ARL-TN-0981 ● Nov 2019

DEVCOM ARMY RESEARCH LABORATORY

Machine Learning for Predicting Properties of Silicon Carbide Grain Boundaries

by Dennis Trujillo, Matthew Guziewski, and Shawn Coleman

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-TN-0981 ● Nov 2019

Machine Learning for Predicting Properties of Silicon Carbide Grain Boundaries

Dennis Trujillo
Department of Materials Science and Engineering, University of Connecticut

Matthew Guziewski and Shawn Coleman
Weapons and Materials Research Directorate, CCDC Army Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
November 2019	Technical Note	1 May–30 September 2019

4. TITLE AND SUBTITLE
Machine Learning for Predicting Properties of Silicon Carbide Grain Boundaries

5a. CONTRACT NUMBER
W911NF-16-2-0008

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Dennis Trujillo, Matthew Guziewski, and Shawn Coleman

5d. PROJECT NUMBER
HIP-19-009

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
CCDC Army Research Laboratory
ATTN: FCDD-RLW-ME
Aberdeen Proving Ground, MD 21005-5068

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TN-0981

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
US Army Engineer Research and Development Center
HPCMP Internship Program
3909 Halls Ferry Road, Vicksburg, MS 39180

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
ORCID ID(s): Dennis Trujillo, 0000-0001-9259-3744; Matthew Guziewski, 0000-0002-5761-720X; Shawn Coleman, 0000-0002-5542-3161;

14. ABSTRACT
Statistical techniques are utilized to determine the efficacy of physics-based descriptors to predict the energetic properties of silicon carbide grain boundaries. These descriptors are utilized in kernel ridge regression models with a radial basis function kernel for the prediction of grain boundary energetics. Models derived from this approach have been implemented as a replacement to the insertion, removal, and replacement probability functions in a Monte Carlo–based selection scheme for sampling the microscopic degrees of freedom in silicon carbide grain boundaries. Preliminary results show these models increase the overall computational efficiency of finding low-energy minimized states compared to current techniques.

15. SUBJECT TERMS
machine learning, atomistic simulation, grain boundary, Monte Carlo, optimization

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Shawn Coleman
Unclassified	Unclassified	Unclassified	UU	26	19b. TELEPHONE NUMBER (Include area code) (410) 306-0697

Contents

List of Figures

List of Tables

Preface

Dennis Trujillo is a third-year PhD student in materials science and engineering at the University of Connecticut. He holds a Bachelor of Science degree in physics awarded by New Mexico State University. Current research interests include atomistic simulations of surfaces and interfaces via classical molecular dynamics and density functional theory for the design of novel materials exhibiting enhanced chemical, electronic, and mechanical properties. Likewise, the application of machine learning and deep learning techniques for materials discovery and computer vision problems is of great interest and the subject of multiple concurrent studies conducted by Trujillo. Trujillo intends to complete his PhD in December 2020 and afterward will pursue a postdoctoral position that matches his research interests.

This work serves as a proof of concept for the design of a machine-learning-based framework for the prediction of energetic properties of bicrystal interfaces. This is one of the first attempts made to predict energetic properties of a grain boundary interface in a multi-element system, as known to the authors at the time of writing. Implementing these models in a Monte Carlo framework to sample the microscopic degrees of freedom in a grain boundary interface (replacing the standard energy-based partition function) is a novel application of a machine-learning-derived model. This work stands as a successful application of descriptor-based machine learning models in a novel and relevant engineering problem. This work fulfills the original project goals, which included deriving a machine-learning-based approach to optimizing Monte Carlo–based site selection for silicon carbide grain boundaries.

This internship has provided Trujillo the ability to combine machine learning models and experiment with his own ideas regarding structural descriptors for predicting energetic properties of materials. The skills expanded on through this project will aid the remainder of his PhD, as well as potential postdoc or employment opportunities.

Acknowledgments

This research was sponsored by the High-Performance Computing Modernization Program (HPCMP) and the HPC Internship Program (HIP-19-009). Dennis Trujillo acknowledges Shawn Coleman and Matt Guziewski for their mentorship.

1. Introduction

Designing ceramic composite systems that exhibit enhanced hardness and fracture toughness is a key consideration in creating next-generation armor systems. Interfaces (e.g., grain boundaries and phase boundaries) are one component of the material design space due to their effects on the macroscopic mechanical properties.[1] Atomistic simulations are often exploited to investigate the structure–property relationships in individual interfaces constructed from bicrystal models. Many prior works have limited their studies to thermodynamic stable interface structures as they are more likely to occur naturally or with a greater frequency in experimental growth conditions.[2] However, in designing ceramic composites, both thermodynamically stable and metastable states must be considered due to their active role in brittle fracture and failure.[3] Srolovitz et al. determined via Monte Carlo methods that the multiplicity of metastable grain boundaries is indeed extensive and employed a statistical mechanics framework to predict finite temperature equilibrium and nonequilibrium physical properties.[4] Therefore, an improved method to systematically study interfaces using atomistic simulations is needed in order to design ceramic composite.

