-IVAR+ can be found in Supplementary Figs. 13, and 17. **c** Extrapolation to LWR conditions (60–100 years) of the model fit to CD-IVAR+. **d** The residual vs. kernel density estimation plot.

circle represents the prediction under ATR2 condition using the OWAY model, and the blue crosses are the corresponding prediction from our ML model, again plotted vs. EF. The standard deviation shown as the green shaded area in the Fig. 2a is estimated from a bootstrap ensemble to represent prediction uncertainty associated with the specific data used for training, which is discussed in detail in Supplementary Note 2. The prediction errors are close to the CV test error and suggest no particular sensitivity to the exact training data used. All the predicted hardening curves were positive valued, smooth, and trend upward, even when extrapolating to the high EF regime. In addition, all hardening curves showed values similar to the IVAR+ measured and predicted hardenings for the same alloy at the same EF. Thus, the results suggest no obvious unphysical behavior for prediction of behavior under LWR conditions. Specifically, Fig. 2a, b, Supplementary Fig. 16 and Supplementary Fig. 17 show the hardening curves for the six alloys (i.e., CM6, LG, LH, LI, LC, and LD) which had high fluence data (i.e., fluence $>1.7 \times 10^{19}$ n cm$^{-2}$, with a maximum value was $1.1 \times 10^{21}$ n cm$^{-2}$) (see details in Supplementary Note 4). The six alloys, which span a wide and systematic range of Cu and Ni, are broadly representative of RPV steels. Hardening increases with both Cu and Ni as predicted both by the ML and CD model. All the alloys show the expected hardening upswing in the high fluence region. In the way of a brief physical explanation, increases in yield strength are due to CRPs and MNPs[3]. CRPs quickly precipitate reaching 63% of full phase separation at fluence $<10^{19}$ n cm$^{-2}$ for IVAR conditions; the CRP volume fraction increases with Cu and subsequently plateaus below fluence $\approx 2 \times 10^{19}$ n cm$^{-2}$. In contrast, MNPs precipitate much more slowly due to lower thermodynamic chemical potential differences. MNPs reach ≈63% of full phase separation at fluence $\approx 2 \times 10^{20}$ n cm$^{-2}$. The volume fraction of MNPS is mainly controlled by Ni and they form in both Cu bearing and effectively Cu-free steels. Hardening is proportional the square root of the precipitate volume fraction. Since there is typically much more Ni + Mn + Si alloying elements compared to impurity Cu in RPV steels, MNPs can contribute a large amount of hardening at high fluence. Indeed, the ML model captured these precipitation effects even though the model was not informed by any microstructure information. The ability of the ML model to correctly predict the onset and impact of the CRPs and MNPs in many cases is actually quite remarkable. Even though we only had 42 high fluence data points in total (from the BR2 and ATR1 conditions) out of 1501 in the training data set, they played an important role in helping the ML model to capture the correct physics, especially in the high fluence region. Along with the results shown in Test S3 to S5, it is clear that the accuracy of high fluence predictions depends critically on the inclusion of high fluence data.

Although the increase in hardening at EF values a little over $10^{19}$ n cm$^{-2}$ is expected, the very strongly positive slopes of the hardening for higher EF values above $10^{20}$ n cm$^{-2}$ is likely an artifact of the model. Real alloys show a saturation of the hardening at very high EF due to solute limits imposed by phase boundaries above about an EF $> 2 \times 10^{20}$ n cm$^{-2}$, although the exact saturation EF depends on the alloy. The ML model does not capture this saturation correctly as there is far too few training data points reaching saturation for the model to correctly learn this physics sufficiently to predict the curve shape. Specifically, for an EF of $>\approx 1.66 \times 10^{19}$ n cm$^{-2}$ there are data for just six alloys and those at varied temperatures from 290 °C to 320 °C. Therefore, the model is only modestly constrained at a higher EF region. Even if the model predicted hardening value is reasonable for alloys similar to those with the maximum training EF of $1.84 \times 10^{20}$ n cm$^{-2}$, it is not constrained for values above that. However, the ML is reasonably accurate within the EF domain of values $\leq 1.84 \times 10^{20}$ n cm$^{-2}$.

Parallel tests on CD-IVAR+ data set shown in Test S7 suggest that the CD-IVAR+ data set behaved very similarly, in a semi-quantitative manner, to the IVAR+ dataset. We therefore think it is informative to assess the ML model using the CD-IVAR+ data set in ways we could not with the experimental data. We plotted the hardening curves for every alloy we had in the CD-IVAR+ condition and show an example of alloy LD in Fig. 2b (see all curves in Supplementary Fig. 13, and Supplementary Fig. 17). The red points are the CD-simulated LWR data while the blue line is the predicted one by the ML model. Most of the curves show positive valued, smooth, trends upward, and agreed with the CD-simulated LWR conditions. Except for alloys A37(L-VH), A53(M-VH), A54(L-VH), and A55(H-VH), which had a generally large error compared to the CD-simulated data at longer LWR life-extensions, all the other residuals were rather modest. The reason for the hardening underprediction of alloys A37(L-VH), A53(M-VH), A54(L-VH), and A55(H-VH) was likely different Ni and P composition in these alloys. The normal Ni composition range of RPV steels was <1.3 wt.%. The Ni composition for A37(L-VH), A53(M-VH), A54(L-VH), and A55(H-VH) were 1.70, 1.71, 1.65, and 1.66 wt.%. These four alloys were thus all high Ni alloys with compositions that fell outside the normal range of RPV steels. It is therefore expected that they would form a large volume fraction of precipitates and had very large hardening at higher fluence. The A37(L-VH) alloy also had ultrahigh P content. As a result, these alloys may show a different hardening behavior from the others. Thus, it is thus not surprising to see that the present ML model did not well capture their hardening effect. The present ML model could accurately capture the complex hardening effects, even at LWR life-extension conditions.

