

ARTICLE OPEN



Decoding defect statistics from diffractograms via machine learning

Cody Kunka¹, Apaar Shanker², Elton Y. Chen¹, Surya R. Kalidindi¹ and Rémi Dingreville¹✉

Diffraction techniques can powerfully and nondestructively probe materials while maintaining high resolution in both space and time. Unfortunately, these characterizations have been limited and sometimes even erroneous due to the difficulty of decoding the desired material information from features of the diffractograms. Currently, these features are identified non-comprehensively via human intuition, so the resulting models can only predict a subset of the available structural information. In the present work we show (i) how to compute machine-identified features that fully summarize a diffractogram and (ii) how to employ machine learning to reliably connect these features to an expanded set of structural statistics. To exemplify this framework, we assessed virtual electron diffractograms generated from atomistic simulations of irradiated copper. When based on machine-identified features rather than human-identified features, our machine-learning model not only predicted one-point statistics (i.e. density) but also a two-point statistic (i.e. spatial distribution) of the defect population. Hence, this work demonstrates that machine-learning models that input machine-identified features significantly advance the state of the art for accurately and robustly decoding diffractograms.

npj Computational Materials (2021)7:67; <https://doi.org/10.1038/s41524-021-00539-z>

INTRODUCTION

Diffraction techniques can probe large volumes of material while maintaining high resolution in both space and time^{1–4}. Hence, these techniques are widely used to provide structural characterizations across a variety of scientific fields, including biology^{5–7}, materials science^{1,8,9}, and polymer physics^{10,11}. However, the difficulty of decoding diffractograms has greatly limited their utility¹². Challenges concern both steps of the decoding process: (i) identifying the key features in the diffractogram and then (ii) modeling their relationships to the desired structural characterizations. To identify key features, most studies employ a common set of “human-identified features” (HIFs). Assessments of 2D diffraction patterns, such as from selected area electron diffraction (SAED), have focused on the positions, areas, and shapes of the spots. Likewise, assessments of 1D diffraction line profiles, such as from conventional X-ray diffraction (XRD), have focused on the positions and widths of the peaks. Unfortunately, neither the 1D nor the 2D HIFs was developed systematically, and they cannot comprehensively summarize a diffractogram. Hence, models that input HIFs can only predict a limited set of structural characterizations. As a result, models are numerous, highly empirical, and often conflicting. For example, we recently demonstrated that fitting two popular width models with the same XRD data yielded opposite trends in characteristic size¹³.

Because of these difficulties, researchers have often supplemented their limited analysis of diffractograms with results from other techniques, such as transmission electron microscopy^{14–16}, thermal measurements^{17–19}, ultraviolet-visible absorption^{20,21}, and atomistic simulations^{22–26}. However, if a reliable and complete decoding strategy were available, these researchers could have efficiently obtained the desired structural information from the diffractograms alone. Machine-learning techniques could resolve this problem, working on the assumption that the ability to input a large number of features would more robustly link comprehensive diffractogram features to an improved set of characterizations^{27–42}. For instance, several research groups recently employed

convolutional neural networks to reconstruct images from coherent XRD^{37,42}. Likewise, researchers recently analyzed XRD^{27,36,39} and SAED^{31,38} via machine-learning techniques in order to identify crystallography and phase transformation. In the current study, we focus on the ability of machine-learning models to interpret defect statistics from diffraction data.

To enable complete decoding, the current work systematically demonstrates (i) how to compute machine-identified features (MIFs) that fully summarize a 2D diffraction pattern and (ii) how to employ machine learning to reliably connect these features to an expanded set of structural characterizations. For instructional purposes, we examined virtual SAED patterns produced from atomistic simulations of irradiated copper. This simulation-based approach bypassed experimental complications, such as dynamic diffraction conditions, enabling a 1:1 comparison between diffractogram features and defect statistics. We chose irradiation damage as an exemplar because it provided a “defect laboratory” that offered numerous defect statistics to be related to the diffractogram features. We chose SAED specifically because it exemplified 2D diffraction patterns, which contain more information than 1D diffraction line profiles (as is common in conventional XRD). However, 2D diffraction patterns could also be constructed for XRD⁴, so our findings would apply to that as well.

As summarized in Fig. 1, we first characterized defect statistics of the irradiated microstructures to serve as evaluation metrics (i.e. outputs) for our models. We computed one-point statistics (i.e. point-defect density and dislocation density) as is common in diffraction studies. We also computed a two-point statistic (i.e. the pair-correlation function (PCF) of the point-defect distribution) to demonstrate that diffractogram decoding could be extended. A two-point statistic effectively captures all of the information of the corresponding one-point statistic while adding higher-order information. Our particular two-point statistic captured both the density and the spatial distribution of the point defects. To produce the inputs for our models, we characterized the virtual SAED patterns in two ways. For HIFs, we tracked the area,

¹Center for Integrated Nanotechnologies, Sandia National Laboratories, Albuquerque, NM, USA. ²Georgia Institute of Technology, Atlanta, GA, USA. ✉email: rdingre@sandia.gov

