ARL-TR-10377 ● JUNE 2026

Improved Mesoscale Grain Boundary Energy Models for Silicon Carbide from Atomistic-Informed Regression

by Alexander S. Hauck, Matthew Guziewski, and Efraín Hernández–Rivera

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-10377 ● JUNE 2026

Improved Mesoscale Grain Boundary Energy Models for Silicon Carbide from Atomistic-Informed Regression

Alexander S. Hauck
Department of Nuclear Engineering, The Pennsylvania State University

Matthew Guziewski and Efraín Hernández–Rivera
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
June 2026	Technical Report	June 2025	August 2025

4. TITLE AND SUBTITLE
Improved Mesoscale Grain Boundary Energy Models for Silicon Carbide from Atomistic-Informed Regression

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER AH80	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Alexander S. Hauck, Matthew Guziewski, and Efrain Hernandez-Rivera

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
DEVCOM Army Research Laboratory ATTN: FCDD-RLA-MB Aberdeen Proving Ground, MD 21005-5066	ARL-TR-10377

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
ORCID(s): Alexander S. Hauck, 0000-0002-6402-2343; Matthew Guziewski, 0000-0002-5761-720X; Efrain Hernandez-Rivera, 0000-0002-7855-1027

14. ABSTRACT
Grain boundaries have significant influence over material properties such as strength, toughness, fatigue, creep resistance, and thermal conductivity. This is magnified within extreme environments as temperature and pressure control microstructure or radiation impacts local chemistry. Modeling techniques of grain boundary structures have been well explored but are often limited to atomistic techniques such as molecular dynamics. To fully understand material behavior, information from microscale techniques should be passed to macroscale representations of the grain boundaries. This enables their use in higher-scale modeling such as kinetic Monte Carlo to support processing studies such as grain growth with increased accuracy compared to the currently used Read–Shockley model. Here, decision-tree-based regression models are trained using a dataset of 332 [1 0 0], [1 1 0], and [1 1 1] tilt and twist grain boundaries generated in β-SiC. Models are generated and tuned to achieve mean absolute percentage errors below 10.0% with reasonably high coefficients of determination. These models are transformed from Python to ONNX, which enables their use in the open-source C++ kinetic Monte Carlo software SPPARKS.

15. SUBJECT TERMS
Molecular dynamics, ONNX, Potts Monte Carlo, Read-Shockley, SiC, Sciences of Extreme Materials

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UU	33

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Efrain Hernandez	(520) 691-4849

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures viii

List of Tables ix

Acknowledgments x

1. Introduction 1

2. Methods 4

2.1 Grain Boundary Models and Descriptors 4

2.2 Machine Learning Methods 7

3. Results 8

4. Discussion 14

5. Conclusion 15

6. References 17

Appendix. Impact of Summer Research Experience 22

List of Symbols, Abbreviations, and Acronyms 24

Distribution List 25

List of Figures

Figure 1. Prediction error from (a) CatBoost grain boundary (GB) model using macroscopic descriptors and (b) Read–Shockley (RS) model calculated by misorientation angle and fitted parameters ($\gamma_0 = 7.13\ \mathrm{J/m^2}$ and $A = 0.158$). Molecular dynamics grain boundary energy (GBE) values are considered the actual values, and the descriptor data from molecular dynamics passed through each model creates the predicted values. Both actual and predicted GBE are in units of $\mathrm{J/m^2}$. In (a), the predictions from training data are shown in green while the test set predictions are shown in blue. 9

Figure 2. Probability density functions representing the spread of macroscopic descriptor values from the molecular dynamics data (blue) and EBSD data converted to descriptors (red). Here, the bicrystal compatibility is referred to as the "Bicrystal Plane Alignment Ratio," the maximum Schmid factors along $\{h\ k\ l\}\langle h\ k\ l\rangle$ are given as "Max hkl hkl," and the overall maximum Schmid factor is "Max Schmid." 10

Figure 3. (a) Pearson correlation and (b) Spearman correlation heatmaps for macroscopic descriptors and GBE. 11

Figure 4. Probability density functions representing the spread of SiC GBE values from molecular dynamics (blue) and predicted values from measured GB orientation data converted to descriptors (red). The corresponding distribution means are marked by the colored dashed lines with μ, and the bounds for one standard deviation are included as $\mu \pm \sigma$..... 13

Figure 5. Graph of `ONNX` model for an Extra Trees regression model. The model was trained on eight descriptors: bicrystal compatibility, interface–$\{1\,0\,0\}$ plane alignment, interface–$\{1\,1\,0\}$ plane alignment, interface–$\{1\,1\,1\}$ plane alignment, Schmid ratio, Schmid factor $\{1\,1\,1\}\langle 1\,1\,0\rangle$, Schmid factor $\{1\,1\,1\}\langle 1\,1\,2\rangle$, and Schmid factor $\{1\,1\,0\}\langle 1\,1\,0\rangle$. The model output is an array of GBE values. 13

Figure 6. Feature importance comparison for CatBoost, Random Forest, and Extra Trees models. The two most important features for all three models were the Schmid factor $\{1\,1\,1\}\langle 1\,1\,0\rangle$ and interface–$\{1\,1\,0\}$ plane alignment. 15

List of Tables

Table 1. Pearson correlation values with a magnitude greater than 0.4 between pairs of macroscopic descriptors..... 12

Acknowledgments

This work was supported in part by high-performance computer time and resources from the Department of Defense (DoD) High Performance Computing Modernization Program in collaboration with an appointment to the DoD Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the DoD. ORISE is managed by ORAU under DOE contract number DE-SC0014664. All opinions expressed in this report are the author's and do not necessarily reflect the policies and views of DoD, DOE, or ORAU/ORISE. This material is also based upon the work supported under a University Nuclear Leadership Program Graduate Fellowship. Computations for this research were performed on the DoD HPCMP using subproject ARLAP38753387.

A short biography and description of the internship program's impact on the primary author can be found in the Appendix.

1. Introduction

Silicon carbide (SiC) is a wide-bandgap semiconductor with a variety of desirable properties leading to its widespread applications across industries like nuclear power, electronic devices, and armor. A common thread is the robustness of the material, often prompting its use in extreme environments. For nuclear power applications, SiC has been considered for many uses including TRi-structural ISOtropic (TRISO) particle fuel, radiation detectors, and various reactor core structural components.[1,2] TRISO fuel and structural components such as control rod sleeves and fuel rod cladding require mechanical and thermal stability under the intense radiation present in a reactor core. Furthermore, the low neutron absorption cross-section of SiC is beneficial for these uses.[3] Radiation detectors require electronic reliability under similar settings, and a clear link has been drawn between wide-bandgap semiconductors and radiation resistance.[4] SiC's stability at high temperatures and relatively high electron mobility have led to its use in power device applications requiring high voltage, power, and frequency. In the case of armor, hardness, strength, and toughness are key properties, all of which SiC excels at.[5] Correct processing to achieve the desired material properties for each application is essential.