As 60–100 years operation at the LWR condition is of the most interest in the present study, we pulled out the prediction for this range and constructed the parity plot between the model-predicted and the CD-simulated hardening, as shown in Fig. 2c. Note that we did not have the exact 60, 80, and 100 year operation fluence data in our CD simulation due to the step size we took (see details in Supplementary Note 1). We therefore pulled the data at fluence of $6 \times 10^{19}$, $7.71 \times 10^{19}$ and $9.6 \times 10^{19}$ n cm$^{-2}$, which corresponded to 63.42, 81.39, and 101.47 years, respectively, and viewed these three data as representing the 60, 80, 100 years operation. The RMSE (MAE) ($R^2$) scores for the 60-, 80- and 100-year groups were 77.3 (35.2) (0.47), 89.3 (41.2) (0.39), and 97.4 (46.3) (0.33), respectively. Outlier alloys seem to be clearly identifiable, which are A37(L-VH), A53(M-VH), A54(L-VH), and A55 (H-VH). If we removed the outliers, the RMSE (MAE) ($R^2$) scores for the 60, 80 and 100 years data would be 21.5 (18.2) (0.97), 24.3 (20.8) (0.96), and 27.0 (23.6) (0.94), respectively.

As shown in Fig. 2d, it was likely that the four outlier alloys (A37 (L-VH), A53(M-VH), A54(L-VH), and A55(H-VH)) were not similar with the other alloys in the CD-IVAR+ training data set (i.e., mainly due to high Ni content). As a result of the unusual composition of these alloys a low hardening and large error occurred when predicting these alloys at the LWR and even just IVAR+ conditions for a wide range of EF. It is therefore clearly unreliable to predict these four alloys from the model and they were viewed as outlier cases. Of course, one must be very careful about removing outliers from one's analysis based only on their predicted errors on test data. However, in this case we have clear evidence from errors on training data and on metrics of closeness to training data that these area outlier systems, so they (or similar cases) could be readily identified before any predictions were made and are therefore appropriate to remove. Overall, the prediction to LWR conditions without the outlier cases is accurate at <27 MPa.

It should be stressed that the CD data, while having similar characteristics to actual measurements, is not equivalent to experimental data. Due to the approximations used in the CD model the relationship between the composition and irradiation conditions and hardening is likely simpler than for real alloys, which could make it easier for the ML models to fit and predict the data. In particular, the CD model does not include changing sink densities from precipitates which likely underly the reduced flux effects at high fluence, and therefore may have incorrect flux dependence at high fluence. This type of fluence dependent flux dependence is exactly what the ML model is likely to get incorrect due to limited training data, so the CD model in some ways has errors that are likely to help the ML model be more accurate. Therefore, while failure to make reasonable predictions on the CD data would have suggested the ML approach had seriously limitations, successful predictions on the CD data cannot be taken as proof that that the ML models will work on real alloy data. However, the successes observed in this study are an encouraging sign that ML models may be highly effective on similar data from real alloys.

**Test 3 (Prediction on ATR2 conditions)**

Test 3 further shows how well the hardening can be predicted for alloys under ATR2 conditions (i.e., EF of $5.27 \times 10^{19}$ n cm$^{-2}$ and flux of $3.68 \times 10^{12}$ n cm$^{-2}$ s$^{-1}$) when the model was fitted to the IVAR+ data set. To demonstrate the effectiveness on real alloy compositions we first use a set of surveillance alloy compositions for this comparison (see Supplementary Note 1 for details). Figure 3a shows results when the high fluence data (data above $1.7 \times 10^{19}$ n cm$^{-2}$ in fluence, which consisted of six alloys and 42 total data points) was not included in the training data set (i.e., using only the IVAR data set). The error bars here are the standard deviation estimated from a bootstrap ensemble. The RMSE ($R^2$) score of the fit was 85.9 MPa (0.47), indicating that the prediction was not very good. More significantly the data is severely and non-conservatively biased. These poor results are expected since the IVAR database alone does not have enough high fluence data to capture the physics at ATR2 irradiation conditions. Results from Test S3–S5 also suggest that we need to include the same level or near the same level of the EF data in the training data set if we would like to make accurate high-fluence predictions. Figure 3b shows the same predictions when we included the high fluence BR2 and ATR1 data back into the training data set (i.e., using the full IVAR+ data set). The RMSE ($R^2$) score of the fit was 37.2 MPa (0.71), which is reasonably good although there are some significant errors. However, much of the poorly predicted data has training data compositions and/or irradiation conditions far from the training data set, as can be seen in Fig. 3c, where we used a kernel density estimation distance metric on features to assess if alloys are at risk of being outside the models' domain. If we excluded these risky alloys with a single distance cutoff, the RMSE ($R^2$) score is 25.9 MPa (0.89), which is a quite small error, approaching that of the CF model itself (i.e., 18.9 MPa)[3]. The results also show a much more modest underprediction compared to the IVAR based model, with a mean error of −20 MPa. The origin of this underprediction is not totally clear, but is likely due to a well-established decrease in the flux effect at the ATR2 high fluence conditions, that is not fully captured by the present model with a constant flux scaling $p$, since it is trained on primarily lower fluence data. Future studies, with larger data sets, will explore flux-fluence interactions, which can be expressed as $p = f(\text{fluence})$. Just using flux and fluence and avoiding EF all together could avoid this issue, but as shown in Test S1 (see Supplementary Note 3) this choice yields less accurate models on the present IVAR+ database. Preliminary explorations of fluence dependent $p$ values with the present training data also yielded less accurate models. In both cases the failure to obtain the proper flux scaling at very high fluence is likely due to the limited training data with varied high flux and a fluence. Another possible reason for the bias in the ATR2 predictions is that our data is biased to lower fluences. Therefore, we performed Test S6 to rebalance the training data to have equal weights to low and high fluence data. However, this balancing did not turn out to improve the predictions. We further compared with