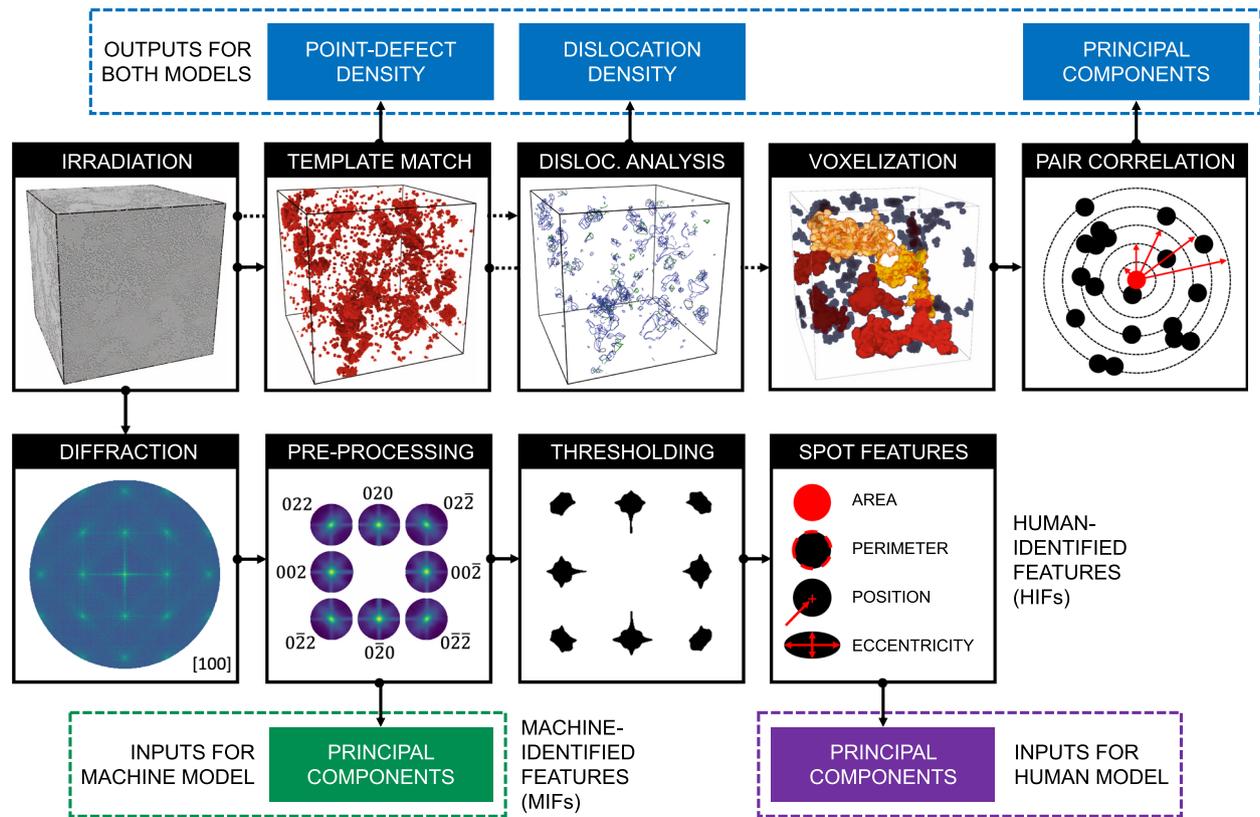


Fig. 1 Overview of the process for producing the inputs (i.e. SAED features) and the outputs (i.e. defect statistics) for the two machine-learning models (i.e. the human model and the machine model). Atomistic representations visualized via OVITO⁶³. SAED pattern visualized via ParaView⁶⁴.

perimeter, eccentricity, and position for each family of diffracted spots. For MIFs, we simply computed the principal components of the SAED patterns. Finally, we used Gaussian Process Regression to construct a “human model” from the HIFs and a “machine model” from the MIFs. Both machine-learning models captured the one-point statistics, but only the machine model (which used MIFs to characterize the virtual SAED patterns) captured the two-point statistic. Therefore, our results demonstrate how to decode advanced structural information unavailable to conventional diffractogram decoding.

RESULTS

This study comprised three major parts: (i) simulation of irradiated microstructures to provide defect statistics, (ii) simulation of SAED patterns to provide HIFs and MIFs, and (iii) construction of machine-learning models to decode defect statistics from the HIFs and MIFs. The Methods section provides further technical details on each step.

Choosing defect statistics

To understand the defect statistics generated from our atomistic simulations, we first examined the defect mechanisms as a function of the number of irradiation events, which we iteratively introduced over time. We observed two primary stages in the evolution of point defects (e.g., Fig. 2a, b) and dislocations (e.g., Fig. 2c, d). In the first stage (S1), small clusters of point defects and small dislocation loops rapidly accumulated in a mostly isolated fashion. In the second stage (S2), the microstructures reached a saturation point for both types of defects. Hence, isolated clusters and small dislocation loops tended to coalesce into more complex defect structures. In both stages, the predominant types of

dislocations were Shockley dislocations (i.e. $1/6\langle 112 \rangle$) and stair-rod dislocations (i.e. $1/6\langle 100 \rangle$ and $1/3\langle 100 \rangle$). The former tended to comprise complex networks of dislocation loops while the latter mainly were present within stacking-fault tetrahedra.

Figure 3a, b quantify our one-point statistics: point-defect density and dislocation density, respectively. In the accumulation stage (S1), both densities sharply increased with the number of irradiation events. Due to the sparse distribution of defects, interactions between neighboring defect clusters and between neighboring dislocation loops were infrequent. However, as the system saturated with defects in the saturation stage (S2), the likelihood of defect interaction increased significantly, and defects tended to coalesce or annihilate rather than nucleate. Note that the point-defect density and the dislocation density followed a nearly identical trend, which suggests that these types of defects were physically coupled, as previously observed⁴³. Further, Fig. 3b shows that the major constituent dislocations (i.e. Shockley and stair-rod) followed a trend nearly identical to that of the total dislocations.

Figure 3c plots our two-point statistic of the irradiation damage: the PCF of the point-defect distribution. Because of the radial averaging, each PCF curve represents the correlation between point defects as a function of their separation distance. For explanation purposes, imagine a random reference point defect. The PCF then reveals the probability of finding another defect at a specified distance away from that reference defect. As the number of irradiation events increased, so did the point-defect density and correspondingly the PCF between 0 (blue) and 38k (red) events. Now consider the general trend in each PCF curve in Fig. 3c: a rapid decrease within region R1, a transition within region R2, and finally a flatline throughout region R3. In R1, the PCF sampled the separation distances of defects within close proximity (i.e. within

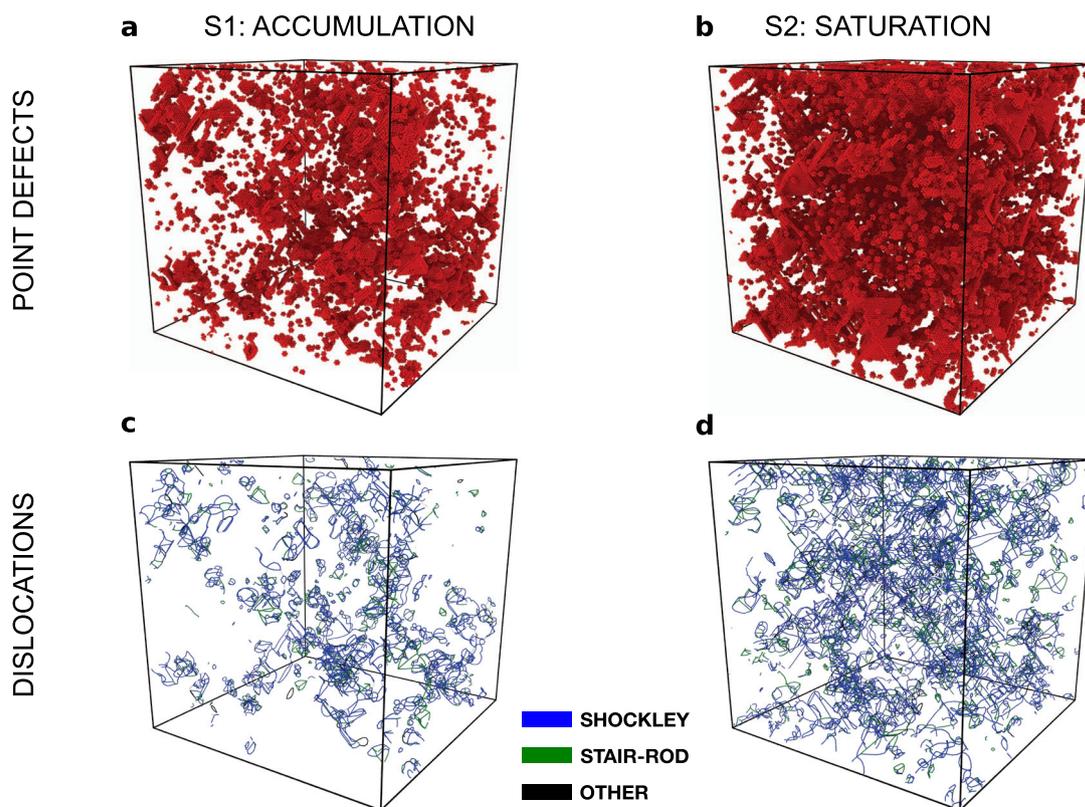


Fig. 2 Evolution of point defects and dislocations. Representative snapshots of **a** and **b**: point defects and **c** and **d**: dislocations in the 4-million-atom simulation cell within the two different stages of damage. In stage 1 (S1), the microstructures experienced a rapid accumulation of point defects and dislocations in a largely isolated fashion. In stage 2 (S2), the microstructures experienced a saturation point and mainly exhibited the coalescence of pre-existing defects into complex defect structures.

small defect clusters). Because the curves were nearly identical in this region, most defect clusters must have contained similar core structures. In R3, the PCF sampled defects that were separated by large distances, usually between separate clusters or within very large clusters. Therefore, the PCF successfully captured the characteristic spacing both within defect clusters (largely in R1) and among defect clusters (largely in R3). Overall, the evolution of the one-point statistics (Fig. 3a, b) and the two-point statistic (Fig. 3c) demonstrated that the accumulation of defects was not a simple matter of density but rather a complex evolution resulting from two competing processes.