Grain boundaries (GBs) significantly influence the properties and processes in SiC, e.g., strengthening through the inverse Hall–Petch relationship. Further, the use of sintering additives is commonly employed to control microstructure and the resulting mechanical response.[6–8] Wang et al. studied radiation-induced segregation (RIS) of the constituent elements in SiC to GBs.[9] Thermal transport is also affected by GBs, with Watari et al. reporting markedly reduced low-temperature thermal conductivity in polycrystalline SiC relative to single crystals due in part to phonon–boundary scattering.[10] Because direct GB characterization is difficult and costly, modeling is widely used to study how GB structure and thermodynamics influence these properties, with multiscale approaches required to bridge the atomistic scale at which GBs form with the meso- and macroscales at which their collective behavior is observed.

At the mesoscale, GB-mediated microstructural evolution–in particular grain growth–is commonly simulated using curvature-driven kinetic Monte Carlo methods such as the Potts Monte Carlo model implemented in `SPPARKS`.[11] The driving force for boundary migration is the reduction of the total interfacial energy of the system.[11–14]

In these simulations, the energy of each boundary segment is typically computed from a single macroscopic descriptor, the misorientation angle θ, via the Read–Shockley (RS) relation of Equation 1, where γ_0 and A are material-dependent constants.[15]

$$\gamma_{\mathrm{GB}} = \gamma_0 \begin{cases} \theta\left(A - \ln\left(\theta\right)\right), & \theta \leq 15^\circ \\ 1, & \theta > 15^\circ \end{cases} \tag{1}$$

The RS form is derived assuming an isotropically elastic body, low-angle dislocation description of the boundary and reduces the full five-parameter macroscopic GB character to a single scalar, which limits the fidelity with which it can represent anisotropy or chemistry-specific behavior in covalent crystals such as SiC. Developing a more descriptive surrogate for γ_{GB} that can be trained on atomistic data and evaluated cheaply at the mesoscale is therefore an active area of research, and is the focus of the present work (i.e., integration of such a surrogate into a Potts model is left as future work).

On the atomistic side, the GB energy landscape of SiC has been mapped using molecular dynamics with empirical interatomic potentials. Kohler examined $\langle 0\,0\,1\rangle$ and $\langle 1\,1\,0\rangle$ symmetrical tilt GBs in β-SiC constructed from structural units and relaxed to a minimum-energy configuration, finding that boundaries containing antiphase boundaries were generally—but not always—energetically unfavorable.[16] Wojdyr et al. later developed an iterative MD relaxation scheme employing repeated NVT cycles and conjugate-gradient minimizations on overlapping zincblende lattices for $\langle 0\,0\,1\rangle$ symmetric tilt GBs and showed that the resulting optimized structures frequently contain 6- and 7-atom rings with energies below the structural-unit models obtained with the same potential.[17] Building on a Monte Carlo strategy originally developed by Banadaki et al. for Al, Guziewski et al. extended an atomistic Monte Carlo optimization scheme—using trial atom removals, insertions, type swaps, and rigid GB translations accepted via a Metropolis-like criterion—to diamond-cubic Si and zincblende SiC, recovering tilt-GB structures with energies comparable to or below those of prior studies at reduced computational cost.[18,19] For a multi-element crystal such as SiC, the relative arrangement of the atomic bases across the boundary must additionally be enumerated, yielding several inequivalent chemistry configurations for each macroscopic GB orientation. These atomistic datasets provide the ground-truth energies needed for data-driven GB-energy modeling.

Machine learning (ML) has emerged as a practical route to surrogate models that map descriptors of a grain or phase boundary directly to its interfacial energy or mechanical response, sidestepping repeated atomistic relaxations. For SiC GBs, Trujillo et al. trained kernel ridge regression models on microscopic descriptors to predict the per-operation energy change associated with individual Monte Carlo trial moves, then used those predictions inside the optimization loop to accelerate the search for low-energy structures by roughly an order of magnitude.[20] Montes de Oca Zapiain et al. combined polyhedral template matching and Smooth Overlap of Atomic Positions (SOAP) descriptors with boosted regression trees to predict tensile strength distributions of metastable $\Sigma 5$ and $\Sigma 9$ tilt GBs in SiC, identifying GB excess free volume and C–C bond density as the most informative microscopic features.[21] Guziewski et al. subsequently produced a broader SiC GB dataset spanning 344 unique symmetric tilt and twist coincident site lattice (CSL) orientations (with $\Sigma < 200$) and three atomic-basis variants per orientation, and trained random-forest regressors using a macroscopic descriptor set (15 variables), a microscopic set (10 variables), and their union to predict both GB energy and tensile strength.[22] While combined and microscopic feature sets yielded the most accurate models, the macroscopic-only random forest still predicted GB energy to within 9.9% on average, indicating that quaternion-derived macroscopic descriptors alone retain sufficient information to be useful surrogates for γ_{GB}. Analogous studies in other materials systems—for example, the systematic evaluation by Owens et al. of descriptor transformations and regression algorithms across $7\,304$ Al GBs—further emphasize the importance of feature selection, multicollinearity control, and the use of interpretable, tree-based models when training GB-energy surrogates from finite atomistic datasets.[23] Similar tree-ensemble approaches have proven effective for related materials-property regression problems, lending additional confidence to this class of methods.[24,25]

The present work builds on this body of literature with two contributions. First, a generalized, largely automated regression pipeline is developed in `Python` (using `scikit-learn` and `PyCaret`) that ingests a GB descriptor–energy dataset, screens features for multicollinearity, selects and hyperparameter-tunes a regression model, and emits an interpretable, decision-tree-based surrogate for γ_{GB}. The pipeline is applied to the macroscopic-descriptor subset of the Guziewski et al. SiC dataset,[22] averaged across the three atomic-basis variants since macroscopic descriptors cannot encode local chemical information, to train a SiC-specific GB-energy surrogate

that is then evaluated on macroscopic descriptors derived from electron backscatter diffraction (EBSD) measurements of polycrystalline β-SiC. Second, the trained model is exported to the framework-agnostic `ONNX` format,[26] enabling it to be loaded directly from `C++` (or any language for which `ONNX Runtime` is available) so that it can ultimately replace the RS energy call inside mesoscale GB-evolution codes such as `SPPARKS`; that integration is reserved for future work, but the model and interface developed here provide the necessary building block.

2. Methods

2.1 Grain Boundary Models and Descriptors

Models of SiC GBs were generated within `LAMMPS` as described in Guziewski et al.[22,27] These simulations used the screened empirical bond-order potential for Si–C by Pastewka et al., a Tersoff–Brenner type potential.[22,28] This choice was originally made to ensure successful modeling of brittle fracture in SiC and comparable mechanical response to that from density-functional theory. A summary of the approach and relevant macroscopic GB descriptors is presented below.