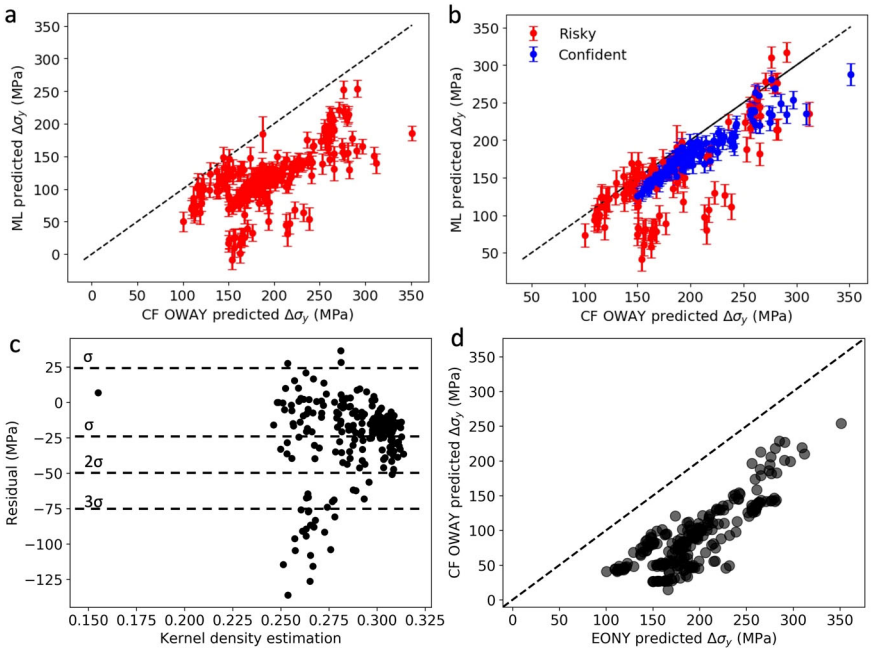


**Fig. 3 High fluence data effect on hardening prediction. a** The parity plot of hardening prediction to surveillance alloys under ATR2 conditions when not fitting to high fluence data (RMSE ($R^2$) = 85.9 (0.47)). **b** The parity plot of hardening prediction to surveillance alloys under ATR2 conditions when fitting to high fluence data (RMSE ($R^2$) = 37.2 (0.71)). **c** Residual plot vs. kernel density estimation when machine learning model was fitted to high fluence data. **d** EONY model prediction to surveillance alloys under ATR2 conditions vs. CF OWAY model predictions (RMSE ($R^2$) = 107.9 (0.57)). The error bars were the standard deviation predicted by the bootstrap ensemble method.

prevailing EONY model (based on lower fluence surveillance data) prediction to these surveillance alloys and conditions, as shown in Fig. 3d. The RMSE ($R^2$) score is 107.9 (0.57), coupled with a very large unconservative bias. Clearly our ML model has a much better predictive ability than the EONY model.

In additional to the tests on surveillance alloy compositions done above, we also predict behavior under ATR2 conditions for CM6, LG, LH, LI, LC, and LD, which are the six alloys for which the most high fluence data is available for training (i.e., these are the alloys with BR2 and ATR1 high fluence data). These alloys are expected to perform particularly well under ATR2 conditions because we have high fluence training data and they are also at compositions similar to other training data, allowing for robust interpolation from the training data. Figure 4 shows the parity plot, demonstrating a very good agreement between the OWAY model predicted hardening and that predicted by the ML model, with an RMSE of 18.9 MPa, essentially identical to the error of OWAY itself of 18.9 MPa[3].

In Fig. 5 we include the complete hardening vs. fluence behavior for these same six alloys, including the full set of training data (IVAR+), model fits to the training data (ML predicted IVAR+), model predicted behavior vs. EF under ATR2 flux and temperature conditions (ML predicted ATR2), and the OWAY predictions under ATR2 conditions (OWAY predicted ATR2). The agreement of the dashed lines and red points show the good agreement of the model predictions with the OWAY models, providing an equivalent but different view on the results shown in Fig. 4. Figure 5 also shows how the ML model predictions under ATR2 flux differ from the training data values, which cover a range of flux values. In particular, for the very high EF values from ATR1 experiments (EF of $1.84 \times 10^{20}$ n cm$^{-2}$) we see that the hardening values increase rapidly due to the precipitation of MNPs. However, for the same EF but at the lower ATR2 flux, the model predicts a much lower hardening for these alloys. A similar difference in hardenings for the same alloys at the same EF is seen in Supplementary Fig. 16, which is equivalent to Fig. 5 but shows the ML model predictions for LWR flux in place of ATR2 flux. The hardening difference at the same EF is disconcerting since different flux-fluence combinations are expected to give similar hardening at the same EF when the latter is properly determined. The hardening discrepancy is likely due to the issue already mentioned above in this sub-section, which is that the $p$ value treatment in this work is oversimplified. Specifically, the EF definition used in this work uses a single constant effective $p$ scaling of 0.2 to capture flux effects. However, the $p$ value should move toward zero for higher fluence (like ATR1 conditions)[3]. If a more accurate EF were defined and used for the $x$-axis in Fig. 5 then a major effect would be that the ATR1 data would move to the right relative to the model predictions. As a qualitative estimate of the impact of this effect one can consider holding the ATR2 flux predictions from the model constant, since $p$ is still not too far from 0.2 at this flux, and then moving the ATR1 data on Fig. 5 to a new EF value of $1.1 \times 10^{21}$ n cm$^{-2}$, which is the actual fluence for ATR1 and therefore the EF for $p = 0$. Such a shift would bring the ML model predictions and experimental measurements into fairly good agreement for all the alloys. More work is needed to integrate fluence dependent $p$ values into the ML model.