Grain boundary orientations are generally described by five macroscopic degrees of freedom that illustrate the relative orientation of the grains and the alignment of the interface plane. However, at the microscale, there are countless microscopic degrees of freedom that can result in different local atomic interface structures. As a result, the sampling of all combinations of these microscopic degrees of freedom can rapidly increase the computational cost for even a single grain boundary. Prior researchers have approached the local optimization of interfaces using translational search techniques,[5] evolutionary/genetic algorithms,[6] and Monte Carlo–based sampling.[7] While each of these methods proved useful for their associated studies, the scope of these works was often limited to studying single-component metals or only exploring ideal thermodynamically stable interfaces. Monte Carlo–based optimization, however, has recently shown promise in both exploring multi-element ceramics and extracting relevant metastable interface structures.[8]

During Monte Carlo–based optimization of ceramic interfaces,[8] structures are probed by pseudo-randomly inserting, removing, or replacing individual atoms within the interface regions. To more efficiently probe likely favorable states, the randomness of these operations is biased based on probabilities tuned by the user at the start of the search. Once an operation type is chosen, the location of the operation is determined based on probabilities defined by the local structure. After each operation, the interface structure is relaxed using a three-step process of quenching, equilibration, and minimization within the framework of classical

molecular dynamics. Energetically favorable operations are accepted using a Boltzmann weighted probability, which enables efficient sampling of metastable interface structures along the way to the energetic minimum. Monte Carlo optimization of a single interface structure will often involve thousands of interface operations and produce hundreds of accepted metastable states. The amount of data generated in these studies opens up opportunities for using machine learning methods to accelerate sampling.

In this work, a machine learning model is constructed to cheaply predict the change in the interface energy after an individual Monte Carlo operation. This prediction can help accelerate the overall interface optimization by replacing the hand-tuned operation probabilities with one that takes into account the current system configuration. This work focuses on silicon carbide grain boundary data obtained from classical atomistic simulations coupled with descriptor-based machine learning. The implementation of a statistical (via Pearson correlation) and regression-based scheme (kernel ridge regression [KRR]) to predict the energetic properties of grain boundaries in silicon carbide is discussed in the remainder of this report.

2. Materials and Methods

2.1 Monte Carlo Interface Optimization

Monte Carlo optimization routines for atomic interfaces were first developed by Banadaki et al. for single-element metals[9] and expanded by Guziewski[8] for multi-element ceramic systems.[8] In these works, the initial configuration is perturbed by performing an insertion, removal, or atomic species replacement operation in the interface region. If the operation is energetically favorable (i.e., decreases relative to the original state), the state is accepted as the new grain boundary. If the energy of the state is less energetically favorable than the initial state, then a Boltzmann weighted probability function is utilized to determine acceptance or rejection of the perturbed state.

To more efficiently explore multi-element interfaces, Guziewski[8] allowed the user to hand-tune the probability of each operation type based on their understanding of the system. For example, in the silicon carbide system used in this study, the insertion, removal, and replacement operation probabilities were weighted at 50%, 25%, and 25%, respectively. Once the operation type is decided, the specific site for the operation to occur was chosen based on additional probability functions that account for the local structure. Descriptions of the probability functions that

Guziewski used to identify the individual operation site are included in Appendix A.

After each insertion, removal, and replacement operation, the atomic structure is relaxed using a three-part process before evaluating its energetic favorability. The relaxations included a quench, equilibration, and minimization at 0 K. Quenching involves applying a Nose–Hoover thermostat[10] and Parrinello–Rahman barostat[11] at 0.9 T_m, and dropping the temperature to a minimum temperature of 5 K over 500 time steps. Equilibration involves maintaining a 5 K thermostat and 0-Pa barostat over an additional 500 time steps to allow the atoms and simulation domain to evolve further. The final minimization procedure involves utilizing a conjugate gradient minimization of the atomic forces and energy at 0 K over an additional 500 time steps. The intense 1500-step relaxation increases the overall computational time of each operation; however, it was determined to be necessary to reduce the interface energy and excess interfacial strain.

During Monte Carlo optimization of the interface, thousands of operations are tried in the exhaustive search for the minimum energy structure. While costly, the amount of data explored makes it possible to train machine learning models, which can potentially accelerate the Monte Carlo–based sampling techniques. Specifically, this work develops machine learning models using local structural descriptors near the Monte Carlo operation site to predict the final relaxed energy after the proposed operation. If successful, this machine learning model will simplify and speed up a future Monte Carlo interface optimization algorithm by simultaneously predicting the operation type and site of maximum likelihood to reduce the interface energy.