Training an ML model for this application makes grain boundary energy (GBE) the most appropriate choice of target variable. Consistent with the previous work by Guziewski et al., Equation 2 gives the energy for a GB in a multi-element system based on the grand canonical potential, as defined by Kohyama et al.[29] The difference is taken between the total system energy, E_{tot}, and the pristine SiC energy (total number of Si/C atoms $N = N_{\mathrm{Si}} + N_{\mathrm{C}}$ times the per-atom cohesive energy of SiC, E_{SiC}). This is scaled by the area of the GB, as is the following term that is subtracted from the first. This second term accounts for the change in chemical potential due to stoichiometric changes, where the difference in chemical potential is given by $\Delta\mu = \mu_{\mathrm{Si}} - \mu_{\mathrm{C}}$ and the difference in atom numbers is $\Delta N = N_{\mathrm{Si}} - N_{\mathrm{C}}$.

$$\gamma_{\mathrm{GB}} = \frac{E_{\mathrm{tot}} - NE_{\mathrm{SiC}}}{A_{\mathrm{GB}}} - \frac{\Delta N \Delta\mu}{2A_{\mathrm{GB}}} \tag{2}$$

Exact chemical potential values depend heavily on the local environment, and instead, upper and lower bounds may be determined by the relationship below from the bulk energies of silicon, graphite, and SiC as discussed in Guziewski et al.[22] To simplify modeling efforts, local chemical environments were assumed to

be stoichiometric in all cases, and therefore, using values from the Pastewka et al. potential, $\Delta\mu = -2.78\mathrm{eV}$.[28]

$$\mu_{\mathrm{SiC}}^{\mathrm{bulk}} - 2\mu_{\mathrm{C}}^{\mathrm{bulk}} \leq 2\mu_{\mathrm{Si}}^{\mathrm{bulk}} - \mu_{\mathrm{SiC}}^{\mathrm{bulk}}$$

The misorientation angle of a GB simply refers to the rotation angle between the two grains demarcating the boundary.[17] The rotation angle and rotation axis define the relative lattice misorientation—3 of the 5 macroscopic degrees of freedom of a GB. Using a quaternion representation, the element-wise notation for each grain orientation in the 4-D vector space is given by $\mathbf{q} = q_0 + q_1\hat{\mathbf{i}} + q_2\hat{\mathbf{j}} + q_3\hat{\mathbf{k}}$. The rotation quaternion between two quaternions $\mathbf{q_a}$ and $\mathbf{q_b}$ is calculated by $\mathbf{q_r} = \mathbf{q_b}\mathbf{q_a^*}$, where $\mathbf{q}^* = q_0 - q_1\hat{\mathbf{i}} - q_2\hat{\mathbf{j}} - q_3\hat{\mathbf{k}}$ denotes the conjugate. Assuming $\mathbf{q}$ is a unit quaternion (versor), an equivalent representation highlights the axis-angle parameterization $\mathbf{q} = (q_0, \hat{q}) = [\cos(\omega/2), \hat{n}\sin(\omega/2)]$.[30,31] Here, ω represents the rotation of 3-D space about the axis of rotation $\hat{n}$. The rotation angle ω may be calculated by $\omega = 2\arccos(|q_0|)$ for the range $[0, \pi]$, or by the four-quadrant inverse tangent $\omega = 2\arctan2(||\hat{q}||/|q_0|)$, which ranges over $[-\pi, \pi]$.[31] In the current work, equivalent rotations are taken by utilizing symmetry to simplify the domain and ensure the scalar part, q_0, is positive.[30] A kd-tree of rotation quaternions spanning all combinations of pairs within a relevant set of low-index plane normal vectors is generated using `SciPy`.[32] Querying the kd-tree facilitates efficient determination of nearest neighbor low-index planes to the GB orientation.

Rotation quaternions between the grain rotation and two nearest low-index plane rotations are determined. For SiC, low-index planes in the $\{1\,0\,0\}$, $\{1\,1\,0\}$, and $\{1\,1\,1\}$ families are considered due to their previously demonstrated impact on energetic properties with GB alignment.[33–35] The bicrystal plane alignment, ψ, is determined as the angle of rotation between the GB and the nearest low-index plane.[36] This is shown generally for any given plane in Equation 3, where $\mathbf{r}$ is the rotation quaternion between the GB quaternion and the plane quaternion.

$$\psi = 2\arctan2\left(\frac{||\hat{r}||}{|r_0|}\right) \tag{3}$$

The ratio, or bicrystal compatibility, is taken between this value for the nearest and

second nearest low-index planes (ψ_1 and ψ_2, respectively). The bicrystal compatibility serves as a measure of strain accommodation by the GB.[22]

$$r_p = \frac{\psi_1}{\psi_2} \tag{4}$$

Similarly, the angle between these low-index plane normals (corresponding to $\{1\,0\,0\}$, $\{1\,1\,0\}$, and $\{1\,1\,1\}$) and the bicrystal interface normal are of interest due to the fracture behavior of SiC.[37] Preferential crack propagation occurs within these planes, and therefore, the minimum angle, $\kappa_{\min}$, is determined for each plane family by Equation 5. Here, the given plane normal is expressed as the unit vector $\hat{p}_i$ while $\hat{q}$ refers to the GB rotation axis.

$$\kappa_{\min} = \min\left(\arccos\left|\hat{p}_i \cdot \hat{q}\right|\right) \tag{5}$$

GBs strengthen materials based on type, structure, orientation, and spacing.[38] Preferential crack propagation in zincblende SiC and diamond cubic Si occurs along the $\{1\,1\,0\}\langle 1\,1\,0\rangle$, $\{1\,1\,0\}\langle 1\,1\,1\rangle$, $\{1\,1\,1\}\langle 1\,1\,0\rangle$, and $\{1\,1\,1\}\langle 1\,1\,2\rangle$ slip systems.[37,39–41] Five descriptors related to the Schmid factor are included to capture the critical slip systems for resolved shear stress.[22] These include the maximum Schmid factor for each of the preferred directions, the maximum Schmid factor overall, and the ratio between the two highest Schmid factors.[42] The maximum values for each slip system are found by Equation 6 and account for the largest resolved shear stress on the system. Practically, these may be calculated in a similar fashion as the interface–plane alignment descriptors where i and j are selected such that only the independent slip systems are considered (not all combinations of i and j).

$$m_{\max} = \max\left(\cos\left(\phi\right)\cos\left(\lambda\right)\right) = \max\left(\left|\hat{p}_i \cdot \hat{q}\right| \cdot \left|\hat{p}_j \cdot \hat{q}\right|\right) \tag{6}$$

The ratio of the primary (m_1) and secondary (m_2) Schmid factors across the systems evaluated accounts for the possibility of multi-slip occurring, given by Equation 7.

$$r_m = \frac{m_1}{m_2} \tag{7}$$

The directional Young's modulus along the grain rotation axis, $E_{\hat{q}}$, is computed using Equation 8.[43] This descriptor accounts for fracture behavior further by considering the elastic anisotropy between the two grains, which can cause localization effects at the GB and crack propagation.[42]

$$E_{\hat{q}}^{-1} = \frac{C_{11} + C_{12}}{(C_{11} - C_{12})\,(C_{11} + 2C_{12})} + \left(\frac{1}{C_{44}} - 2\frac{1}{C_{11} - C_{12}}\right) \frac{q_1^2 q_2^2 + q_2^2 q_3^2 + q_1^2 q_3^2}{||\hat{q}||^4} \quad (8)$$

The reported elastic tensor values for the Pastewka et al. potential—with those primarily of interest being $C_{11} = 405$ GPa, $C_{12} = 145$ GPa, and $C_{44} = 247$ GPa for β-SiC—agree well with experimental results and are used here.[44]

2.2 Machine Learning Methods

An ML regression script was constructed in `Python` using primarily `scikit-learn` and `PyCaret`.[45,46] This script was designed to be sufficiently general for a number of GB ML applications that are appropriately addressed by regression methods. Namely, several applications were tested including predicting change in energy due to per-atom removal and insertion operations in Si[20] and predicting GBE based on macroscopic descriptors of the GB.[22] The specific application is almost exclusively based on the dataset provided, and little to no alterations to the function calls are necessary. Either a user-defined train-test split or automated split of the data may be used. In `PyCaret`, the training set is used for both training and validation, and the testing data is used as a hold-out set. The types of regression models considered by `PyCaret` should be defined as well as if graphics processing unit (GPU) training is to be enabled for supported algorithms.