These test results show relatively small errors when predicting hardening at higher fluence lower flux conditions of ATR2 when the fitting includes a range of higher fluence and flux data when the domain of the model is carefully constrained (which can be done by a simple test prior to prediction for new data points). These tests also show the significant limitations of models like EONY on the same kind of test. However, the results do suggest issues with the model, particularly illustrating the limitations of the present model's ability to capture flux effects with an EF defined by a fixed $p$ of 0.2. Therefore, overall these test results support the conclusion that a ML model can yield useful extrapolation to lower flux conditions at high fluence, and that the results are likely far superior to the widely used regulatory models like EONY.

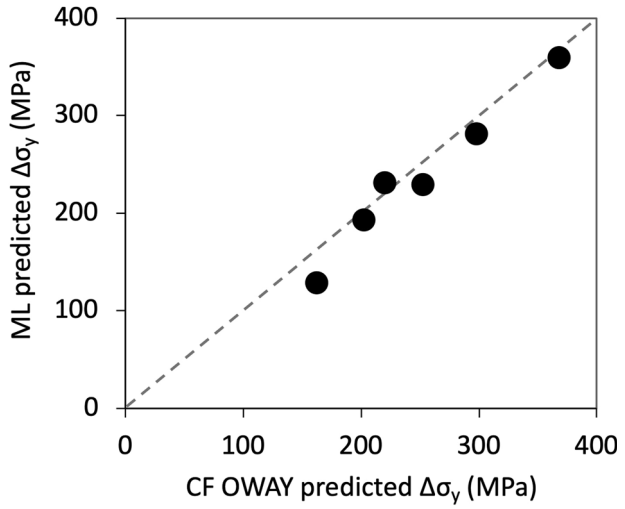


**Fig. 4 Hardening prediction under ATR2 conditions.** The parity plot of hardening prediction to alloy CM6, LG, LH, LI, LC, and LD under ATR2 conditions. The RMSE = 18.9 MPa, and $R^2 = 0.96$.

However, the models need further refinement with additional data and more flexible EF $p$ scaling to properly capture flux effects. Such ML models have the potential to be a tremendous compliment to more physically-based approaches.

### Test 4 (EONY and E900 comparison)

In Test 4, the prediction for the yield strength after irradiation of 80 years at the LWR flux and temperature was made using the present GKRR model fit to the IVAR+ dataset. The prediction was then compared with the ones made by E900 and EONY model. As shown in Fig. 6, our predicted values seem to be lying between these two models. Supplementary Figure 18 plots the differences E900-GKRR and EONY-GKRR. We note here that a large disagreement exists between the E900 model and the EONY model for the alloys of A36(L-VH), A37(L-VH), A53(M-VH), A54(L-VH), and A55(H-VH). The discrepancies for these alloys are likely due to the fact that they are high Ni alloys and E900 and EONY were fit to data on alloys without comparably high Ni. In general, we are above EONY, which is expected given that the EONY model is expected not to include some precipitation from MNPs that was not generally present in the training data used to construct EONY. However, our predicted values are significantly lower than EONY for some alloys, specifically A13(H-M), A22(H-M), and A50(H-M). The predicted hardening curves for these alloys by the GKRR model shown in Supplementary Fig. 9a-c (fits to IVAR+ experimental data) and Supplementary Fig. 13 (fits to CD-simulated data) show generally reasonable behavior, and the source of this inverted ordering is not clear. However, as with some of the alloys discussed above, the unusual behavior is likely due at least in part to these alloys being outside the domain of the EONY model. In particular, they have high Cu (A13(H-M) and A50(H-M)) or low Mn (A22(H-M)). It was also shown in our model that it seems risky to predict hardening for these alloys because they were too far away from other alloys. Nevertheless, in general our model agreed well with present physically-based models despite the fact that the prediction was only informed by the composition and the irradiation conditions of a given alloy.