To facilitate model construction, we performed principal component analysis (PCA) on the PCF to reduce its dimensionality. Figure 3d shows that the first two principal components (i.e. PCF₁ and PCF₂) accounted for 93.5% and 6.3% of the variation in the PCF, respectively. Because these components effectively summarized the point-defect distribution, we selected them as our final evaluation metrics. By definition, these principal components are orthogonal, so each one captures a unique aspect of the PCF. Further, these components resulted from a linear transformation, so we evaluated the linear correlation of each component to the overall PCF to evaluate the utility of each component (Fig. 3e). Because PCF₁ and PCF₂ had significant Pearson linear-correlation coefficients in R1 and R2, both components must be modeled to capture the short-range order within clusters. In contrast, only PCF₁ was significant in R3, so only PCF₁ would be needed for modeling the long-range order among isolated clusters and within large clusters.

Figure 3f summarizes the four defect statistics that ultimately served as outputs for our machine-learning models: point-defect density, dislocation density, PCF₁, and PCF₂. To compare these statistics, we transformed them into Z-scores by shifting each data

point by the corresponding mean and normalizing by the corresponding standard deviation. Hence, each Z-score represents the number of standard deviations that each point deviated from the mean. Interestingly, the total defect density, total dislocation density, and PCF₁ exhibited strong linear correlations (with linear-correlation coefficients above 0.97). These results confirm that the point defects were physically coupled with the dislocations. As for PCF₁, recall that this principal component had a significant correlation coefficient with the PCF for a broad range of correlation lengths (Fig. 3e). If PCF₁ can well capture the defect distribution, then it should certainly well capture the defect density. In contrast, PCF₂ correlated with a much smaller range of correlation length and exhibited a different trend in Fig. 3f.

Identifying key SAED features

Now consider the SAED patterns simulated from the irradiated microstructures. Figure 4a–d presents the Z-scores for the HIFs of the SAED patterns: area, perimeter, eccentricity, and position for the first two families of diffracted spots. The spot areas (Fig. 4a) followed a nearly identical trend as the point-defect (Fig. 3a) and dislocation densities (Fig. 3b) because defects generally smear a diffraction pattern by interrupting the interferences that would otherwise yield small spots. The spot perimeters (Fig. 4b) followed a similar trend as the spot areas because the spots almost uniformly broadened, as evidenced by the lightly oscillating values of the eccentricity (Fig. 4c). The spot positions (Fig. 4d) also jumped in S1 (i.e. toward the center of the SAED pattern) but then oscillated much more than the area or the perimeter in S2. The position of a spot indicates the length of the diffraction vector in reciprocal space and therefore the corresponding interplanar separation in real space. As defects accumulated and wedged themselves among the crystallographic planes, the planes

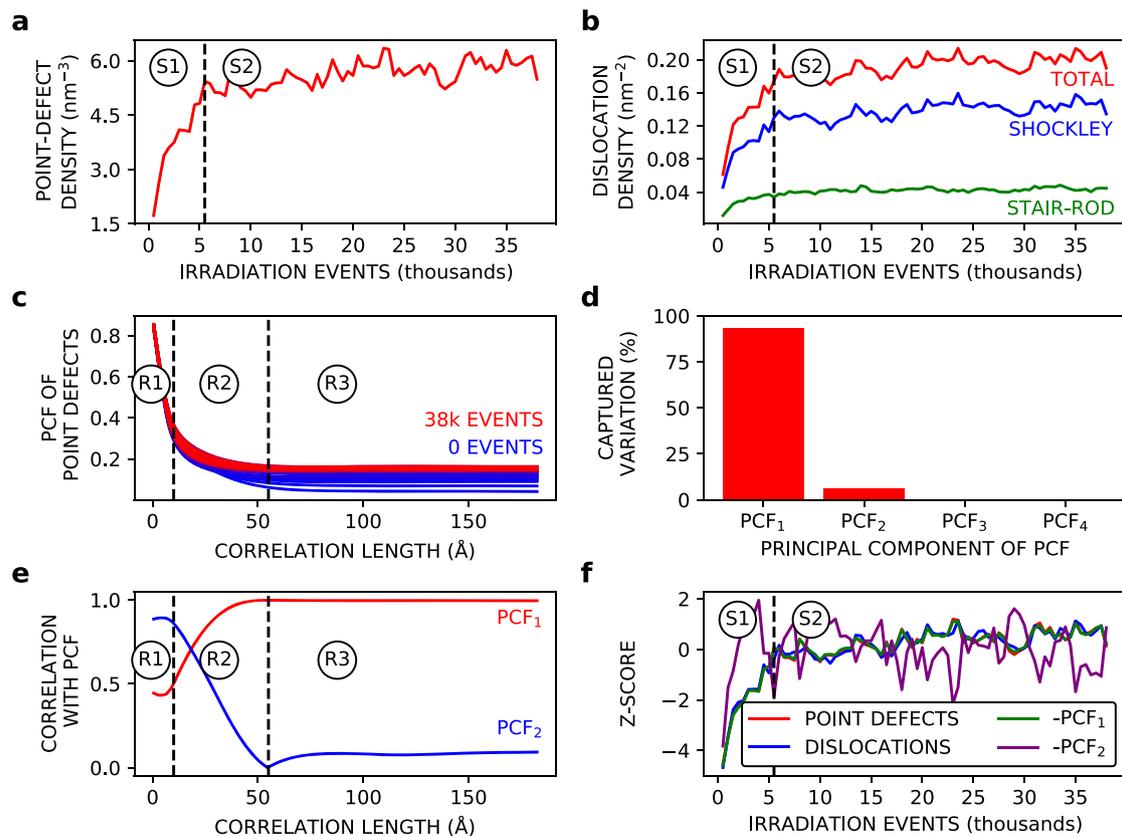


Fig. 3 Defect statistics to be used as outputs in the machine-learning models. Raw defect statistics for **a** point-defect density, **b** dislocation density, and **c** pair-correlation function (PCF) of the point-defect distribution. **d** PCF variation captured by its principal components (PCF_i). **e** Linear-correlation coefficient for PCF₁ and PCF₂ with the overall PCF. **f** Total point defects, total dislocation density, PCF₁, and PCF₂.

expanded, so the spots moved inward. However, as the system saturated, defect annihilation and defect coalescence sometimes reduced the strain to allow the planes to approach their original positions. But then new defects would accumulate and induce strain again, hence the oscillation of the position curves.