Prior to model creation, a Pearson correlation matrix is created using the training data including the target variable (e.g., GBE). The matrix describes the Pearson correlation coefficients between pairs of descriptors, indicating the strength of linear correlation between them (0–1) and direction (+/−). If multiple descriptors are correlated above a certain threshold, all but the descriptor most correlated with the target variable are removed from the dataset. This is done to prevent multicollinearity and maintain model simplicity, improving performance.[47] The choice of threshold value above which descriptors are considered sufficiently correlated allows for more

or less information to be kept.

The selected regression models are created and trained by `PyCaret`. All trained models are evaluated against several metrics, and the best performing is selected based on primarily the mean absolute percent error (MAPE). MAPE was selected to focus on prediction accuracy, and the lack of outliers and near-zero values in the dataset ensure its usefulness. Other metrics tracked include the coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and root mean squared logarithmic error (RMSLE). The hyperparameters of the selected model are tuned for 100 iterations utilizing the Tree-structured Parzen Estimator search as implemented in Optuna.[48] K-Fold cross-validation is employed during hyperparameter tuning, and the tuned model is evaluated again using primarily MAPE. With the tuned model, predictions are generated, performance figures are created, and the model is saved.

3. Results

Several regression models (namely the CatBoost, Extra Trees, Random Forest, AdaBoost, Extreme Gradient Boosting, Decision Tree, Bayesian Ridge, Support Vector, Ridge, Kernel Ridge, and Multi-layer Perceptron Regressors) were created during training and evaluated against metrics to determine the best performing. K-Fold cross-validation is performed with $K = 10$. Consistently, three decision-tree-based ensemble learning methods outperform all other chosen model types across metrics: Random Forest, Extra Trees, and CatBoost.[47,49–51] These models are predicated upon the creation of a tree structure with nodes representing decisions based on data features. The root node denotes the first decision and branches into internal (decision) nodes. The branching process continues from each internal node until leaf (terminal) nodes are reached that represent possible outcomes. These specific types of decision-tree-based methods represent specific adjustments to address their disadvantages, often using a set of decision trees and methods such as bagging and boosting. For example, the Random Forest method uses bagging to randomly select a sample of data where each data point can be chosen more than once and their aggregated predictions are taken for more accurate results.[47] Additionally, it employs feature randomness to perform a similar action across features and generate a random split. This reduces the chance of overfitting by minimizing correlation and creates a model with interpretability through feature importance.

To improve the GBEs calculated in the Potts Monte Carlo simulations, the ML model implemented should predict GBE values more aligned with the "true" (molecular dynamics) values than those calculated via the RS model. This can be compared most clearly using the prediction error format, where the line of best fit for a perfect model matches the identity, meaning the actual and predicted values are identical. The MAPE values are included, where a lower value shows higher accuracy. Additionally, a high R^2 closer to 1 indicates that the descriptors used capture the variability of GBE. Figure 1 compares the fit quality of a trained CatBoost ML model (Figure 1(a)) against the RS model given by Equation 1 (Figure 1(b)).

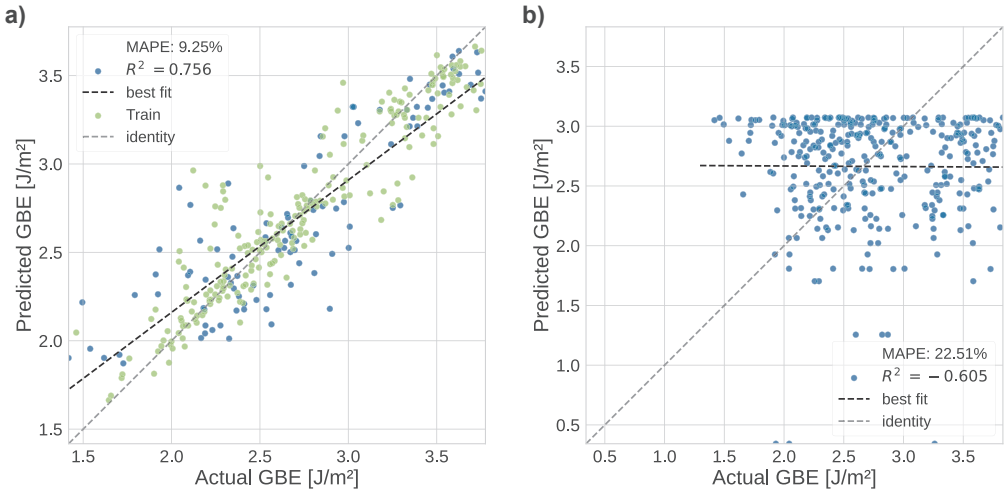

Figure 1. Prediction error from (a) CatBoost grain boundary (GB) model using macroscopic descriptors and (b) Read–Shockley (RS) model calculated by misorientation angle and fitted parameters ($\gamma_0 = 7.13\ \mathrm{J/m^2}$ and $A = 0.158$). Molecular dynamics grain boundary energy (GBE) values are considered the actual values, and the descriptor data from molecular dynamics passed through each model creates the predicted values. Both actual and predicted GBE are in units of $\mathrm{J/m^2}$. In (a), the predictions from training data are shown in green while the test set predictions are shown in blue.

Distributions for the descriptors used are shown in Figure 2. Here, the data generated from the GB models created in molecular dynamics is used as reference data. Bunge–Euler angles ((ϕ_1, Φ, ϕ_2) using ZXZ convention) were measured for each pair of grain orientations across 25,282 grain boundaries in β-SiC using EBSD. The angles were converted to quaternions using `SciPy`, and in turn, quaternions were transformed into descriptors as detailed in Section 2.1.