## DISCUSSION

The present ML model fit to the well-controlled IVAR+ database demonstrated the potential benefits and risks of using ML model to predict the complex compositional and flux dependence of hardening in RPV alloys, including extrapolations to LWR extended life conditions. We summarize key aspects and results of the present work here:

1. EF as a useful input feature: The EF used here is a simple approximation that has proven successful in fitting data over a wide range of flux. It is generally consistent with a model for the flux dependent effect of recombination on RED. This work explored EF as a feature for ML and demonstrated that it is effective and that a subtle physical quantity like the $p$-factor scaling of the EF can be extracted from ML fitting. However, we also found that the known reduction in $p$ factor with high fluence could not be extracted from the present training data, demonstrating the need for a modified training database to capture this important physics. A more accurate treatment of flux effects in RPV ML hardening models is the subject of ongoing research.
2. Experimental data with high fluence data: The present work used significantly more data and data at much higher fluence and EF than any previous ML RPV study. It should be stressed that even though these high fluence data points were small in number, their existence in the training data set was indispensably important in directing the model to capture the correct high-fluence physics. This work was therefore able demonstrate the potential efficacy of ML models at EF values consistent with LWR life-extension conditions.
3. Synthetic data for directly assessing LWR conditions: This work used a physically informed hardening model to create synthetic data across all relevant conditions, in particular including LWR conditions, and then use fits to that synthetic data to demonstrate the ability of the ML approaches to extrapolate to LWR conditions, as described further below.
4. Successful prediction of extrapolated and interpolated conditions: The model showed accurate prediction of LWR conditions on synthetic data for all but four rather unusual outlier alloys, without which the model yields an RMSE of 27.0 MPa at about 100 years of simulated LWR life-extension (Test 2). This represents accurate extrapolation capabilities to high fluence and low flux, albeit only on synthetic data that is expected to have simpler flux dependence at high fluence that the true experiments. The model also yielded accurate prediction of hardening on surveillance alloys under the new ATR2 irradiation conditions, with an RMSE of 23.3 MPa, suggesting strong interpolation capabilities, and potentially good extrapolation capabilities, to new flux and fluence conditions (Test 3).
5. Independent of preconceived physics, and corresponding models, this ML study confirmed the highly non-conservative predictions of an existing embrittlement model (EONY) for high extended life fluence conditions with a reasonable adjustment for flux effects ($p \approx 0.2$). Reliable embrittlement predictions are critical to the safe operation of the worldwide fleet of nuclear reactors.

While the model is overall quite promising, it does show significant discrepancies in many cases. There are some likely sources for these discrepancies and we here enumerate them along with some suggestions for model improvements.

- The data is poorly sampled at higher fluence and more modest flux, which makes extrapolation to low-flux and high-fluence particularly challenging. Including of data from the recently completed ATR2 irradiations in the training data (in this work it was only used for testing) will be a significant step in resolving this issue.
- Some compositions are poorly sampled and may have very different physics that the other alloys in the database. For

