THE JOURNAL OF PHYSICAL CHEMISTRY C

Article

Cite This: J. Phys. Chem. C 2019, 123, 20715−20722

pubs.acs.org/JPCC

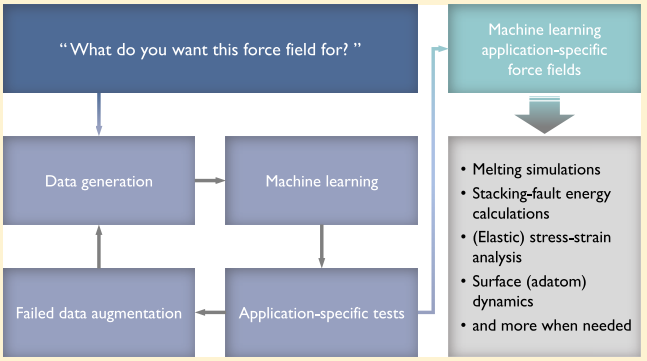
Iterative-Learning Strategy for the Development of Application-Specific Atomistic Force Fields

Tran Doan Huan,[†] Rohit Batra,[†] James Chapman, Chiho Kim, Anand Chandrasekaran, and Rampi Ramprasad*

School of Materials Science and Engineering, Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta, Georgia 30332, United States

Supporting Information

ABSTRACT: Emerging data-driven approaches in materials science have triggered the development of numerous machine-learning force fields. In practice, they are constructed by training a statistical model on a reference database to predict potential energy and/or atomic forces. Although most of the force fields can accurately recover the properties of the training set, some of them are becoming useful for actual molecular dynamics simulations. In this work, we employ a simple iterative-learning strategy for the development of machine-learning force fields targeted at specific simulations (applications). The strategy involves (1) preparing and fingerprinting a diverse reference database of atomic configurations and forces, (2) generating a pool of machine-learning force fields by learning the reference data, (3) validating the force fields against a series of targeted applications, and (4) selectively and recursively improving the force fields that are unsuitable for a given application while keeping their performance on other applications uncompromised. We demonstrate this strategy by developing a series of machine-learning-based Al and Cu force fields that can simultaneously be used for various applications, e.g., (elastic) stress/strain analysis, stacking-fault energy calculations, and melting simulations. This strategy is also generalizable, i.e., it may be used for other materials as well.



1. INTRODUCTION

In molecular dynamics (MD) simulations, physical properties of interest, e.g., melting temperature and dislocation density, are extracted from the MD trajectory, which is obtained by integrating Newton's equations of motion using the atomic forces evaluated by a given method.[1,2] Therefore, the fidelity of a MD simulation is rooted in the accuracy of the atomic force predictions. Because MD requires the forces to be evaluated at every single time step, an ideal method for this purpose should be accurate, fast, and transferable, so that it can easily be used for different systems in different conditions. Although first-principles-based methods like density functional theory (DFT)[3,4] are considered to be accurate and transferable, they are computationally prohibitive for realistic simulations in which large systems (composed of thousands to tens of thousands of atoms) are monitored over long-time scales (from nanoseconds to milliseconds). The force evaluation can be ≃4−6 orders faster if an empirical force field (FF) is used. Force fields of this class[5−10] are developed by fitting an assumed functional form to reference (experimental and/or computed) data to determine the predefined parameters. Therefore, the parametric settings (of an empirical force field) used for a material at a given condition are generally not applicable for other materials and/or at other conditions.

Data-driven methods, which recently joined the main stream of material science and engineering,[11−20] offer a potential alternative pathway for the development of force fields.[21−42] Unlike their traditional counterparts, machine-learning (ML) force fields do not have a well-defined functional form, or at least a physically meaningful one, that can be optimized. Instead, they are constructed by training a statistical model on a reference database to predict potential energy or atomic forces. This approach offers great flexibility in the development of ML force fields for different materials with minimal human interference. During the last several years, numerous ML force fields have been developed for Al,[25−28,34,42] C,[30,31,33] Na,[31,34] Li,[32] Si,[34−36] W,[37] Mo,[38] Fe,[39] Zr,[40] and SiO_2,[41] whereas some ML force field creation packages have also been released.[43] The developed ML force fields are generally very fast and accurate in reproducing the reference data, e.g., the force prediction errors of an adaptive, generalizable, and neighborhood informed (AGNI) ML FF can be as low as 0.01−0.10 eV/Å.[25−29] ML force fields are similarity-based, i.e., they generally may not work for applications (simulations) involving configurational domains that are different from