Given the similarities in the trends of the areas and perimeters, we produced independent inputs for the human model by performing PCA on the set of 8 HIFs. If we had not ensured independent inputs, the human model would have been susceptible to overfitting error. As shown by Fig. 4e, the first principal component (HIF₁) accounted for 98.6% of the variability in the HIFs, and the second (HIF₂) captured the remaining 1.4%. Predictably, the trend in the dominant HIF₁ resembled that of the point defects, dislocations, and PCF₁. In contrast, HIF₂ exhibited a fairly unique trend, which could be helpful in modeling the elusive PCF₂. Regardless, we replaced the 8 individual HIFs with HIF₁ and HIF₂ as inputs for our human model because these principal components captured the same structural information but in an independent fashion.

To contrast the HIFs and produce the inputs for the machine model, we directly computed the 76 principal components of the 76 pre-processed SAED images to serve as MIFs. Interestingly, the distribution of these principal components was far more spread out than in the HIF analyses. For example, the first two principal components accounted for 99.8% of the PCF (Fig. 3d) and 99.95% of the HIFs (Fig. 4e) but only 46% of the SAED patterns (Fig. 4g). In fact, the first 15 components accounted for only ~70% of the variability of the SAED patterns. Fortunately, we only needed to capture the SAED variability resulting from the salient defects. Further, because MIF₁ (Fig. 4h) alone highly correlated with the point-defect density, total dislocation density, and PCF₁, there

were many more MIFs available to capture the elusive PCF₂. As explained in the next subsection, we used correlation coefficients to downselect which MIFs to include in the machine model because selecting too many MIFs would risk error via overfitting.

Modeling defect statistics from SAED features

For the human model, we had two obvious inputs: HIF₁ and HIF₂. These two principal components effectively summarized the evolution in the area, perimeter, eccentricity, and position for the first two families of SAED spots. As shown by Table 1, HIF₁ had a strong, linear correlation with the point-defect density, dislocation density, and PCF₁. Hence, the human model would certainly capture these defect statistics. In contrast, neither HIF₁ nor HIF₂ had a strong linear correlation with PCF₂, which related to the close-range order of the point-defect distribution. Surprisingly, several of the individual HIFs fared better. For example, the eccentricity for {200} had a correlation coefficient of 0.28 for PCF₂ in comparison to the 0.02 and 0.06 of HIF₁ and HIF₂, respectively. Recall that the principal components of the HIFs captured 99% of the variation of the individual HIFs, including the eccentricity for {200}. Therefore, we deduced that the relationship between the HIFs and the PCF₂ was likely non-linear if present. In this case, a non-linear human model (e.g., based on Gaussian Process Regression) could potentially still predict PCF₂.

In contrast to the two obvious selections for inputs in the human model, we had numerous significant MIFs available for the machine model (recall Fig. 4g). To increase the likelihood of convergence and decrease the likelihood of overfitting error, we downselected to MIFs that had high linear correlation coefficients with the desired defect statistics. This feature-selection technique can obscure non-linear relationships but is a common first

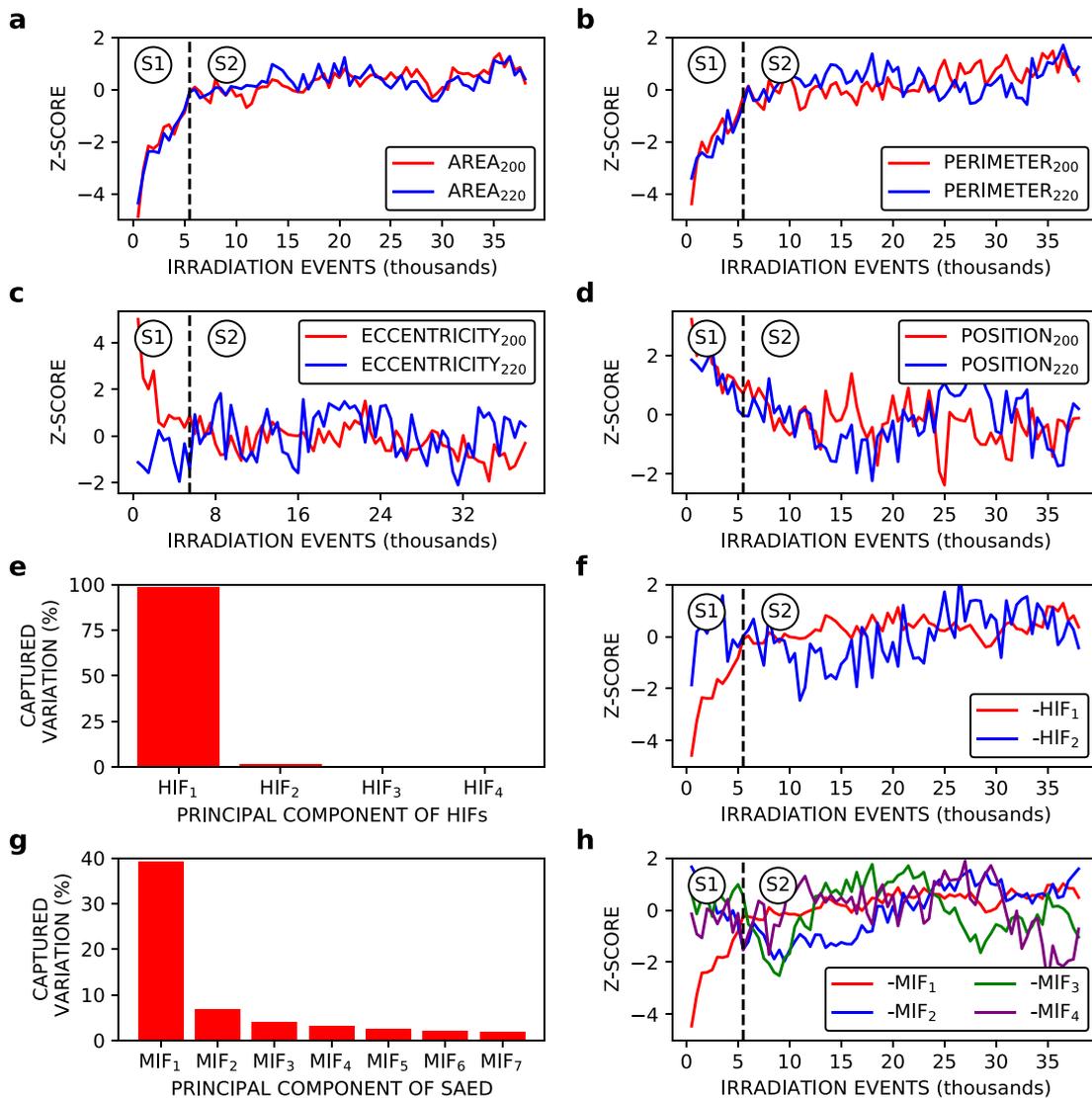


Fig. 4 SAED features to be used as inputs to the machine-learning models. Human-identified features (HIFs), comprise **a** area, **b** perimeter, **c** eccentricity, and **d** position for the first two families of SAED spots. **e** Variation of the HIFs captured by the principal components of the HIFs (i.e. HIF_i). **f** The first two principal components of the HIFs (i.e. HIF₁ and HIF₂). **g** Variation of the SAED patterns captured by their principal components, which were also called the “machine-identified features” (MIFs). **h** The first four MIFs. When plotted against irradiation events, the principal components were negated in order to enhance the inspection of trends.

approach within the machine-learning community. For the current work, we simply desired a proof of concept, around which a fully optimized model could later be conceived. As with HIF₁, MIF₁ had a strong linear correlation with point-defect density, total dislocation density, and PCF₁. Hence, we chose to model these defect statistics by using only the first three MIFs. Overfitting was unlikely with just three inputs, and the additional 2 MIFs might have been useful in a non-linear fashion. As for PCF₂, we noted that several higher MIFs had significant linear correlation coefficients, so we used the first 12 MIFs to model this defect statistic.