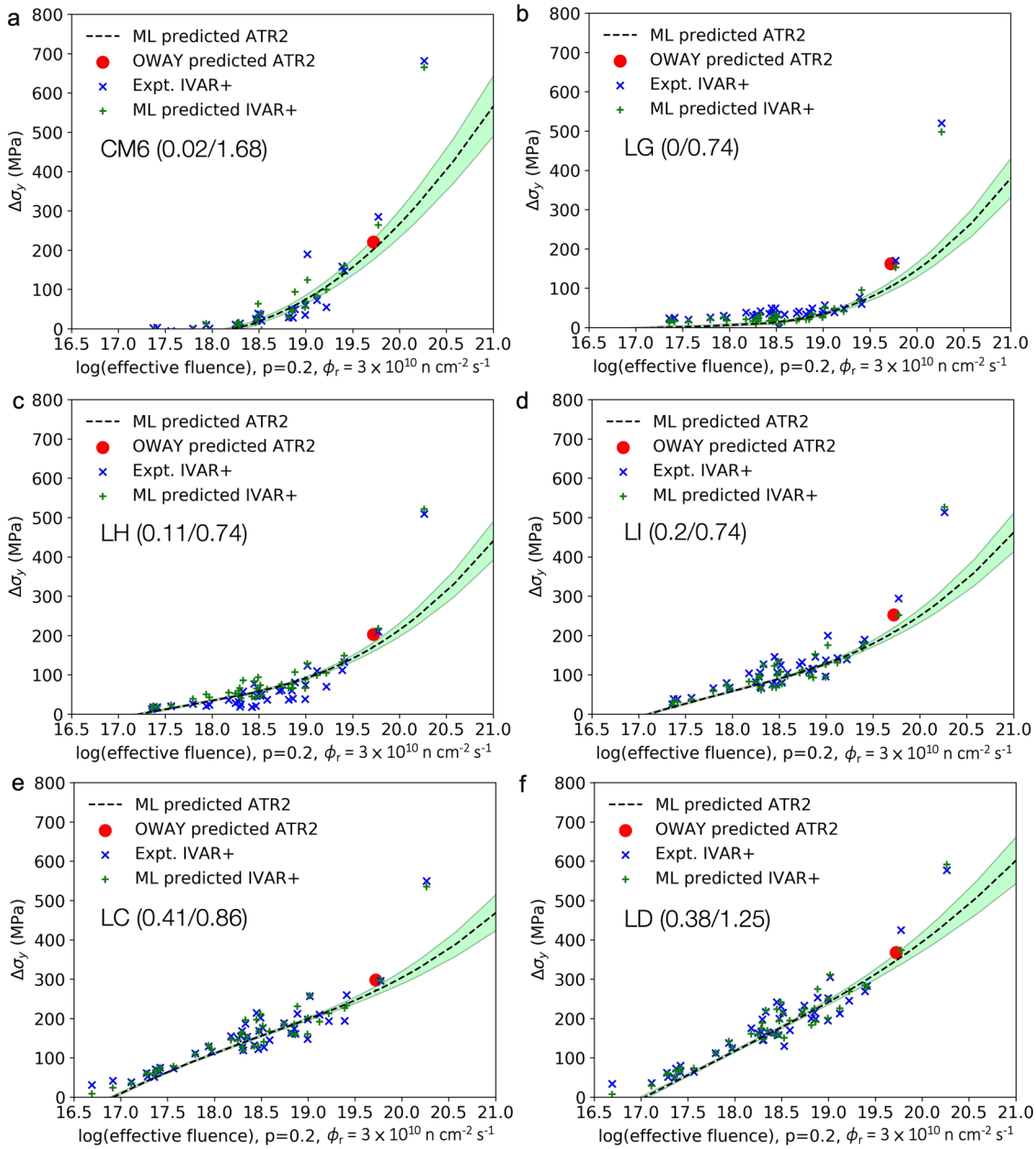


**Fig. 5 Hardening curves at ATR2 conditions.** The full fit vs. effective fluence (hardening curve) for ATR2 conditions with an optimized *p* value for the optimized GKRR model fit on the IVAR+. Alloys referred in the plots were (**a**) CM6, (**b**) LG, (**c**) LH, (**d**) LI, (**e**) LC and (**f**) LD. The ratio following the alloy name represent the Cu over Ni content (e.g., CM6 (0.02/1.68) means Cu and Ni content of CM6 were 0.02 and 1.68 in wt.%, respectively) (see Supplementary Note 4). The bands represent one standard deviation error bars predicted by the bootstrap ensemble method.

example, high Ni alloys represent only a small fraction of the alloys in the database and could form different precipitates and therefore have quite different hardening behavior compared to lower Ni alloys. One might consider either adding more such data or developing models without these alloys that use a more restricted training data and have more restricted domains of applicability. One could even envision using the Fe-Cu-Mn-Ni-Si phase diagram to determine where the precipitation physics is fairly consistent and restricting the model to this domain of compositions.

- The present ML model uses a simple EF feature that uses a single constant *p* value, which is known to be only an approximate treatment of flux effects. Using multiple *p* values, fluence dependent *p* values, or even just flux and fluence when combined with more fitting data would potentially yield more robust ML models.
- No processing data or information is used in the features of the present ML model. While this is a reasonable approximation for the IVAR+ database due to efforts to keep the processing consistent, some improvements are likely possible by including processing-based features. Furthermore, for other alloys with more processing variation a set of processing features might be essential. One example of a processing based insight that could be useful is that Cu concentrations in the alloy are generally limited to ≈0.25 at.% due to pre-precipitation of any supersaturation above this value

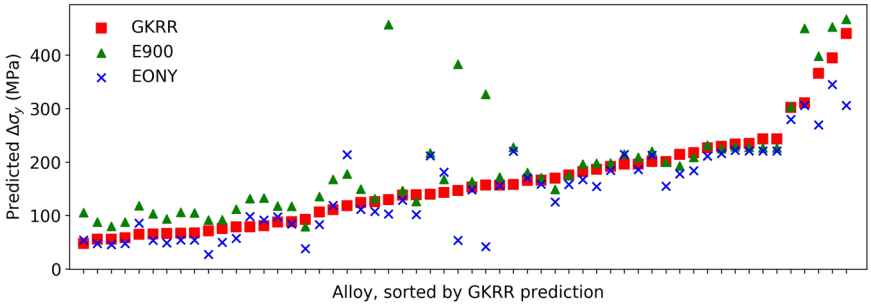


**Fig. 6 LWR life-extension hardening prediction.** LWR life-extension at 80 years (fluence of $7.71 \times 10^{19}$ n cm$^{-2}$) fit to IVAR+ data with a reference flux of $3 \times 10^{10}$ n cm$^{-2}$ s$^{-1}$ at T = 290 °C (comparison with E900 and EONY models).

during the steel processing.

- Some simple hardening physics is not enforced, including that hardening vs. fluence has a non-negative slope and that is saturates to zero slope at very high fluence. Enforcing these constraints may be difficult due to algorithmic challenges of their application and due to uncertainties in the exact values (e.g., the saturation fluence varies with alloy and flux), but would likely constrain the model further.
- No thermodynamic data is presently used in the model, making it easy to violate basic thermodynamic trends. Including thermodynamic information, e.g., by adding in triple products[3] to the feature list, could enhance the model accuracy.

Despite these many possible improvements, we believe that model developed shows the ability to predict hardening at new flux and fluence conditions in ways critical to possible future application of ML to commercial RPV alloys in LWRs.