2.2 Data Utilized

Data for this work were obtained from individual Monte Carlo optimization steps examining a series of 150 silicon carbide symmetric tilt grain boundaries exhibiting (100) and (110) tilt axes with varying degrees of misorientation. The data contain specific structural information related to the region around each Monte Carlo operation step (insertion, removal, or replacement) and the system energies after the three-part relaxation. In total, 959,296 unique Monte Carlo operation steps were captured. For each operation step, seven atomic metrics are evaluated at the operation site as well as the four closest neighboring atoms to provide 35 structural descriptors to describe the local region. The atomic descriptors used in this study provide both energetic and structural information about the local environment at the operational site and are listed in Table 1. Because it is unlikely that one machine learning model would be applicable for all operations, the data set was discretized

based on both the operation type (insert, remove, or replace) as well as the atomic species operated upon (silicon [Si] or carbon [C]). These subsets are then used to train individual models. The subsets are separated into training and test sets with a 0.9/0.1 split that was initially chosen at random. Throughout this study, silicon carbide was modeled using classical atomistic descriptions predicted by the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS)[8] using the modified Tersoff potential developed by Kohler.[12]

Table 1 Descriptors utilized in building predictive models via KRR

Atomic descriptor	Description
Energy (eng)	Per-atom energy of site atom (eV)
Common-neighbor parameter (CNP)	Common neighbor parameter of site atom
Voronoi volume (vol)	Voronoi volume of site atom or void ($Å^3$)
Si–Si bonds	Number of Si–Si bonds on site atom
Si–C bonds	Number of Si–C bonds on site atom
C–C bonds	Number of C–C bonds on site atom
Nearest-neighbor distance	Distance to N^{th} nearest neighbor from site atom (Å)

All structural metrics used in this study are fairly inexpensive to compute, so that the models built from these descriptors could potentially reduce the overall computational cost. Additionally, these particular metrics were also chosen to account for various factors that often contribute to grain boundary energy. The energy metric is particularly useful as it is the atomic energy metric for the local configuration as determined by the interatomic potential, and often the removal of high-energy atoms can reduce the energy of the system. The common neighbor parameter (CNP) can be used to identify defects and provides a measure of the variation of the local crystal structure around an atom to that of the bulk, which would be the lowest energy configuration. Voronoi volume and nearest-neighbor distance relate to the amount of space available, or conversely the local strain that can be accommodated. Nearest-neighbor distances in particular also give insight into bond lengths in the local neighborhood. Lastly, the total number of bonds of each type provides a measure of the local stoichiometry associated with each atom.

The endpoint property being tracked for each Monte Carlo operation is the change in the system's grain boundary energy after the three-part relaxation. This metric is used as it provides a means of deriving a probabilistic relationship based on energy (i.e., the probability of the operation being accepted is related to the change in energy associated with the operation, as described in Appendix A).

2.3 Pearson Correlation for Model Development

The importance of descriptors is evaluated by its Pearson correlation with the endpoint variable. Perfect correlation is observed at values of 1, anticorrelation is observed at values of –1, and no correlation is observed when the value is equal to 0. Correlation maps for each operational data set are provided in Figs. 1–3 for insertion, removal, and replacement operations, respectively.

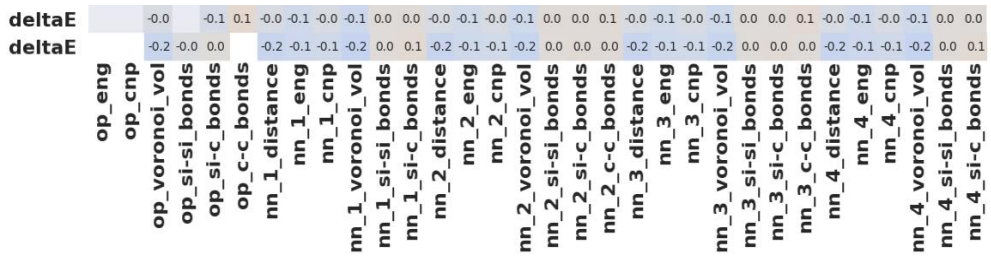

Fig. 1 Pearson correlation map for data representing a C atom (top) and a Si atom inserted (bottom) at the interface. Descriptors for the operation site do not contain values for eng or CNP as they are unoccupied, while the vol corresponds to the void size and the number of neighbors to the atomic species of the atoms at the void vertices.