Figure 5 provides parity plots that reveal how effectively both machine-learning models decoded the four desired defect statistics from the chosen SAED features. Predictably, both the human model and the machine model captured the following defect statistics that exhibited strong linear correlations with SAED features: point-defect density, dislocation density, and PCF₁ (Fig. 5a–f). However, only the machine model captured PCF₂, which must have been non-linearly correlated with MIFs (Fig. 5g, h).

This difference in performance for predicting the two-point statistic (particularly at short-range order) indicated that SAED patterns contain non-trivial fingerprints beyond human recognition.

DISCUSSION

By connecting features of virtual SAED patterns to a variety of defect statistics from irradiated microstructures, this work demonstrates the power of machine learning in decoding 2D diffractograms. Specifically, we constructed a “human model” that inputs HIFs and a “machine model” that inputs MIFs. Both models were based on the same machine-learning algorithm, so performance differences were solely due to the inputs. The human model certainly had the advantage in the interpretability of its inputs and captured the one-point statistics (i.e. point-defect density and total dislocation density). However, the HIFs incompletely assessed the SAED patterns, so the human model failed to capture the two-point statistic (i.e. PCF of the point-defect

Table 1. Pearson linear-correlation coefficients of the (top 10 rows) human-identified features (HIFs) and (bottom 15 rows) machine-identified features (MIFs) with the desired defect statistics (i.e., point-defect density, dislocation density, and the first two principal components of the pair-correlation function (PCF) of the point-defect distribution).

SAED feature	Point defects	Dislocations	PCF ₁	PCF ₂
HIF ₁	0.96	0.96	0.97	0.02
HIF ₂	0.14	0.18	0.13	0.06
AREA ₂₀₀	0.97	0.97	0.97	0.03
AREA ₂₂₀	0.95	0.94	0.95	0.01
PERIMETER ₂₀₀	0.88	0.88	0.88	0.13
PERIMETER ₂₂₀	0.81	0.81	0.82	0.09
ECCENTRICITY ₂₀₀	0.76	0.76	0.77	0.28
ECCENTRICITY ₂₂₀	0.34	0.34	0.34	0.07
POSITION ₂₀₀	0.70	0.71	0.72	0.18
POSITION ₂₂₀	0.52	0.50	0.53	0.10
MIF ₁	0.98	0.98	0.99	0.03
MIF ₂	0.02	0.02	0.04	0.19
MIF ₃	0.02	0.00	0.00	0.36
MIF ₄	0.07	0.01	0.04	0.57
MIF ₅	0.01	0.05	0.01	0.05
MIF ₆	0.05	0.07	0.05	0.09
MIF ₇	0.09	0.01	0.07	0.12
MIF ₈	0.03	0.01	0.02	0.27
MIF ₉	0.05	0.05	0.05	0.09
MIF ₁₀	0.01	0.05	0.02	0.31
MIF ₁₁	0.03	0.01	0.03	0.21
MIF ₁₂	0.03	0.06	0.03	0.31
MIF ₁₃	0.03	0.02	0.03	0.03
MIF ₁₄	0.05	0.04	0.04	0.02
MIF ₁₅	0.01	0.03	0.01	0.09

Note that a low coefficient merely indicates the lack of a linear relationship not the lack of any relationship.

distribution). In contrast, by incorporating features beyond human recognition, the machine model captured all of the evaluated defect statistics. The MIFs of the machine model resulted from the comprehensive, direct analysis of the full SAED patterns, so avoiding feature-selection bias was much easier than with the HIFs. Further, our MIFs were independent by definition, so it was easier to avoid overfitting errors with MIFs than with the often-correlated HIFs.

Even as proofs of concept, both of our models more accurately and more comprehensively decoded diffractograms than traditional approaches would. However, our framework could be readily enhanced even further in both steps of the decoding process. Regarding the identification of diffractogram features, we could have expanded the types of inputs. For example, we could have incorporated SAED patterns with other zone axes and/or added space-group information⁴⁴. Instead of identifying MIFs via PCA, we could have used texture/shape statistics^{45,46}, training autoencoders, generative neural networks, or transfer learning (from pre-trained convolutional neural networks). Regarding the model building, we could have optimized feature selection and/or evaluated other machine-learning algorithms. For example, we could have considered other combinations of principal components for our machine model. As for the machine-learning algorithm itself, we successfully used Gaussian Process Regression to capture both linear relationships and non-linear relationships within a relatively

small dataset. Alternatively, we could have evaluated the effectiveness of a variety of other algorithms, including support vector machines, random forests, and k-nearest neighbors.

Going forward, we note that our framework could be extended to other structural statistics, other loading conditions, other materials, and other diffraction techniques. In the current work, we assessed point defects and dislocations in irradiated copper via electron diffraction, but the same methods could also be applied to, for example, disclinations of a fatigued ceramic probed with hard X-rays. Further, the machine-learning models did not have to be built exclusively on simulation-based data. We chose to illustrate our concept via simulation in order to establish comprehensive defect statistics to serve as ground truths for model evaluation. However, experimental data could have been used to explore nuanced effects, such as dynamic diffraction and complex loading conditions within larger systems.

METHODS

Accelerated irradiation simulations

We used the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS)⁴⁷ (which is an open-source, molecular-dynamics code) with the reduced-order atomistic cascade method (ROAC)⁴⁸ to simulate radiation damage in copper. Instead of expensively simulating a series of consecutive cascades initiated by primary-knock-on atoms (PKA), the ROAC method simultaneously simulates multiple cascades at once by approximating collision cascades as randomly-located, core-shell regions, provided that the core regions do not overlap. The shells capture the long-range, athermal, recombination-corrected displacements per atom, and the cores capture the short-range, thermal-spike-induced replacements per atom⁴⁹. Depending on the number of atoms and the energy of the recoil spectrum, the ROAC method can model irradiation events 10,000 times faster than consecutive PKA simulations⁴⁸. Hence, we were able to simulate a large number of irradiation events to produce complex defect structures. Specifically, we introduced a total of 38,000 irradiation events of 1 MeV copper ions (500 recoil events were inserted every 20 ps, resulting in 76 unique defect structures) into a fully periodic, 4-million-atom copper sample at room temperature (300 K). We modeled the interatomic interactions with an embedded-atom-method potential that incorporated both nuclear stopping⁵⁰ and electronic stopping^{51,52}. We calibrated the core-shell structures for a broad range of recoil energies by performing consecutive PKA simulations in bulk copper⁴⁸ and then incorporating these results into an extension of the Norgett-Robinson-Torrens DPA model⁴⁹.