## METHODS

Here we briefly describe methods of the work. The ML model used in this work is GKRR. More details of the GKRR model description, the data sets and the features can be found in the Supplementary Note 1 and details on hyperparameter optimization can be found in Supplementary Note 2. The model analysis and exploration was primarily performed with the MAterials Simulation Toolkit for Machine Learning (MAST-ML)[25], an open-source python package with scikit-learn[26] library to automate ML workflows and model assessment. Data sets used in this study consisted of both experimental (i.e., IVAR+) and CD-simulated (i.e., CD-IVAR+) $\Delta\sigma_y$ under irradiation. To prevent confusion, the IVAR+ data set mentioned throughout this paper refers to the experimental hardening data for IVAR+ alloy compositions and irradiation conditions, while the CD-IVAR+ data set refers to the same set of alloy compositions and irradiation conditions but includes hardening values from only the cluster-dynamics model and none from experiment. The present ML model used the elemental compositions of Cu, Ni, Mn, Si, P, and C, irradiation temperature, fluence, and EF as input features with $\Delta\sigma_y$ in MPa as the response. We added synthetic data of hardening equal to 0 at fluence of $6.0 \times 10^{16}$ n cm$^{-2}$, flux of $1.0 \times 10^{11}$ n cm$^{-2}$ s$^{-1}$, and temperature of 290 °C to the data sets for all alloys to help the model yield the expected zero hardening at low EF condition. The total number of data points was 1501. 55 alloys were included in the data set and were named throughout this study by using A1, A2, A3… to A55. Specific names denoting composition, and/or compositions, are not generally given due to the proprietary nature of the database. However, given their importance in this work, we specifically name and give compositional information on six alloys that were studied to high fluence, namely A36-CM6, A44-LC, A45-LD, A46-LG, A47-LH, and A48-LI. Other alloy compositions are denoted approximately as X–Y, where X and Y represent Cu and Ni content at different content level, respectively. We choose Cu and Ni as they are typically the elemental compositions dominating the hardening behavior. For Cu, X would be L: Cu ≤ 0.08 wt. %; M: 0.08 < Cu ≤ 0.24 wt. %; H: Cu > 0.24 wt. %. For Ni, Y would be L: Ni ≤ 0.5 wt. %; M: 0.5 < Ni ≤ 0.9 wt. %; H: 0.9 < Ni ≤ 1.3 wt. %; VH: Ni > 1.3 wt. %. For instance, A1(M-M) means alloy Al has Cu content of 0.08 < Cu ≤ 0.24 wt. %, and Ni content of 0.5 < Ni ≤ 0.9 wt. %. Supplementary Table 1 tabulates the Cu and Ni content level for each alloy. The hyperparameters of the GKRR model were optimized by using grid search method with a custom cost function taking the LO-multiple-group CV average RMSE as the scoring metric. The bootstrap method to generate and ensemble of models was used in predicting a standard deviation for LWR predictions, as well as all other predictions. Kernel density estimation was used in quantifying how similar a given alloy is with the training data set[27,28]. Cross-plot analysis was used in assessing the performance of the model at the given composition space. The GKRR model was assessed with CV tests, including fivefold CV, and LO-group CV, and comparisons with the CF OWAY[3], EONY[6] and E900[9] models. Details of the grid search, CV methods, cross-plot analysis, bootstrap method, kernel density estimation, the CF OWAY[3], EONY[6] and the E900[9] models can be found in Supplementary Note 2.

## DATA AVAILABILITY

To ensure all publicly available data used in this paper are easily accessible and adequately archived, we have placed the following files in the Supplementary Information and on Figshare with https://doi.org/10.6084/m9.figshare.12816437. The IVAR+ database is not publicly available and therefore this data is not included in any of the shared files. Requests for the IVAR+ data should be sent to G.R.O. at odette@engineering.ucsb.edu. The databases used in this study are still under development. The part analyzed here is available upon request to evaluate the correctness of our results but based on an agreement that there would not be further dissemination.

1. Figures Data: Fig X.csv and Fig SX.csv contain all the data used to make Figure X and Supplementary Fig X in the paper and the Supplementary Information, respectively, except for the data that may directly revealed the experimentally determined $\Delta\sigma_y$ of a given alloy composition.
2. Tables Data: Table X.csv and Table SX.csv contain all the data used to make Table X and Supplementary Table X in the paper and the Supplementary Information, respectively.
3. Model parameters: Model_coef_X.csv and Model_kernel_X.csv contain the $\beta$ coefficient and the $K$ Gaussian kernel matrix that we used in creating the model at the full-fit prediction, where X is CD or Expt, accounting for using CD-IVAR+ and IVAR + as the training data set.

## CODE AVAILABILITY

The main code for MAST-ML software used in the present work is available at https://github.com/uw-cmg/MAST-ML. The code for the custom cost function taking the leave-out (LO)-multiple-group CV average RMSE as the scoring metric for hyperparameter optimization is available at https://github.com/yuchenliu19/dbtt-npj.

Received: 24 August 2020; Accepted: 23 March 2022;
Published online: 27 April 2022

## REFERENCES

1. Administration, U. S. E. I. *U.S. Nuclear Industry - Energy Explained, Your Guide To Understanding Energy*, http://www.eia.gov/energyexplained/index.cfm?page=nuclear_use (2016).
2. Administration, U. S. E. I. *How old are U.S. nuclear power plants, and when was the last one built?*, http://www.eia.gov/tools/faqs/faq.cfm?id=228&t=21 (2016).