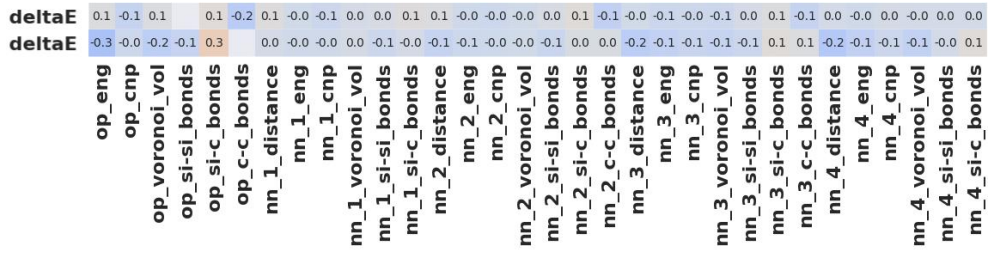

Fig. 2 Pearson correlation map for data representing a C atom (top) and a Si atom (bottom) removed at the interface. Descriptors for the number of Si–Si bonds are empty for C removal as there are no Si–Si bonds present. Likewise, the Si removal data contain no information relative to the number of C–C bonds.

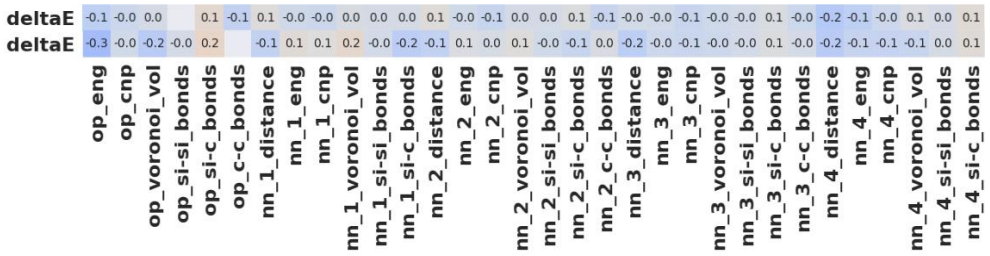

Fig. 3 Pearson correlation map for data representing a C atom replaced with a Si atom (top) and a Si atom replaced with a C atom (bottom) at the interface. Descriptors for the number of Si–Si bonds are empty for C replacement are empty as there are no Si–Si bonds present. Likewise, the Si replacement data contain no information relative to the number of C–C bonds as there is no C atom at the operational site to form C–C bonds.

As shown, weak correlation/anti-correlation was observed between all descriptors and the endpoint, as all values were near zero. This could be due in part to the measurement of the linear relationship defined by the Pearson correlation. Ultimately, these results highlight the difficulty in predicting the proper operation and site based on just a single descriptor and suggests that utilizing a combination of these descriptors in the framework of a nonlinear model is needed to provide a useful regression-based model for predicting energetic properties.

2.4 Kernel Ridge Regression

A KRR technique, as implemented in the scikit-learn package[13] (see Appendix B for a description of the derivation), is employed to develop a nonlinear model for the description of change in energy from an initial to a perturbed structure (*deltaE*) in the framework described by Guziewski and Banadaki et al., respectively.

Hyperparameters (α, γ) were determined via a discrete grid search,[13] which utilized the cross-validated mean absolute error (MAE) as the metric. The MAE is defined accordingly, as

$$MAE = \frac{\sum_{i=1}^{n} |actual_i - predicted_i|}{n}, \quad (1)$$

where the difference between the actual and predicted values for each sample is taken into consideration.

We considered the following table of hyperparameters (Table 2) and optimized the models for each respective operational data set using a five-fold cross-validated MAE as the metric. The test set errors, however, are a measure of the ability of the model to predict the endpoint for a given descriptor set for samples that were not included in the training set. In general, the resulting models had similar error values when applied to both training and test set data, which indicate that overfitting is not an issue within the models.

Table 2 Train and test scores for optimized models determined via a grid search over KRR hyperparameters (α, γ). All models found the same set of hyperparameters to be optimal $(\alpha = 1e^{-6}, = 1e^{-6})$.

Operation	Training MAE (J/m^2)	Test MAE (J/m^2)
Insert C	0.207	0.209
Insert Si	0.305	0.311
Remove C	0.183	0.191
Remove Si	0.184	0.192
Replace C	0.257	0.264
Replace Si	0.181	0.175

The plots for insertion, removal, and replacement of a C or Si atom at the interface (Figs. 4–6) are 2-D histogram representations of the model prediction versus actual value. The insertion models (Fig. 4) illustrate a particularly interesting case where the efficacy of model's ability to predict values accurately was purposely influenced by biasing the training data set. Specifically, a greater number of training data were included in the range *deltaE* < 0.2 J/m^2 to increase the ability of the model to predict in this domain. This was motivated after observing the poor performance of models built with an unbiased representation of the data relative to the performance of the other operational models. Despite biasing, the insertion models typically underpredict *deltaE* and had the highest MAE when compared to all other operations (C insert: 0.209 J/m^2, Si insert: 0.311 J/m^2).