The reference data in Figure 2 was first analyzed for strong correlations using both the

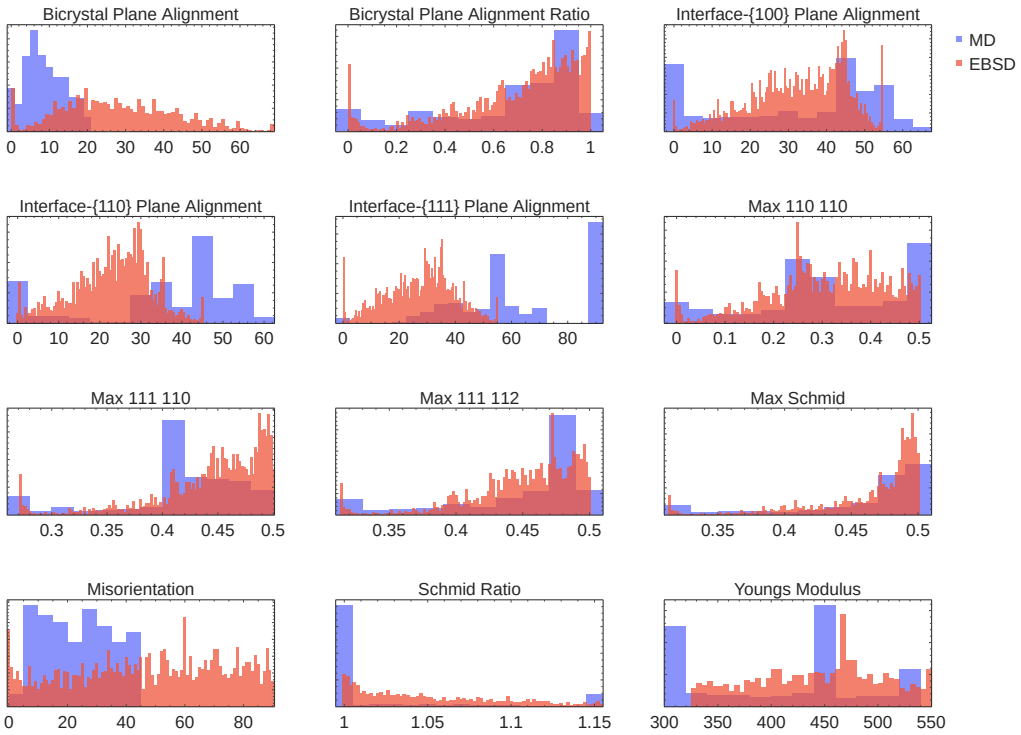

Figure 2. Probability density functions representing the spread of macroscopic descriptor values from the molecular dynamics data (blue) and EBSD data converted to descriptors (red). Here, the bicrystal compatibility is referred to as the "Bicrystal Plane Alignment Ratio," the maximum Schmid factors along $\{h\,k\,l\}\langle h\,k\,l\rangle$ are given as "Max hkl hkl," and the overall maximum Schmid factor is "Max Schmid."

Pearson and Spearman metrics, shown in Figures 3(a) and 3(b), respectively. Several variables can be considered as such, identified by the darker red (correlated) or blue (anti-correlated) squares. Values for both the Pearson and Spearman correlations are quite similar, indicating that the relationships between variables are mostly linear and monotonic.

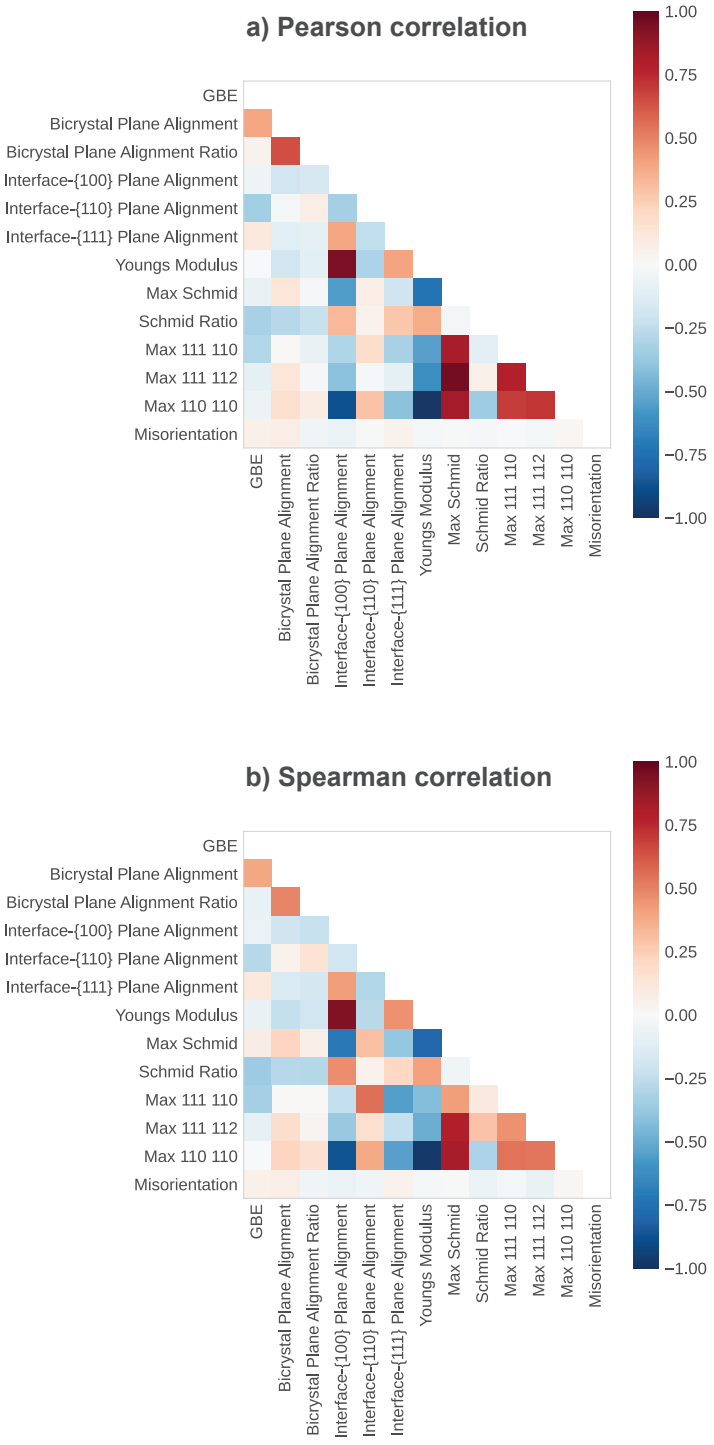

Figure 3. (a) Pearson correlation and (b) Spearman correlation heatmaps for macroscopic descriptors and GBE.

The Pearson correlation was used to identify instances of multicollinearity. Fifteen Pearson correlation values have a magnitude of 0.4 or more, listed in Table 1. A correlation threshold of 0.7 was used to determine if a pair of descriptors was sufficiently correlated to remove one from the dataset prior to training the CatBoost model shown in Figure 1(a). This left a dataset of seven descriptors: bicrystal plane alignment, bicrystal compatibility, interface–$\{1\,1\,0\}$ plane alignment, interface–$\{1\,1\,1\}$ plane alignment, Schmid ratio, Schmid factor $\{1\,1\,1\}\langle 1\,1\,0\rangle$, and misorientation angle.

Table 1. Pearson correlation values with a magnitude greater than 0.4 between pairs of macroscopic descriptors.