3. Odette, G. R. et al. On the history and status of reactor pressure vessel steel ductile to brittle transition temperature shift prediction models. *J. Nucl. Mater.* **526**, 151863 (2019).
4. Nanstad, R. K. & Server, W. L. Reactor Pressure Vessel Task of Light Water Reactor Sustainability Program: Initial Assessment of Thermal Annealing Needs and Challenges. Report No. ORNL/LTR-2011/351, https://www.energy.gov/ne/articles/reactorpressure-vessel-task-light-water-reactor-sustainability-program-initial (Oak Ridge, TN, 2011).
5. News, W. N. *Rosatom launches annealing technology for VVER-1000 units*, https://www.world-nuclear-news.org/Articles/Rosatom-launches-annealing-technology-for-VVER-100 (2018).
6. Eason, E. D., Odette, G. R., Nanstad, R. K. & Yamamoto, T. A physically-based correlation of irradiation-induced transition temperature shifts for RPV steels. *J. Nucl. Mater.* **433**, 240–254 (2013).
7. American Society for Testing and Materials International (ASTM) standard E185-16, *Standard Practice for Design of Surveillance Programs for Light-Water Moderated Nuclear Power Reactor Vessels*, in ASTM International, West Conshohocken, PA. vol. 12.02, p. 9 https://doi.org/10.1520/E0185-15 (2015).
8. Odette, G. & Lucas, G. Embrittlement of nuclear reactor pressure vessels. *JOM* **53**, 18–22 (2001).
9. American Society for Testing and Materials International (ASTM) standard E900-15, *Standard Guide for Predicting Radiation-Induced Transition Temperature Shift in Reactor Vessel Materials*, in ASTM International, West Conshohocken, PA. vol. 12.02, p.4 https://doi.org/10.1520/E0900-15 (2017).
10. Eason, E. D., Wright, J. E. & Odette, G. R. Improved Embrittlement Correlations for Reactor Pressure Vessel Steels, https://books.google.com.tw/books?id=DQOqNAAACAAJ (1998).
11. Wells, P. B. et al. Evolution of manganese–nickel–silicon-dominated phases in highly irradiated reactor pressure vessel steels. *Acta Mater.* **80**, 205–219 (2014).
12. Morgan, D. & Jacobs, R. Opportunities and challenges for machine learning in materials science. *Annu. Rev. Mater. Sci.* **50**, 71–103 (2020).
13. Li, W., Jacobs, R. & Morgan, D. Predicting the thermodynamic stability of perovskite oxides using machine learning models. *Comput. Mater. Sci.* **150**, 454–463 (2018).
14. Wu, H. et al. Robust FCC solute diffusion predictions from ab-initio machine learning methods. *Comput. Mater. Sci.* **134**, 160–165 (2017).
15. Lu, H.-J. et al. Error assessment and optimal cross-validation approaches in machine learning applied to impurity diffusion. *Comput. Mater. Sci.* **169**, 109075 (2019).
16. Liu, Y.-c et al. Exploring effective charge in electromigration using machine learning. *MRS Commun.* **9**, 567–575 (2019).
17. De Jong, M. et al. A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds. *Sci. Rep.* **6**, 34256 (2016).
18. Long, S. & Zhao, M. Theoretical study of GDM-SA-SVR algorithm on RAFM steel. *Artif. Intell. Rev.* **53**, 4601–4623 (2020).
19. Mathew, J. et al. Reactor pressure vessel embrittlement: Insights from neural network modelling. *J. Nucl. Mater.* **502**, 311–322 (2018).
20. Morgan, D. et al. Machine learning in nuclear materials research. *Curr. Opin. Solid State Mater. Sci.* **26**, 100975 (2022).
21. Takamizawa, H., Itoh, H. & Nishiyama, Y. Statistical analysis using the Bayesian nonparametric method for irradiation embrittlement of reactor pressure vessels. *J. Nucl. Mater.* **479**, 533–541 (2016).
22. Odette, G. et al. *Effects of Composition and Heat Treatment on Hardening and Embrittlement of Reactor Pressure Vessel Steels*. (Office of Nuclear Regulatory Research, U.S. Nuclear Regulatory Commission, 2003).
23. Odette, G. R. & Nanstad, R. K. Predictive reactor pressure vessel steel irradiation embrittlement models: issues and opportunities. *JOM* **61**, 17–23 (2009).
24. Mamivand, M. et al. CuMnNiSi precipitate evolution in irradiated reactor pressure vessel steels: Integrated Cluster Dynamics and experiments. *Acta Mater.* **180**, 199–217 (2019).
25. Jacobs, R. et al. The Materials Simulation Toolkit for Machine learning (MAST-ML): An automated open source toolkit to accelerate data-driven materials research. *Comput. Mater. Sci.* **176**, 109544 (2020).
26. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
27. Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **33**, 1065–1076. https://doi.org/10.1214/aoms/1177704472 (1962).
28. Agarwal, N. & Aluru, N. R. A data-driven stochastic collocation approach for uncertainty quantification in MEMS. *Int. J. Numer. Methods Eng.* **83**, 575–597, https://doi.org/10.1002/nme.2844 (2010).

## ACKNOWLEDGEMENTS

D.M., H.W., R.J., and T.M. gratefully acknowledge partial funding from NSF SI2-SSI award 1148011, the Light Water Reactor Sustainability program, and Nuclear Energy University Program (NEUP) 21-24382. Y.-c.L. gratefully acknowledge the financial support from Graduate Student Study Abroad Program (GSSAP) (107-2917-I-006-008), project (110-2222-E-006-008) from the Ministry of Science and Technology (MOST), and the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) and MOST (110-2634-F-006-017) in Taiwan. The UCSB researchers gratefully acknowledge that the IVAR database development was sponsored by the US Nuclear Regulatory Commission; while the US Department of Energy (DOE) Nuclear Scientific Users Program supported the ATR 1 and 2 irradiations. PIE on ATR-1 and 2 was partially supported by the DOE Light Water Reactor Sustainability Program.

## AUTHOR CONTRIBUTIONS

Y.-c.L. performed the bulk of the final analysis and wrote most of the text. G.R.O. and his group independently developed all the databases used in this work over many decades and shared them with the other authors on this paper to help support this work. G.R.O. and collaborators also independently developed the OWAY model and shared its details with the other authors on this paper to help support this work. G.R.O. and D.M. conceived of the work and guided the project. They also heavily edited and contributed significant text to the final paper. P.W., N.A., and T.Y. worked on aspects of the experimental database used in this work. H.W., T.M., B.A., R.J., J.P., J.G., J.C., J.X., H.Y., A.L., H.W., M.P., F.D., A.P., and L.X. performed early versions of the analysis. T.M. and B.A. revised the analysis wrote an early version of the paper. All authors reviewed the complete paper.

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41524-022-00760-4.

**Correspondence** and requests for materials should be addressed to Dane Morgan.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.