Received: May 3, 2019
Revised: August 7, 2019
Published: August 9, 2019

DOI: 10.1021/acs.jpcc.9b04207
J. Phys. Chem. C 2019, 123, 20715−20722

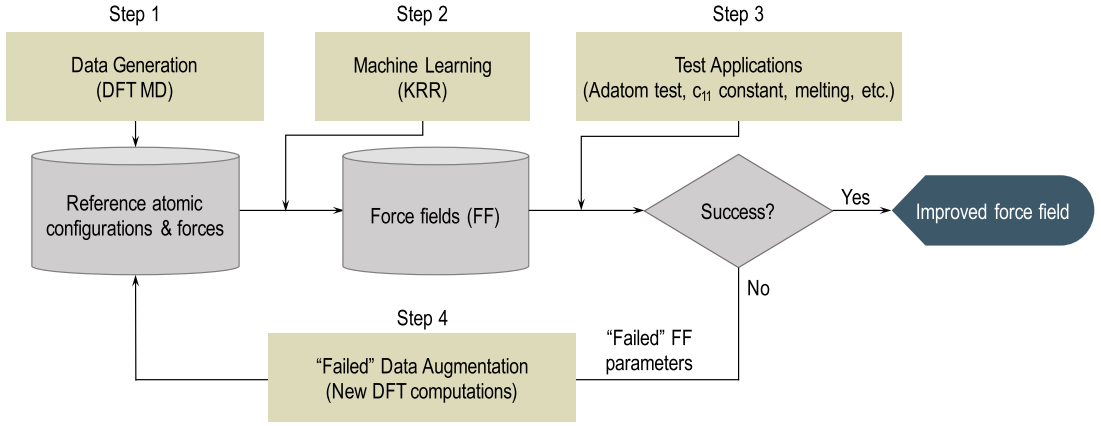


Figure 1. Iterative-learning strategy for AGNI force field generation.

Table 1. Summary of the Two References Data Sets Prepared for Al and Cu Force Field Generation[a]

config class	Al	Cu
defect-free bulk	face-centered cubic (fcc) and body-centered cubic (bcc) (with and without strain)	fcc and bcc (with and without strain)
point defects	supercell with one, two, six random vacancies	supercell with one, two, six random vacancies
planar defects	(100), (110), (111), (200), (333) surfaces; (111), (210), (310), (320), (510) grain boundaries; (111) stacking fault	(100), (110), (111), (200), (311), (333) surfaces; (111), (210), (310) grain boundaries; edge, screw dislocations; (100) with uniaxial strain
point and planar defects	adatom(s) on (100), (110) and (111) surfaces	adatom(s) on (100), (110) and (111) surfaces
clusters	radius 3, 5, 8, 10, 12 Å clusters	radius 5, 7 Å clusters, and cylinder
liquids	(111) slab at liquid state	(100) slab at liquid state

[a]Each of them is categorized into five types of configurations (configs) based on the types of defects contained. The number of configurations in each category excesses 1 000 000.

their training domain. However, the ML approach, unlike the traditional methods, does provide the opportunity to systematically improve the versatility of the force fields.

The present work demonstrates the last point by using an iterative-learning strategy to systematically improve the performance of an AGNI ML force field when being used for "new" applications (configurations). The key aspect of this strategy is "failed" data augmentation, during which a volume of data needed for a new application is systematically generated, identified, and appended to the original training data, and then the respective ML model is retrained on the updated training data. The proposed iterative-learning strategy can be used to systematically enlarge the domain of applicability of the developed ML force fields. We demonstrate this strategy by creating and validating AGNI ML force fields for Al and Cu solids, which can be used for a wide range of applications, including (elastic) stress/strain analysis, stacking-fault energy calculations, and melting simulations.

2. METHODS

The iterative-learning strategy used in this work aims specifically at upgrading a developed ML force field, making it suitable for targeted (new) applications, while keeping its current behavior uncompromised. This simple and generalizable strategy, visualized in Figure 1, starts with "reference data preparation" and "force field generation", details of which can also be found in our previous works.[25−27,29] The last two steps, i.e., "application-specific tests" and "failed data augmentation", are the essential components of the iterative-learning strategy.