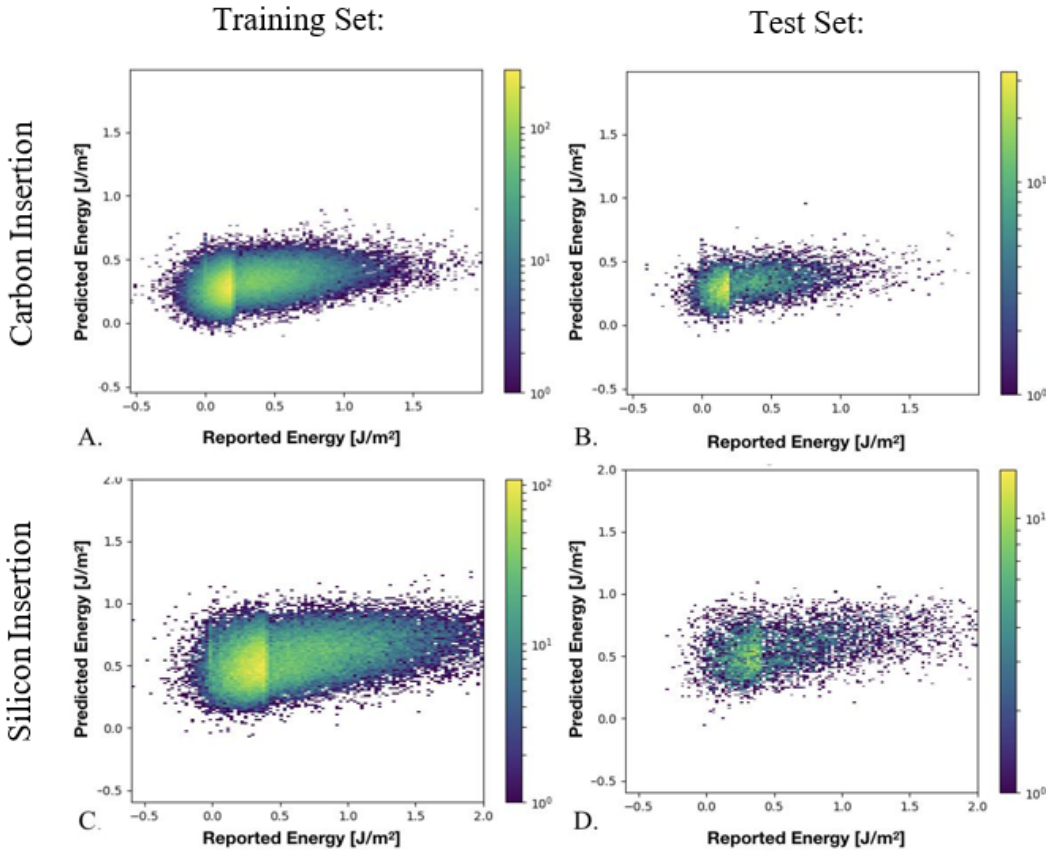

Fig. 4 Predicted vs. true plots illustrating the ability of the KRR model for predicting *deltaE* for data representing an atom inserted at the grain boundary. A and C correspond to training data, while those B and D correspond to test data. A and B are for C atoms, while C and D is representing Si atom insertion.