Defect characterizations

For each irradiated structure, we identified the point defects via polyhedral template matching⁵³. Using this technique, we compared the local neighborhood of each atom with that of a pristine face-centered-cubic crystal. Atoms that deviated by at least 0.1 Å were considered defects. We also identified the dislocations (i.e. line defects) via the dislocation analysis algorithm in OVITO⁵⁴. We considered Shockley partials (i.e. 1/6<112>), stair-roads (i.e. 1/6<100> and 1/3<100>), and others. To characterize the morphology of the irradiation damage, we computed the PCF^{55,56} from a voxelized representation of the point-defect distribution. This approach had a similar effect as computing a radial distribution function on the discrete data and then performing a smoothing⁵⁷. For the voxelization, each atom not in its lattice position was assigned a spherical volume with a 3-Å radius (which was slightly larger than the first nearest-neighbor distance), and then overlapping volumes were combined. The autocorrelation function was computed on this continuous domain and radially averaged to produce the PCF. Because of the high dimensionality of this two-point statistic, we performed PCA on the PCF via sci-kit learn⁵⁸.

Electron diffraction simulations

For each irradiated structure, we produced a virtual SAED pattern by employing the LAMMPS user-diffraction package^{59,60}. This package computes kinematic electron diffractograms directly from atomistic simulations without prior knowledge of the underlying crystal structure. This package had already been successfully used to detail displacement-cascade formation in a prior study⁵¹. We simulated SAED patterns along the [100] zone axis with an incident electron wavelength of 0.0251 Å,

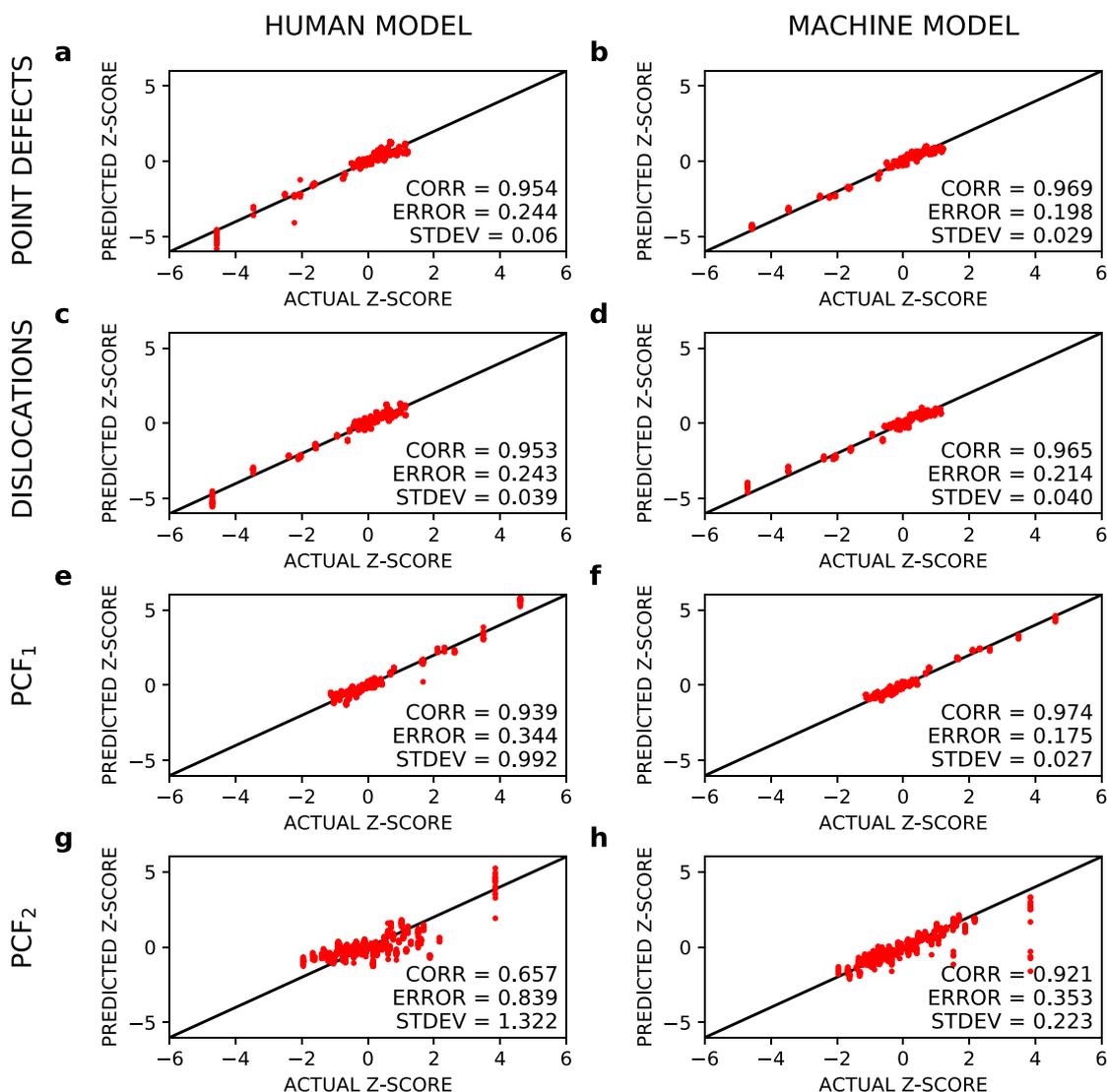


Fig. 5 Parity plots for the human model (left column) and machine model (right column). Evaluated metrics are the Z-scores for the following: total point-defect density (a and b), total dislocation density (c and d), and the first two principal components of the pair-correlation function (PCF) of the point-defect distribution (e–h). The black line is the $y=x$ reference. The Pearson correlation coefficient (CORR), root-mean-squared error (ERROR), and standard deviation (STDEV) as compared to that reference line are provided.

which corresponded to a 200-keV electron beam. To ensure that the reciprocal space in the SAED patterns was sufficiently sampled and to increase the likelihood that the Bragg condition was satisfied, we used a reciprocal lattice spacing of 0.001 \AA^{-1} with an Ewald sphere thickness of 0.005 \AA^{-1} . The maximum region of reciprocal space explored was limited to 1.25 \AA^{-1} in order to capture the diffracted spots corresponding to the two most dominant families of planes.

Diffractogram characterizations

To prepare each SAED pattern for characterization, we performed a series of preprocessing steps. First, we applied both a base-10 logarithmic transformation and a square crop to emphasize the intensities of the spots from the two most dominant families of planes (i.e. $\{200\}$ and $\{220\}$). Then we smoothed the intensity map with a Gaussian blur to ease feature identification. Finally, we removed the direct central spot (which cannot be measured experimentally) and lessened the impact of relrods (which are more prominent in simulation than in experiment) by applying a fixed mask. We constructed this mask by fitting circular regions to the centroids of the diffracted spots in the SAED pattern of the non-irradiated copper.

After preprocessing, we characterized the diffractograms. For the HIFs, we identified the diffraction spots (by constructing a binary map via

thresholding at 5% of the maximum intensity) to compute the area, perimeter, position (via its centroid), and eccentricity (via `sci-kit image`⁶¹) for each spot. Averaging these values across each of the two families of spots yielded a total of eight HIFs per SAED pattern. Incorporating correlated inputs into a model would have complicated its convergence and risked its accuracy, so we performed PCA on the full set of HIFs. The resulting principal components would ultimately serve as the inputs to the human model. For the MIFs, we simply performed PCA directly on the pre-processed SAED patterns. As compared to the human-based approach, this machine-based approach was more straightforward and therefore more robust. For example, there is no need to determine which features to characterize, which threshold to use, or how to ensure feature independence. However, we did need to downselect MIFs to avoid overfitting. As detailed in the Results section, we used the Pearson linear correlation coefficient to inform our feature selection.