The overall workflow of this strategy can be briefly summarized as follows. Starting from a diverse reference dataset of configurations and associated atomic forces (obtained in step 1), a pool of (≃100) low-root-mean-square error (RMSE) ML force fields is generated in step 2. In step 3, these force fields are subjected to a series of predefined application-specific tests. The force fields that pass a test are advanced to the next test, whereas those that fail are discarded. For situations, wherein none of the force field passes a test, the model that performed the best is selected for step 4, i.e., failed data augmentation. Two critical pieces of information are carried over from step 3 to step 4, including (1) the atomic configurations encountered during the unsuccessful test and (2) all of the ML parameters, e.g., the fingerprint parameters, the learning algorithm, and the training data, of the selected force field. Although the former is used to generate new reference data (by performing fresh computations using DFT in this work), the latter transfers the current performance of the original force field to those to be trained on the augmented training data. This procedure is repeated until a ML force field that successfully passes all of the application-specific tests is generated and can be confidently used for any of the applications they are designed for. Here, we showcase the validity of iterative-learning strategy by creating ML force fields for Al and Cu, which are capable of accurately predicting many static and dynamic properties. Details of each of the steps involved in the construction of ML force fields are discussed in the next sections.

This strategy is similar to the active learning approach for ML force fields developments[31,32,36,37] in terms of ideas but its end goal is different. Going beyond the intention of minimizing

DOI: 10.1021/acs.jpcc.9b04207
J. Phys. Chem. C 2019, 123, 20715−20722

force/energy prediction errors, the present work aims at developing FFs that can be used in realistic simulations, and a series of accompanied tests were designed for this purpose.

2.1. Reference Data Preparation. Two comprehensive reference data sets, summarized in Table 1, were prepared for Al and Cu in an accurate and consistent manner to minimize intrinsic numerical noise. DFT-based MD simulations, performed using the Vienna Ab initio Simulation Package (VASP),[44−47] were utilized to generate the reference atomic configurations and forces. Following ref 29, we used the Perdew, Burke, and Ernzerhof exchange−correlation functional[48] and the carefully calibrated energy cutoff, *k*-point mesh, and projection scheme[29] to ensure numerical convergence in calculating the atomic forces with VASP.

Diversity is critical for the reference data sets, i.e., they must contain the atomic environments that may be encountered during MD simulations. Therefore, for both Al and Cu, numerous atomic configurations were prepared: periodic or nonperiodic, crystalline or noncrystalline, and defect-free or having at least one of the regular (point, line, and planar) defects. Starting from these configurations, microcanonical DFT-based MD simulations were performed at various temperatures (up to ≃200% of the melting temperature T_m), thereby, providing atomic configurations and forces along the trajectories to form the reference data sets. Then, by rotating along the azimuthal and polar angles, the images of the original data were added to the reference data. The diversity of the data sets prepared for Al and Cu is then progressively augmented during step 4, wherein force fields are improved to handle new tests/applications. All of the reference data prepared for this work can be downloaded freely at https://khazana.gatech.edu/ .

2.2. Force Field Generation. Force field generation was performed in three substeps. First, the atomic configurations in the reference dataset are transformed into fingerprints, a numerical representation of the atomic environment.[21,22,25−28,37,49] Then, a given volume of training data is selected from the reference data using a sampling method. Finally, the training data is used to train the kernel ridge regression (KRR) model,[50,51] creating a force field, which maps the atomic fingerprints (or the atomic configurations) onto the atomic forces. All of these substeps are described in detail below.

2.2.1. Fingerprinting. Two related fingerprints, denoted by $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$, were used for our ML force fields. Their definitions are[22,25,26]

$$\mathbf{v}^{(1)}_{i,\alpha;k} = \sum_{j\neq i} \frac{r^{\alpha}_{ij}}{r_{ij}} \exp\left(-\frac{r^2_{ij}}{\eta^2_k}\right) f_c(r_{ij}) \quad (1)$$

and[29]

$$\mathbf{v}^{(2)}_{i,\alpha;k} = \sum_{j\neq i} \frac{r^{\alpha}_{ij}}{r_{ij}} \exp\left[-\frac{1}{2}\left(\frac{r_{ij}-a_k}{w}\right)^2\right] f_c(r_{ij}) \quad (2)$$

Here, $\mathbf{r}_i$ and $\mathbf{r}_j$ are the Cartesian coordinates of atoms i and j, $r_{ij} = |\mathbf{r}_j - \mathbf{r}_i|$, r^{α}_{ij} is the projection of $\mathbf{r}_j - \mathbf{r}_i$ onto the α Cartesian direction, and the sum runs over the neighbor list $\{j\}$ of atom i. The damping function $f_c(r_{ij}) = 1/2[\cos(\pi r_{ij}/R_c) + 1]$, where the cutoff radius R_c was chosen to be 8 Å, smoothly diminishes contributions from distant atoms. The key parameters of $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ are $\{\eta_k\}$ and $\{a_k, w\}$, respectively, which are determined by minimizing the force error in a separate step.