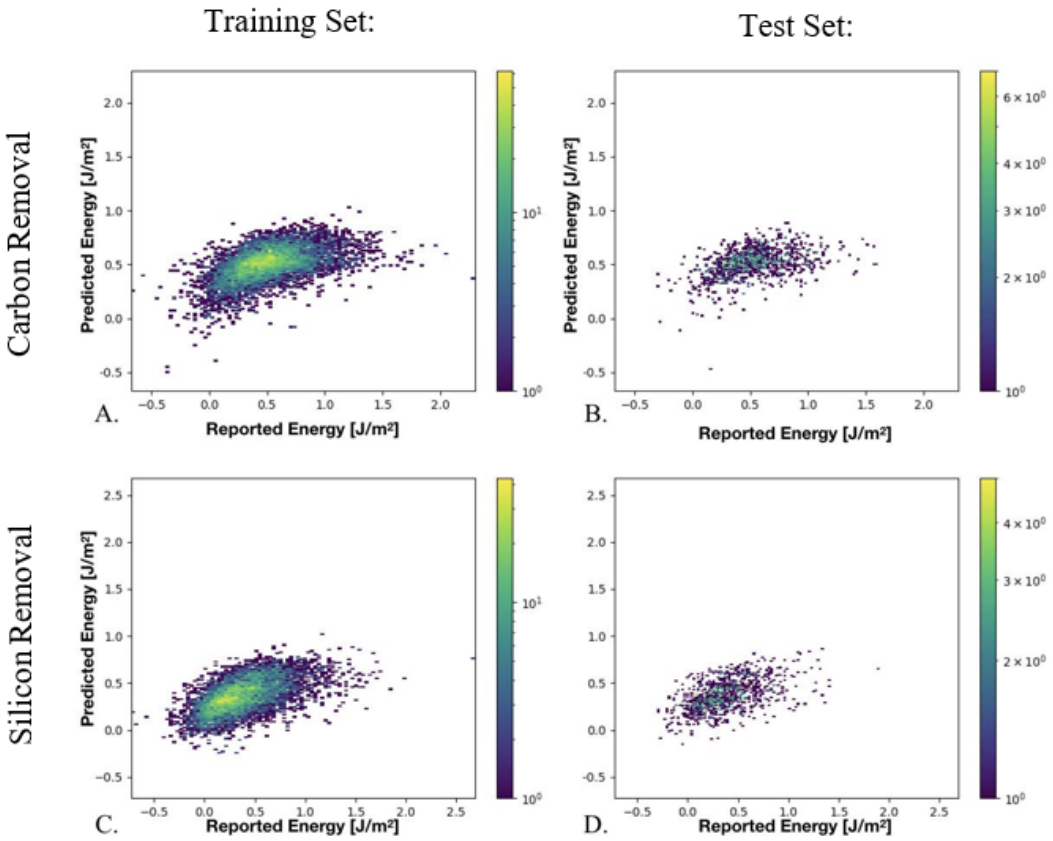

Fig. 5 Predicted vs. true plots illustrating the ability of the KRR model for predicting *deltaE* for data representing removal at the interface. A and C correspond to training data, while B and D correspond to test data. A and B are for C atoms, while C and D represent Si atoms removal.

Models for predicting *deltaE* in removal (Fig. 5) and replacement operations (Fig. 6) utilize all the data in a nonbiased manner and show relatively good performance with a lower amount of underprediction. Models developed for removal operations display the greatest performance overall with a test set MAE of 0.191 and 0.192 J/m^2 for C and Si removal, respectively. Replacement models performed better in Si replacement as compared to C replacement operations, respectively (0.175 vs. 0.264 J/m^2), indicating the descriptors for Si replacement better represent the system energetics. Replacement and removal operations used the full operational data set without bias as there was relatively good performance overall in terms of predicting *deltaE* over the full domain of data.

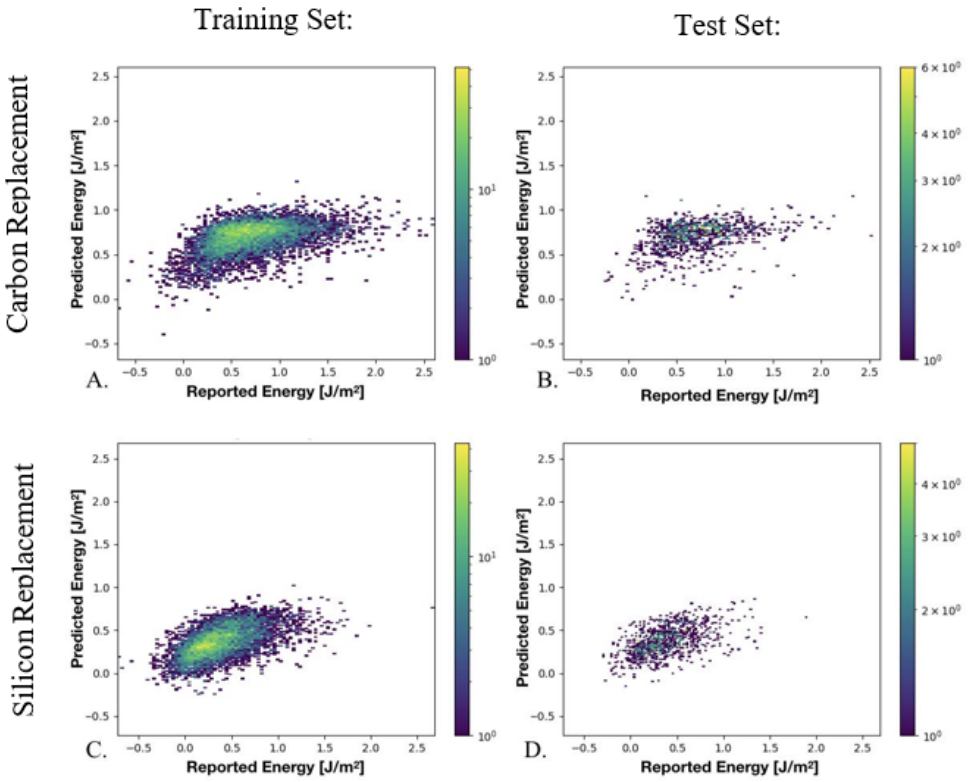

Fig. 6 Predicted vs. true plots illustrating the ability of the KRR model for predicting *deltaE* for data representing atom replacement at the interface. A and C correspond to training data, while B and D correspond to test data. A and B are for C atoms, while C and D represent Si atom replacement.