Model construction

We used Gaussian Process Regression as implemented in `sci-kit learn`^{58,62} to link the SAED features to the desired defect statistics. This nonparametric, Bayesian approach was chosen for its ability to capture non-linear relationships in relatively low amounts of data. The dataset comprised

SAED features (HIFs and MIFs) and their corresponding defect-statistic labels. To construct both the human model and the machine model and to compute their error distributions, we used a radial basis function kernel, which encodes for smoothness of functions (such that the similarity of inputs corresponds to similarity of outputs). This kernel has two hyperparameters: a signal variance and a density scale. We tuned these hyperparameters by maximizing the log marginal likelihood of the training data while using a gradient-based optimizer for efficiency. Because the log marginal likelihood was not necessarily convex, multiple restarts of the optimizer with different initializations were used. We used k-fold cross validation across 50 data splits (80% training and 20% testing).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

CODE AVAILABILITY

The codes used to calculate the results of this study are available from the corresponding author upon reasonable request.

Received: 21 February 2021; Accepted: 15 April 2021;

Published online: 17 May 2021

REFERENCES

- Zewail, A. H. 4D ultrafast electron diffraction, crystallography, and microscopy. *Annu. Rev. Phys. Chem.* **57**, 65–103 (2006).
- Mo, M. Z. et al. Heterogeneous to homogeneous melting transition visualized with ultrafast electron diffraction. *Science* **360**, 1451–1455 (2018).
- Cao, J. et al. Femtosecond electron diffraction for direct measurement of ultrafast atomic motions. *Appl. Phys. Lett.* **83**, 1044–1046 (2003).
- Holt, M., Harder, R., Winarski, R. & Rose, V. Nanoscale hard X-ray microscopy methods for materials studies. *Annu. Rev. Mater. Res.* **43**, 183–211 (2013).
- Shankar, S. S. et al. Biological synthesis of triangular gold nanoprisms. *Nat. Mater.* **3**, 482–488 (2004).
- Vanacore, G. M., Fitzpatrick, A. W. P. & Zewail, A. H. Four-dimensional electron microscopy: Ultrafast imaging, diffraction and spectroscopy in materials science and biology. *Nano Today* **11**, 228–249 (2016).
- Clabbers, M. T. B. & Abrahams, J. P. Electron diffraction and three-dimensional crystallography for structural biology. *Crystallogr. Rev.* **24**, 176–204 (2018).
- Lippens, B. C. & De Boer, J. H. Study of phase transformations during calcination of aluminum hydroxides by selected area electron diffraction. *Acta Crystallogr.* **17**, 1312–1321 (1964).
- Schwartz, A. J., Kumar, M., Adams, B. L. & Field, D. P. *Electron Backscatter Diffraction in Materials Science*, vol. 2 (Springer, 2009).
- Geiss, R. H., Street, G. B., Volksen, W. & Economy, J. Polymer structure determination using electron diffraction techniques. *IBM J. Res. Dev.* **27**, 321–329 (1983).
- Kolb, U., Gorelik, T. E., Mugnaioli, E. & Stewart, A. Structural characterization of organics using manual and automated electron diffraction. *Polym. Rev.* **50**, 385–409 (2010).
- Weidenthaler, C. Pitfalls in the characterization of nanoporous and nanosized materials. *Nanoscale* **3**, 792–810 (2011).
- Kunka, C., Boyce, B. L., Foiles, S. M. & Dingreville, R. Revealing inconsistencies in X-ray width methods for nanomaterials. *Nanoscale* **11**, 22456–22466 (2019).
- Nam, K.-W. et al. Combining in situ synchrotron X-ray diffraction and absorption techniques with transmission electron microscopy to study the origin of thermal instability in overcharged cathode materials for lithium-ion batteries. *Adv. Funct. Mater.* **23**, 1047–1063 (2013).
- Seymour, T. et al. Evolution of dislocation structure in neutron irradiated Zircaloy-2 studied by synchrotron x-ray diffraction peak profile analysis. *Acta Mater.* **126**, 102–113 (2017).
- Gong, X., Noh, H., Gianneschi, N. C. & Farha, O. K. Interrogating kinetic versus thermodynamic topologies of metal–organic frameworks via combined transmission electron microscopy and X-ray diffraction analysis. *J. Am. Chem. Soc.* **141**, 6146–6151 (2019).
- Escobar, D. P., Depover, T., Duprez, L., Verbeken, K. & Verhaege, M. Combined thermal desorption spectroscopy, differential scanning calorimetry, scanning electron microscopy and X-ray diffraction study of hydrogen trapping in cold deformed TRIP steel. *Acta Mater.* **60**, 2593–2605 (2012).
- Rosenthal, M. et al. High-resolution thermal imaging with a combination of nanofocus X-ray diffraction and ultra-fast chip calorimetry. *J. Synchrotron Radiat.* **21**, 223–228 (2014).
- Khafizov, M. et al. Combining mesoscale thermal transport and x-ray diffraction measurements to characterize early-stage evolution of irradiation-induced defects in ceramics. *Acta Mater.* **193**, 61–70 (2020).
- Kern, J. et al. Taking snapshots of photosynthetic water oxidation using femto-second X-ray diffraction and spectroscopy. *Nat. Commun.* **5**, 1–11 (2014).
- Grygiel, C. et al. In-situ kinetics of modifications induced by swift heavy ions in Al₂O₃: Colour centre formation, structural modification and amorphization. *Acta Mater.* **140**, 157–167 (2017).
- Mendelev, M. I., Sordelet, D. J. & Kramer, M. J. Using atomistic computer simulations to analyze x-ray diffraction data from metallic glasses. *J. Appl. Phys.* **102**, 043501 (2007).
- Li, S. et al. Alkyl chain density and temperature effects on structural properties of pyrrolidinium-based ionic liquids: a combined atomistic simulation and small-angle X-ray scattering study. *J. Phys. Chem. Lett.* **3**, 125–130 (2012).
- Niekiel, F., Bitzek, E. & Spiecker, E. Combining atomistic simulation and x-ray diffraction for the characterization of nanostructures: a case study on fivefold twinned nanowires. *ACS Nano* **8**, 1629–1638 (2014).
- Kim, G. et al. First-principles and machine learning predictions of elasticity in severely lattice-distorted high-entropy alloys with experimental validation. *Acta Mater.* **181**, 124–138 (2019).
- Mason, D. R. et al. Observation of transient and asymptotic driven structural states of tungsten exposed to radiation. *Phys. Rev. Lett.* **125**, 225503 (2020).
- Kusne, A. G. et al. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Sci. Rep.* **4**, 1–7 (2014).
- Iwasaki, Y., Kusne, A. G. & Takeuchi, I. Comparison of dissimilarity measures for cluster analysis of X-ray diffraction data from combinatorial libraries. *npj Comput. Mater.* **3**, 1–9 (2017).
- Ke, T.-W. et al. A convolutional neural network-based screening tool for X-ray serial crystallography. *J. Synchrotron Radiat.* **25**, 655–670 (2018).
- Stanev, V. et al. Unsupervised phase mapping of X-ray diffraction data by non-negative matrix factorization integrated with custom clustering. *npj Comput. Mater.* **4**, 1–10 (2018).
- Aguiar, J. A., Gong, M. L., Unocic, R. R., Tasdizen, T. & Miller, B. D. Decoding crystallography from high-resolution electron imaging and diffraction datasets with deep learning. *Sci. Adv.* **5**, eaaw1949 (2019).
- Garcia-Cardona, C. et al. Learning to predict material structure from neutron scattering data. In *2019 IEEE International Conference on Big Data (Big Data)*, 4490–4497 (IEEE, 2019).
- Oviedo, F. et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Comput. Mater.* **5**, 1–9 (2019).
- Song, Y., Tamura, N., Zhang, C., Karami, M. & Chen, X. Data-driven approach for synchrotron X-ray Laue microdiffraction scan analysis. *Acta Crystallogr. A* **75**, 876–888 (2019).
- Sullivan, B. et al. BraggNet: integrating Bragg peaks using neural networks. *J. Appl. Crystallogr.* **52**, 854–863 (2019).
- Sun, S. et al. Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis. *Joule* **3**, 1437–1451 (2019).
- Cherukara, M. J. et al. AI-enabled high-resolution scanning coherent diffraction imaging. *Appl. Phys. Lett.* **117**, 044103 (2020).
- Kaufmann, K. et al. Crystal symmetry determination in electron diffraction using machine learning. *Science* **367**, 564–568 (2020).
- Lee, J.-W., Park, W. B., Lee, J. H., Singh, S. P. & Sohn, K.-S. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic XRD powder patterns. *Nat. Commun.* **11**, 1–11 (2020).
- Tiong, L. C. O., Kim, J., Han, S. S. & Kim, D. Identification of crystal symmetry from noisy diffraction patterns by a shape analysis and deep learning. *npj Comput. Mater.* **6**, 1–11 (2020).
- Wang, H. et al. Rapid identification of X-ray diffraction patterns based on very limited data by interpretable convolutional neural networks. *J. Chem. Inf. Model.* **60**, 2004–2011 (2020).
- Chan, H. et al. Real-time 3D nanoscale coherent imaging via physics-aware deep learning. Preprint at <https://arxiv.org/abs/2006.09441> (2020).
- Caturla, M. J. et al. Comparative study of radiation damage accumulation in Cu and Fe. *J. Nucl. Mater.* **276**, 13–21 (2000).
- Liu, C.-H., Tao, Y., Hsu, D., Du, Q. & Billinge, S. J. L. Using a machine learning approach to determine the space group of a structure from the atomic pair distribution function. *Acta Crystallogr. A* **75**, 633–643 (2019).
- Haralick, R. M., Shanmugam, K. & Dinstein, I. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* 610–621 (1973).