Descriptor 1	Descriptor 2	Pearson correlation
Schmid factor $\{1\,1\,0\}\langle 1\,1\,0\rangle$	Young's modulus	-0.98
Schmid factor $\{1\,1\,1\}\langle 1\,1\,2\rangle$	maximum Schmid factor	0.96
Young's modulus	interface–$\{1\,0\,0\}$ plane alignment	0.94
Schmid factor $\{1\,1\,0\}\langle 1\,1\,0\rangle$	interface–$\{1\,0\,0\}$ plane alignment	-0.87
Schmid factor $\{1\,1\,0\}\langle 1\,1\,0\rangle$	maximum Schmid factor	0.84
Schmid factor $\{1\,1\,1\}\langle 1\,1\,0\rangle$	maximum Schmid factor	0.83
Schmid factor $\{1\,1\,1\}\langle 1\,1\,2\rangle$	Schmid factor $\{1\,1\,1\}\langle 1\,1\,0\rangle$	0.79
maximum Schmid factor	Young's modulus	-0.74
Schmid factor $\{1\,1\,0\}\langle 1\,1\,0\rangle$	Schmid factor $\{1\,1\,1\}\langle 1\,1\,2\rangle$	0.72
Schmid factor $\{1\,1\,0\}\langle 1\,1\,0\rangle$	Schmid factor $\{1\,1\,1\}\langle 1\,1\,0\rangle$	0.69
bicrystal plane alignment	bicrystal compatibility	0.64
Schmid factor $\{1\,1\,1\}\langle 1\,1\,2\rangle$	Young's modulus	-0.61
maximum Schmid factor	interface–$\{1\,0\,0\}$ plane alignment	-0.56
Schmid factor $\{1\,1\,1\}\langle 1\,1\,0\rangle$	Young's modulus	-0.54
Schmid factor $\{1\,1\,1\}\langle 1\,1\,2\rangle$	interface–$\{1\,0\,0\}$ plane alignment	-0.41

The descriptors calculated from EBSD data were passed to the CatBoost model to predict GBE values, presented in Figure 4. The GBE values from the molecular dynamics training data range from 1.42–$3.83\ \mathrm{Jm}^{-2}$ with a mean value of $(2.73\ \pm\ 0.57)\ \mathrm{Jm}^{-2}$. In comparison, the CatBoost predictions for the EBSD data yield a mean GBE of $(2.70\ \pm\ 0.22)\ \mathrm{Jm}^{-2}$. The minimum CatBoost prediction is $1.92\ \mathrm{Jm}^{-2}$ and the maximum is $3.69\ \mathrm{Jm}^{-2}$.

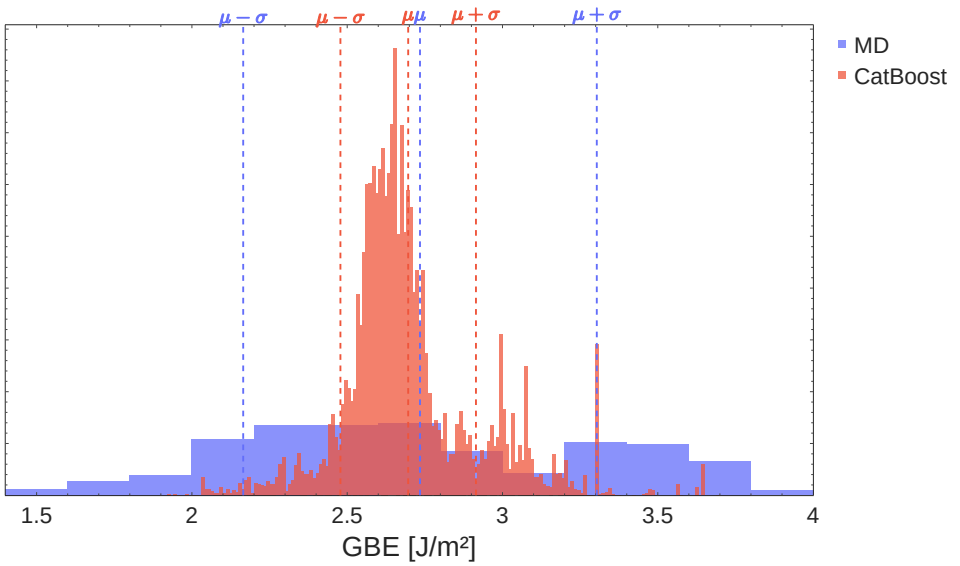

Figure 4. Probability density functions representing the spread of SiC GBE values from molecular dynamics (blue) and predicted values from measured GB orientation data converted to descriptors (red). The corresponding distribution means are marked by the colored dashed lines with μ, and the bounds for one standard deviation are included as $\mu \pm \sigma$.

Figure 5. Graph of `ONNX` model for an Extra Trees regression model. The model was trained on eight descriptors: bicrystal compatibility, interface–$\{1\,0\,0\}$ plane alignment, interface–$\{1\,1\,0\}$ plane alignment, interface–$\{1\,1\,1\}$ plane alignment, Schmid ratio, Schmid factor $\{1\,1\,1\}\langle 1\,1\,0\rangle$, Schmid factor $\{1\,1\,1\}\langle 1\,1\,2\rangle$, and Schmid factor $\{1\,1\,0\}\langle 1\,1\,0\rangle$. The model output is an array of GBE values.

Trained models are converted to the `ONNX` format in `Python`. The graph for a converted Extra Trees regression model is visualized using `Netron` in Figure 5. The X matrix is required to have eight columns to correspond with the number of descriptor variables used during training. The descriptor names are passed along in the model description during the conversion. X may have any number of rows (denoted by "?"). X is passed to the trained model object corresponding to the regression function to compute the prediction values y. This column vector is created with the same number of rows as X (one prediction per one row of data). The "Cast" node ensures this vector is formatted to contain double-precision floating-point numbers and names the y vector

to be the same as the training data target variable ("GBE" in this case). Using this format, the model can be loaded into any software that is coded in a language where `ONNX Runtime` is available.[26] Here, it is used to transfer the model from `Python` to `C++`.

4. Discussion

As with all ML applications, the dataset provided is of the utmost importance to the quality of the model generated. Removing descriptors to prevent multicollinearity and overfitting can also lead to underfitting, and therefore, careful selection of this value is critical. In addition, certain variables with significantly higher correlation to the target variable may heavily impact results. Only four of the descriptors used have a Pearson correlation (or anticorrelation) with GBE above 0.30: bicrystal plane alignment (0.39), interface–$\{1\,1\,0\}$ plane alignment (-0.34), Schmid ratio (-0.32), and Schmid factor $\{1\,1\,1\}\langle 1\,1\,0\rangle$ (-0.30). While the quaternion to macroscopic descriptor distributions in Figure 2 (red) show a similar distribution to that of the training data (blue) for most of these, the bicrystal plane alignment values from the training data skew significantly toward smaller angles.

This is related to a significant challenge of decision-tree-based regression models like Random Forests, which is the determination of a prediction interval.[52] Figure 2 is used to ensure a reasonable range for each descriptor variable. If even a single feature is markedly outside the range presented in the training data, the prediction interval may be skewed, particularly if the feature has a high importance. As regression models inherently struggle with extrapolation, the distribution of the model predictions against the training data target distribution provides a useful indication of the model efficacy.

Feature importance, shown in Figure 6, enables some interpretation of the physical importance of the macroscopic descriptors on GBE. Tree-based models determine feature importance based on which features contribute most to the prediction. Importances of each node are calculated by considering what portion of the overall data reaches and is split by the node. How many nodes split based on a certain feature and their corresponding importances are used to yield a feature's importance. Here, conclusions can be drawn about the relative importance of a GB's value for each descriptor on the GBE value. This does not imply causation between the descriptors of higher importance and the GBE. Figure 6 shows that three different models

place high importance on the maximum Schmid factor along $\{1\,1\,1\}\langle 1\,1\,0\rangle$ and the interface–$\{1\,1\,0\}$ plane alignment, ranking them first or second (in either order). Other important features include the bicrystal plane alignment, bicrystal compatibility, and the maximum Schmid factors along $\{1\,1\,0\}\langle 1\,1\,0\rangle$ and $\{1\,1\,1\}\langle 1\,1\,2\rangle$.