2.5 Implementation of Models in Monte Carlo–based Selection

The KRR models were implemented into the Monte Carlo grain boundary optimization code in place of the user-based probability functions as a potentially more accurate and efficient means of determining potential low-energy operation sites. Initial comparison of the two approaches was performed on three silicon carbide tilt grain boundaries: Σ9 [110] (122), Σ7 [111] (123), and Σ5 [100] (120). For comparison, the Monte Carlo algorithm was run for 1500 operations using both the user-based and KRR model–based operation probabilities. The resulting minimum grain boundary energy (GBE) was predicted, as well as any speed-up that accounts for the relative number of steps for the new machine language models to outperform the previous models. The results in Table 3 show that there is significant improvement using the KRR models. The KRR models were able to speed up the Monte Carlo optimizations by 6.79 to 9.68 times. Even more surprisingly, even lower GBE structures were found using the KRR models than the user-based probabilities. This initial implementation shows promise for extension to other systems with the potential for significantly minimizing the number of

computational steps required to minimize a system and producing a more energetically favorable minimized state.

Table 3 Comparison of physics-based model and model developed in this work

Grain boundary	User-based GBE (J/m^2)	Speed up	ML model GBE (J/m^2)
Σ9 [110] (122)	2.17	6.79	1.94
Σ7 [111] (123)	2.43	8.08	2.01
Σ5 [100] (120)	2.65	9.68	2.20

3. Conclusions

A descriptor-based approach to predicting energetic properties for silicon carbide grain boundaries was effectively designed and implemented. It was determined that structurally relevant descriptors can be used as a viable descriptor set within a nonlinear model framework to build accurate models to predict the change in energy given an insertion, removal, or replacement operation performed at the interface of a silicon carbide bicrystal.

Even though the correlation between the descriptors and the endpoint was minimal, it was still possible to build representative models that exhibited an acceptable amount of error. The implementation of these models in place of the traditional probability functions has shown promise in terms of reducing the total number of steps necessary to minimize a grain boundary structure and has yielded lower energy minimized states. These results are highly encouraging and provide a proof of concept, which can be utilized in similar works with a comparable descriptor space and endpoint.

4. References

1. Zhang JY, Liu G, Sun J. Strain rate effects on the mechanical response in multi-and single-crystalline Cu micropillars: grain boundary effects. International Journal of Plasticity. 2013;50:1–17.
2. Trahanovsky ME. Bicrystal-array fabrication [thesis]. [Berkeley (CA)]: University of California; 2012.
3. Han J, Vitek V, Srolovitz DJ. Grain-boundary metastability and its statistical properties. Acta Materialia. 2016;104:259–273. https://doi.org/10.1016/j .actamat.2015.11.035
4. Kumar N, Choudhuri D, Banerjee R, Mishra RS. Strength and ductility optimization of Mg–Y–Nd–Zr alloy by microstructural design. International Journal of Plasticity. 2015;68:77–97.
5. Olmsted DL, Foiles SM, Holm EA. Survey of computed grain boundary properties in face-centered cubic metals: I. Grain boundary energy. Acta Materialia. 2009;57(13):3694–3703.
6. Zhu Q, Samanta A, Li B, Rudd RE, Frolov T. Predicting phase behavior of grain boundaries with evolutionary search and machine learning. Nature Communications. 2018;9(1):467.
7. 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.
8. Guziewski M, Banadaki AD, Patla S, Coleman SP. Application of Monte Carlo techniques to grain boundary structure optimization in silicon and silicon carbide. Computational Materials Science. Forthcoming 2019.
9. Plimpton S. Fast parallel algorithms for short-range molecular dynamics. Journal of Computational Physics. 1995;117(1):1–19.
10. Hoover WG. Canonical dynamics: equilibrium phase-space distributions. Physical Review A. 1985;31(3):1695.
11. Nosé S. A unified formulation of the constant temperature molecular dynamics methods. The Journal of Chemical Physics. 1984;81(1):511-519.
12. Tersoff J. Modeling solid-state chemistry: Interatomic potentials for multicomponent systems. Physical Review B. 1989;39(8):5566.
13. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research. 2011;12(Oct):2825–2830.

