Machine Learning Applied to the Selection of Stable Low-Energy Grain Boundaries

Dennis Trujillo,[1, a)] S. Pamir Alpay,[1, 2, b)] Shawn Coleman,[3, c)] and Matthew Guziewski[3, d)]
[1)] *Department of Materials Science & Engineering and Institute of Materials Science, University of Connecticut, Storrs, CT 06269*
[2)] *Department of Physics, University of Connecticut, Storrs, CT 06269*
[3)] *Weapons and Materials Research Directorate, CCDC Army Research Laboratory*

Statistical techniques are utilized to determine the efficacy of physics-based descriptors to predict energetic properties of silicon grain boundaries. These descriptors are utilized in random forest regression models for the prediction of grain boundary energetics. The model performance is compared to kernel ridge regression based and support vector regression based models to compare predictive ability across a variety of regression schemes. Models derived from this approach have been implemented as a replacement to the insertion and removal probability functions in a Monte Carlo based selection scheme for sampling the microscopic degrees of freedom in silicon grain boundaries. Preliminary results show these models increase the overall computational efficiency of finding low energy minimized states compared to current techniques.

I. INTRODUCTION

Polycrystalline silicon is an essential component in the current generation of solar cells, next generation batteries, and semiconductor devices. In particular the electronic properties of silicon and polycrystalline silicon have been the subject of many studies based on the application of p and n doped silicon in semiconductor devices. In addition, the electromechanical properties of silicon also prove immensely interesting given the observance of giant piezoresistance in silicon nanowires /citeHe2006. Interfaces such as grain boundaries and phase boundaries are a component of the material design space which have a significant effect on the electrical and mechanical properties of a polycrystalline material and thus should be evaluated appropriately so as to enhance the functional properties of a material /citeZhang2013.

Grain boundary orientations are generally defined by five macroscopic degrees of freedom that describe 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. Previous research has approached the local optimization of interfaces using translational search techniques /citeOlmstead2009, evolutionary/genetic algorithms /citeZhu2018, and Monte Carlo based sampling /citeBanadaki2018. While each of these methods proved relatively useful for a specific set of conditions that were investigated, the scope of these studies was often limited to the investigation of single component metals or to the exploration of ideal thermodynamic stable interfaces. Monte Carlo based optimization, however, has recently shown promise in both exploring polycrystalline ceramics and extracting relevant metastable interface structures /citeGuziewski2019.

During Monte Carlo based optimization of bicrystal 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 off 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 acceptable metastable states. This amount of data generated in these studies opens up opportunities for machine learning methods for accelerated sampling.

In this study, a machine learning model is constructed to efficiently predict the change in the interface energy after an individual Monte Carlo operation. Here we focus on silicon grain boundary data obtained from classical atomistic simulations coupled with descriptor-based machine learning. We show here this approach can accelerate the overall interface optimization by replacing the hand-tuned operation probabilities with one that takes into account the current system configuration. The implementation of a statistical and regression-based scheme for the prediction of the energetic properties of silicon grain boundaries is also discussed.

[a)] Electronic mail: dennis.trujillo@uconn.edu
[b)] Electronic mail: pamir.alpay@uconn.edu
[c)] Electronic mail: shawn.p.coleman8.civ@mail.mil
[d)] Electronic mail: matthew.c.guziewski.civ@mail.mil

II. MATERIALS AND METHODS

A. 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 et al. 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, then 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 et al. [8] allowed the user to hand-tune the probability of each operation type based off 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 off 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 I.

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 timesteps. Equilibration involves maintaining a 5K 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 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.

B. Data Utilized

Data for this work was 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 contains specific structural information related to region around each Monte Carlo operation step (insertion, removal, or replace) 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 dataset was discretized based on both the operation type (insert, remove, or replacement) as well as the atomic species operated upon (Si or 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 LAMMPS [8] using the modified Tersoff potential as developed by Kohler [12].

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 oftentimes the removal of high-energy atoms can reduce the energy of the system. The common neighbor parameter 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 local strain which 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 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 3-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 1.

C. 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 Figures 1-3 for insertion, removal, and replacement operations, respectively.

As shown, weak correlation/anti-correlation was observed between all descriptors and the endpoint, as all values were near zero. This could be in part to the measurement of 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 non-linear model is needed to provide a useful regression-based model for predicting energetic properties.

III. PRINCIPLE COMPNENT ANALYSIS

IV. RANDOM FOREST IMPORTANCE

V. ACKNOWLEDGEMENTS

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