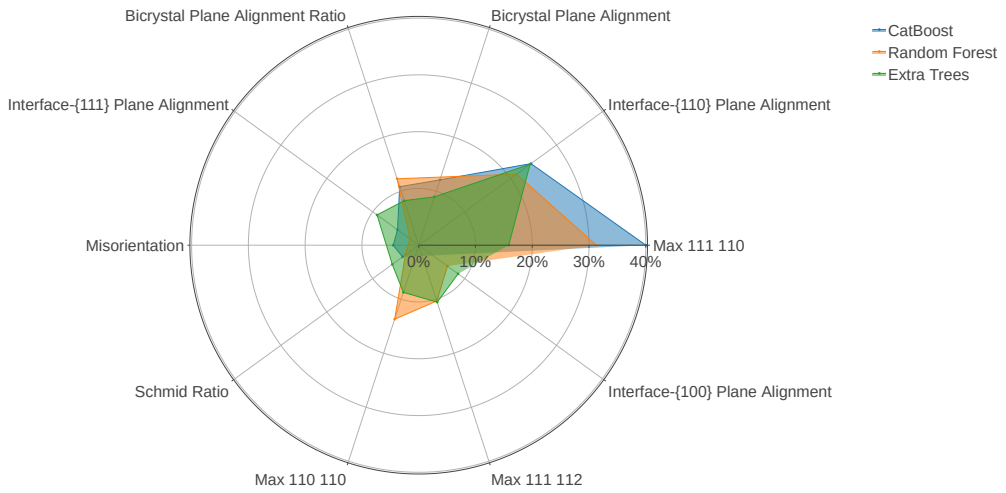

Figure 6. Feature importance comparison for CatBoost, Random Forest, and Extra Trees models. The two most important features for all three models were the Schmid factor $\{1\,1\,1\}\langle 1\,1\,0\rangle$ and interface–$\{1\,1\,0\}$ plane alignment.

While results from some specific trained models are presented, none have yet been applied to the grain growth simulations. Future work will explore this to see the effect of these models on the grain growth behavior.

5. Conclusion

Improvement in GBE predictions was observed between the ML and RS models. A generalized model development approach was created to be adaptable and remove human bias. A dataset of macroscopic descriptors related to the impact of GBs on mechanical properties and corresponding GBEs was used. Decision-tree-based models were chosen for their performance and their potential for backwards interpretability. GPU-enabled training allows for highly efficient and accurate model creation.

Multiple decision-tree-based models (namely Random Forest, Extra Trees, and CatBoost) were able to be generated with relatively high accuracy. Although the distribution of GBEs trends fairly narrow, removing descriptors to prevent multicollinearity broadens the range of values significantly. These values can be generated

within kinetic Monte Carlo simulations to improve accuracy of grain growth simulations and inform GB effects on processing.

6. References

1. Katoh Y, Snead LL. Silicon carbide and its composites for nuclear applications – Historical overview. Journal of Nuclear Materials. 2019;526:151849.
2. Capan I. 4H-SiC Schottky barrier diodes as radiation detectors: a review. Electronics. 2022;11:532.
3. Miftasani F, Su'ud Z, Irwanto D, Permana S. Comparison of neutronic aspects in high-temperature gas-cooled reactor using ZrC and SiC Triso particle with 50 and 100 MWt power. International Journal of Energy Research. 2022;46:4852–4868.
4. Reed FK, Goetz KC, Ericson MN, Sweeney DC, Bull Ezell ND. Wide bandgap semiconductors for extreme temperature and radiation environments. Oak Ridge National Lab. (ORNL). 2022; Report No.: ORNL/TM-2021/2274.
5. Vargas-Gonzalez L, Speyer RF, Campbell J. Flexural strength, fracture toughness, and hardness of silicon carbide and boron carbide armor ceramics. International Journal of Applied Ceramic Technology. 2010;7:643–651.
6. Tanaka H. Sintering of silicon carbide. Springer Netherlands; 1991. p. 213–238.
7. Chen D, Sixta M, Zhang X, De Jonghe L, Ritchie R. Role of the grain-boundary phase on the elevated-temperature strength, toughness, fatigue and creep resistance of silicon carbide sintered with Al, B and C. Acta Materialia. 2000;48:4599–4608.
8. Jang CW, Kim J, Kang SJL. Effect of sintering atmosphere on grain shape and grain growth in liquid-phase-sintered silicon carbide. Journal of the American Ceramic Society. 2002;85:1281–1284.
9. Wang X et al. Radiation-induced segregation in a ceramic. Nature Materials. 2020;19:992–998.
10. Watari K et al. Effect of grain boundaries on thermal conductivity of silicon carbide ceramic at 5 to 1300 K. Journal of the American Ceramic Society. 2003;86:1812–1814.
11. Mitchell JA et al. Parallel simulation via SPPARKS of on-lattice kinetic and Metropolis Monte Carlo models for materials processing. Modelling and Simulation in Materials Science and Engineering. 2023;31:055001.

12. Hillert M. On the theory of normal and abnormal grain growth. Acta Metallurgica. 1965;13:227–238.

13. Humphreys FJ, Hatherly M. Recrystallization and related annealing phenomena. Elsevier; 2004.

14. Hernández–Rivera E, Goins PE, Murdoch HA. Anisotropic grain growth modeling under the SPPARKS framework. Army Research Laboratory (US). 2018 Sep; Report No.: ARL-TR-8507.

15. Shockley W, Read WT. Quantitative predictions from dislocation models of crystal grain boundaries. Physical Review. 1949;75:692–692.

16. Kohler C. Atomistic modelling of structures of tilt grain boundaries and antiphase boundaries in β-silicon carbide. Physica Status Solidi (b). 2002;234:522–540.

17. Wojdyr M, Khalil S, Liu Y, Szlufarska I. Energetics and structure of <0 0 1> tilt grain boundaries in SiC. Modelling and Simulation in Materials Science and Engineering. 2010;18:075009.

18. Banadaki AD, Tschopp MA, Patala S. An efficient Monte Carlo algorithm for determining the minimum energy structures of metallic grain boundaries. Computational Materials Science. 2018;155:466–475.

19. Guziewski M, Banadaki AD, Patala S, Coleman SP. Application of Monte Carlo techniques to grain boundary structure optimization in silicon and silicon-carbide. Computational Materials Science. 2020;182:109771.

20. Trujillo D, Guziewski M, Coleman S. Machine learning for predicting properties of silicon carbide grain boundaries. DEVCOM Army Research Laboratory (US). 2019 Nov; Report No.: ARL-TN-0981.

21. Montes de Oca Zapiain D, Guziewski M, Coleman SP, Dingreville R. Characterizing the tensile strength of metastable grain boundaries in silicon carbide using machine learning. The Journal of Physical Chemistry C. 2020;124:24809–24821.