Appendix A. Description of Operation Probabilities

For insertions, the probability function deciding the insert site and inserted species is based on structure. The insert site location probability ($p_{in,i}$) is derived by taking the difference between the Delaunay polyhedral radius (r_s) of the potential site and the bulk crystal (r_0), where the Delaunay polyhedra are generated via a Voronoi tessellation-based approach, where N_v is the number of atoms composing the Voronoi tessellation:

$$p_{in,i} = \frac{r_{s,i} - r_0}{\sum_{j=1}^{N_v}(r_{s,j} - r_0)}. \quad \text{(A-1)}$$

From this equation, atoms the largest interstitial volume site is the highest likely location for an atom insertion. The species inserted is chosen using another probability function that ensures that the local chemistry best represents the bulk-like structure based on its neighboring atoms. For example, in bulk silicon carbide, every one of the four neighbors identified by the Delaunay polyhedral should be the opposite chemical species. To reflect this zincblende structure, locations with four surrounding silicon (Si) atoms will give an 80% probability for inserting carbon (C) and Si 20%. On the other extreme, if the four surrounding atoms were C, the probability for inserting Si will be 80% and C 20%. The species probabilities for insertions vary linearly between these two based on the neighbors.

For removal and replacements, two probability functions ($p_{r,i}^{I}$ and $p_{r,i}^{II}$) were equally used interchangeably to determine the atomic operation site. The first is based on an excess energy formulation ($p_{r,i}^{I}$):

$$p_{r,i}^{I} = \frac{E_i - E_0}{\sum_{j=1}^{N_{GB}}(E_j - E_0)}. \quad \text{(A-2)}$$

Here the individual atomic energy of the potential operation site is E_i, the atomic energy of a corresponding bulk atom is E_0, and N_{GB} is the number of atoms in the grain boundary region. From this equation, high-energy atoms have the highest probability to be chosen for the removal or replacement operation. In multi-element systems, however, the excess energy formulation can often be biased toward one chemical species.

Therefore, a second probability function for removal and replacements was developed based on structure. In this work, the common-neighbor parameter[1] serves as a useful metric to distinguish between bulk-like atoms and atoms near the interface. A structural probability function for removal and replacement operations

[1] Tsuzuki H, Branicio PS, Rino JP. Structural characterization of deformed crystals by analysis of common atomic neighborhood. Computer Physics Communications. 2007;177(6):518–523.

($p_{r,i}^{II}$) is then defined using the common-neighbor parameter for interfacial (S_i) and bulk like (S_0) atoms:

$$p_{r,i}^{II} = \frac{S_i - S_0}{\sum_{j=1}^{N_v}(S_j - S_0)} \quad \text{(A-3)}$$

From this equation, atoms whose structures were furthest from bulk-like configurations are highest in probability for removal or replacement. To most effectively probe the configurational space removal and replacements operations, we switched between energetic and structural.

Appendix B. Formulation of Kernel Ridge Regression

The generalized formulation for any kernel ridge regression (KRR) as implemented in our studies is discussed. The formulation of these models is similar to linear least squares although with nonlinear functions incorporated via a kernel function. These models are the simplest implementations of the "kernel trick" in machine learning frameworks.

Ridge regression-based models can be expressed using a minimization of the following expression containing $\{(\vec{x}, y)\}$, which represents the feature space and an endpoint:

$$J(w) = (y - Xw)^T(y - Xw) + \lambda\|w\|^2, \tag{B-1}$$

where the optimal solution is given as

$$w = (X^T X + \lambda I_D)^{-1} X^T y. \tag{B-2}$$

Rewriting in terms of inner products,

$$w = X^T(XX^T + \lambda I_D)^{-1} y. \tag{B-3}$$

Rewriting in terms of dual variables, where α is introduced as

$$\alpha = (K + \lambda I_D)^{-1} y \tag{B-4}$$

and K is the Gram matrix replacing XX^T, the weight vector can be written as

$$w = X^T \alpha = \sum_{i=1}^{N} \alpha_i x_i, \tag{B-5}$$

where the solution is a linear sum of N samples such that

$$f(\vec{x}) = \sum_{i=1}^{N} \alpha_j \kappa(x, x_i). \tag{B-6}$$

The $\kappa(\vec{x}_i, \vec{x}_j)$ term is the kernel function. We considered a radial basis function kernel defined accordingly,

$$K(x, y) = e^{-\gamma\|x-y\|^2}. \tag{B-7}$$

List of Symbols, Abbreviations, and Acronyms

C	carbon
CNP	common neighbor parameter of site atom
eng	per-atom energy
GBE	grain boundary energy
HPCMP	High-Performance Computing Modernization Program
KRR	kernel ridge regression
Si	silicon
vol	Voronoi volume of site atom or void

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	CCDC ARL FCDD RLD CL TECH LIB
5 (PDF)	CCDC ARL FCDD RLW ME S COLEMAN M GUZIEWSKI S SILTON FCDD RLW MG B RINDERSPACHER FCDD RLW LB B BARNES