46. Ojala, T., Pietikainen, M. & Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 971–987 (2002).
47. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).
48. Chen, E. Y., Deo, C. & Dingreville, R. Reduced-order atomistic cascade method for simulating radiation damage in metals. *J. Phys. Condens. Matter* **32**, 045402 (2020).
49. Nordlund, K. et al. Improving atomic displacement and replacement calculations with physically realistic damage models. *Nat. Commun.* **9**, 1–8 (2018).
50. Ziegler, J. F. & Biersack, J. P. The stopping and range of ions in matter. In *Treatise on Heavy-Ion Science*, 93–129 (Springer, 1985).
51. Stewart, J. A. et al. Characterizing single isolated radiation-damage events from molecular dynamics via virtual diffraction methods. *J. Appl. Phys.* **123**, 165902 (2018).
52. Lee, C.-W., Stewart, J., Dingreville, R., Foiles, S. & Schleife, A. Multiscale simulations of electron and ion dynamics in self-irradiated silicon. *Phys. Rev. B* **102**, 024107 (2020).
53. Larsen, P. M., Schmidt, S. & Schiøtz, J. Robust structural identification via polyhedral template matching. *Model. Simul. Mater. Sci. Eng.* **24**, 055007 (2016).
54. Stukowski, A., Bulatov, V. V. & Arsenlis, A. Automated identification and indexing of dislocations in crystal interfaces. *Model. Simul. Mater. Sci. Eng.* **20**, 085007 (2012).
55. Torquato, S. Statistical description of microstructures. *Annu. Rev. Mater. Res.* **32**, 77–111 (2002).
56. Talukdar, M. S., Torsaeter, O. & Ioannidis, M. A. Stochastic reconstruction of particulate media from two-dimensional images. *J. Colloid Interface Sci.* **248**, 419–428 (2002).
57. Vizoso, D., Deo, C. & Dingreville, R. Scaling laws and stability of nano-sized defect clusters in niobium via atomistic simulations and statistical analysis. *J. Mater. Sci.* **54**, 14002–14028 (2019).
58. Pedregosa, F. et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
59. Coleman, S., Spearot, D. & Capolungo, L. Virtual diffraction analysis of Ni [0 1 0] symmetric tilt grain boundaries. *Model. Simul. Mater. Sci. Eng.* **21**, 055020 (2013).
60. Coleman, S., Sichani, M. & Spearot, D. A computational algorithm to produce virtual X-ray and electron diffraction patterns from atomistic simulations. *JOM* **66**, 408–416 (2014).
61. Van der Walt, S. et al. Scikit-image: Image processing in python. *PeerJ* **2**, e453 (2014).
62. Ziatdinov, M. et al. Imaging mechanism for hyperspectral scanning probe microscopy via Gaussian process modelling. *npj Comput. Mater.* **6**, 1–7 (2020).
63. Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO – the Open Visualization Tool. *Model. Simul. Mater. Sci. Eng.* **18**, 015012 (2009).
64. Ayachit, U. *The ParaView guide: A parallel visualization application* (Kitware, Inc., 2015).

ACKNOWLEDGEMENTS

We acknowledge J. Weeks from Word Tree Consulting for editing services and D.L. Medlin from Sandia National Laboratories for a review of the manuscript. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Data, Artificial Intelligence and Machine Learning at DOE Scientific User Facilities program under Award Number 34532. A.S.

and S.R.K. acknowledge funding from ONR award N00014-18-1-2879. A.S. and S.R.K. also acknowledge the cyberinfrastructure resources and services provided by the Partnership for Advanced Computing Environment (PACE) at the Georgia Institute of Technology, and NSF which funded the Hive cluster used in this study under grant 1828187. C.K. and R.D. are supported in part by the Center for Integrated Nanotechnologies, an Office of Science user facility operated for the U.S. Department of Energy. E.Y.C. is supported by a Laboratory Directed Research and Development (LDRD) program at Sandia National Laboratories. Sandia National Laboratory is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy National Nuclear Security Administration under contract DE-NA0003525. The views expressed in the article do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

AUTHOR CONTRIBUTIONS

C.K., A.S., S.R.K., and R.D. conceived the idea; C.K. performed the virtual diffraction simulations; E.Y.C. performed the atomistic simulations; A.S. performed feature engineering and the ML training. R.D. and S.R.K. supervised the work. All authors contributed to the discussion and writing of the paper.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to R.D.

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.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2021