22. Guziewski M, Montes de Oca Zapiain D, Dingreville R, Coleman SP. Microscopic and macroscopic characterization of grain boundary energy and strength in silicon carbide via machine-learning techniques. ACS Applied Materials & Interfaces. 2021;13:3311–3324.

23. Owens CB et al. Feature engineering descriptors, transforms, and machine learning for grain boundaries and variable-sized atom clusters. npj Computational Materials. 2025;11:21.

24. Huang SJ, Adityawardhana Y, Sanjaya J. Predicting mechanical properties of magnesium matrix composites with regression models by machine learning. Journal of Composites Science. 2023;7:347.

25. Yousefzadeh R, Kazemi A, Al-Maamari RS. Application of power-law committee machine to combine five machine learning algorithms for enhanced oil recovery screening. Scientific Reports. 2024;14:9200.

26. ONNX Runtime developers. ONNX Runtime. (version 1.22.0) [software]. 2021; URL: https://onnxruntime.ai/.

27. Thompson AP et al. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. Computer Physics Communications. 2022;271:108171.

28. Pastewka L, Klemenz A, Gumbsch P, Moseler M. Screened empirical bond-order potentials for Si-C. Physical Review B. 2013;87:205410.

29. Kohyama M, Kose S, Yamamoto R. Theoretical study of polar interfaces of the (122) Sigma =9 grain boundary in cubic SiC. Journal of Physics: Condensed Matter. 1991;3:7555–7573.

30. Mason J, Schuh C. Hyperspherical harmonics for the representation of crystallographic texture. Acta Materialia. 2008;56:6141–6155.

31. Grimmer H. Disorientations and coincidence rotations for cubic lattices. Acta Crystallographica Section A. 1974;30:685–688.

32. Maneewongvatana S, Mount DM. Analysis of approximate nearest neighbor searching with clustered point sets. arXiv; 1999; URL: http://arxiv.org/abs/cs/9901013.

33. Guziewski M, Coleman SP, Weinberger CR. Atomistic investigation into interfacial effects on the plastic response and deformation mechanisms of the pearlitic microstructure. Acta Materialia. 2019;180:287–300.

34. Guziewski M, Coleman SP, Weinberger CR. Interface energetics and structure of the pearlitic microstructure in steels: An atomistic and continuum investigation. Acta Materialia. 2018;155:1–11.

35. Shi J, Zikry M. Grain–boundary interactions and orientation effects on crack behavior in polycrystalline aggregates. International Journal of Solids and Structures. 2009;46:3914–3925.

36. Kheradmand N, Knorr AF, Marx M, Deng Y. Microscopic incompatibility controlling plastic deformation of bicrystals. Acta Materialia. 2016;106:219–228.

37. Kikuchi H et al. Brittle dynamic fracture of crystalline cubic silicon carbide (3C-SiC) via molecular dynamics simulation. Journal of Applied Physics. 2005;98:103524.

38. Kedharnath A, Kapoor R, Sarkar A. Classical molecular dynamics simulations of the deformation of metals under uniaxial monotonic loading: a review. Computers & Structures. 2021;254:106614.

39. George A. Dislocation dynamics, yield stresses and fracture: from elemental semiconductors to some structural ceramics. Journal de Physique III. 1991;1:909–927.

40. Pérez R, Gumbsch P. Directional anisotropy in the cleavage fracture of silicon. Physical Review Letters. 2000;84:5347–5350.

41. Pérez R, Gumbsch P. An ab initio study of the cleavage anisotropy in silicon. Acta Materialia. 2000;48:4517–4530.

42. Dingreville R, Aksoy D, Spearot DE. A primer on selecting grain boundary sets for comparison of interfacial fracture properties in molecular dynamics simulations. Scientific Reports. 2017;7:8332.

43. Siebörger D, Knake H, Glatzel U. Temperature dependence of the elastic moduli of the nickel-base superalloy CMSX-4 and its isolated phases. Materials Science and Engineering: A. 2001;298:26–33.

44. Lambrecht WRL, Segall B, Methfessel M, Van Schilfgaarde M. Calculated elastic constants and deformation potentials of cubic SiC. Physical Review B. 1991;44:3685–3694.

45. Pedregosa F et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.

46. Ali M. PyCaret: an open source, low-code machine learning library in Python. (version 1.0.0) [software]. 2020; URL: https://www.pycaret.org.

47. Breiman L. Random Forests. Machine Learning. 2001;45:5–32.

48. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019;Anchorage, AK.

49. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning. 2006;63:3–42.

50. Louppe G. Understanding Random Forests: from theory to practice. arXiv; 2014; URL: https://arxiv.org/abs/1407.7502.

51. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. arXiv; 2017; URL: https://arxiv.org/abs/1706.09516.

52. Zhang H, Zimmerman J, Nettleton D, Nordman DJ. Random Forest prediction intervals. The American Statistician. 2020;74:392–406.

Appendix. Impact of Summer Research Experience

Background

Alex Hauck graduated with a dual B.S. in Nuclear Engineering and Materials Engineering from Rensselaer Polytechnic Institute in May 2022. He has been pursuing a Ph.D. in Nuclear Engineering with a Computational Materials Doctoral Minor at The Pennsylvania State University, advised by Dr. Miaomiao Jin and Dr. Blair Tuttle.

Alex performs computational research, primarily atomistic modeling using first-principles techniques and scientific programming and code development with Python. In 2020, he participated in the Westinghouse Fellowship Program at The Pennsylvania State University and utilized machine learning to perform automatic helium bubble identification for TEM-images of irradiated Ni alloys. After graduating in 2022, he joined the NEN-2 (Advanced Nuclear Technology) group at Los Alamos National Laboratory to generate high-fidelity MCNP models of the Flattop critical assembly using Python. His Ph.D. project encompasses numerous ab initio and atomistic modeling techniques to study the effects of radiation-induced defect generation within the wide-bandgap semiconductor materials GaN, AlN, and their ternary alloy system Al–Ga–N.

Impact

Within this project, I have expanded my knowledge of machine learning techniques and applied them to another wide-bandgap semiconductor material, SiC. I have learned about various types of regression models including how to train them utilizing GPUs on HPC systems. Utilizing higher-scale modeling techniques to study grain boundaries increases the number of tools in my computational toolbox and further prepares me to advance my scientific career. This work has also enabled me to begin learning C++ and apply it to scientific challenges.

List of Symbols, Abbreviations, and Acronyms

3-D/4-D	Three-/Four-Dimensional
APB	Antiphase Boundary
CSL	Coincident Site Lattice
EBSD	Electron Backscatter Diffraction
GB	Grain Boundary
GBE	Grain Boundary Energy
GPU	Graphical Processing Unit
HAGB	High-Angle Grain Boundary
MAE	Mean Absolute Error
MAPE	Mean Absolute Percent Error
ML	Machine Learning
MSE	Mean Squared Error
ONNX	Open Neural Network Exchange
RIS	Radiation-Induced Segregation
RMSE	Root Mean Squared Error
RMSLE	Root Mean Squared Logarithmic Error
RS	Read–Shockley
SiC	Silicon Carbide
SOAP	Smooth Overlap of Atomic Positions
SPPARKS	Stochastic Parallel PARticle Kinetic Simulator
TRISO	TRi-structural ISOtropic

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLB CB TECH LIB