The key idea of our scheme is to map the fingerprint ($\mathbf{v}^{(1)}_{i,\alpha}$ or $\mathbf{v}^{(2)}_{i,\alpha}$) of an atom i along a particular direction α to its respective force component along the same direction α, rather than learning the entire force vector at once. It is derived from the fact that all three Cartesian directions x, y, and z (and, in fact, any other direction) are equivalent in terms of the relation between the local atomic environment and the respective atomic force. An advantage of this scheme is that the projections of atomic forces on the x, y, and z directions of any given coordinate frame can be learned and predicted independently. Moreover, information gained from one direction could be used to make predictions for another direction using the same ML force field. However, to allow the ML model to learn rotational invariance, explicit images of the original dataset after rotating along the azimuthal and polar angles[29] are also included in the referenced dataset. A detailed explanation on the above points is included in the Supporting Information. It should be noted that our approach is different from other techniques, e.g., symmetry-adapted kernels,[52,53] symmetry functions,[22] and smooth overlap of atomic positions,[49] in two ways. First, these approaches use fingerprints, which explicitly keep the atomic forces invariant with respect to rotations, allowing them to exclude the uncertainty stemmed from the stochastic nature of our scheme (discussed throughout this work).[52,53] Second, they use fingerprints with additional three-body angular terms, which improves the accuracy of the models. There is, however, trade-off in terms of the computational time required to compute these terms and an increase in the dimensionality of the fingerprint leading to increasing ML model complexity and difficult training process.

Generally, force fields using $\mathbf{v}^{(2)}$ have been found to be slightly better than those using $\mathbf{v}^{(1)}$ in several measures of force prediction errors, including RMSE.[29] A reliable force field based on $\mathbf{v}^{(1)}$ has been developed and tested for Al in the past,[27] whereas in this work, the iterative-learning strategy is demonstrated by developing three force fields, one using $\mathbf{v}^{(2)}$ for Al and the remaining two using $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ for Cu.

2.2.2. Sampling Method. A number of sampling methods for selecting training data from a big reference dataset were developed[27,29] and critically examined.[29,54] The primary criteria for evaluating these methods are various measures of force prediction errors, e.g., RMSE, which should be minimized. One of them, the grid sampling method, recently used to develop an Al force field,[27] was heavily used in the present study. In short, the grid sampling method involves (1) projecting the reference data onto a two-dimensional (2D) manifold spanned by two largest-variance principal components (determined by performing a principal component analysis), (2) defining a 2D grid in this manifold, covering the whole reference dataset, and (3) randomly selecting the training data from all of the grid cells.[27] This sampling method is intuitively expected to appropriately capture the diversity of the reference data and, hence, minimizing the force calculations errors (with respect to the reference data).[27] However, because all of the sampling methods[27,29] are stochastic in the nature, ML force fields trained on different training data are different not only in terms of force evaluation accuracy but also in terms of ability to capture the underlying physics of the materials. This is the reason why we create a pool of (∼100) force fields during the ML training process in step 2.

2.2.3. Learning Algorithm. Given a sampled training dataset of N_t atomic environments (or fingerprints) and forces, we

DOI: 10.1021/acs.jpcc.9b04207
J. Phys. Chem. C 2019, 123, 20715−20722

Table 2. Summary of the Tests Designed To Validate the Force Fields Developed for Al and Cu, Both of Which Adopt fcc as the Ground State[a]

order	test	description	test passed when
1	adatom stability	MD simulations of an adatom on the (111) surface	adatom is held and migrate on the surface for at least 100 ps
2	stress/strain	calculate the c_{11} coefficient and compare with experimental, DFT, and embedded-atom method (EAM) results	c_{11} is within ±5% of the experimental value.
3	phonon bands	calculate the fcc-lattice phonon band structures and compare with DFT and EAM results	force field-computed phonon band structure is closer to the DFT result
4	screw dislocation core structure	calculate and visualize the differential displacement map (DDM) and compare with DFT and EAM results	force field-computed DMM captures the symmetry and the order of the DFT-computed DDM
5	stacking-fault energies	compute (111) stacking-fault energy and compare with DFT and EAM results	force field-computed stacking-fault energy is within ±5% of the DFT value
6	slab melting	calculate the Lindemann index of (111) slab at various T and compare with experimental and EAM data	force field-computed T_m is comparable with the experimental and EAM results
7	adatom diffusion	probe the dynamics of an adatom on the (111) surface and estimate the hopping barrier via the Arrhenius plot	the estimated energy barrier is comparable to that computed by DFT

[a]LAMMPS was used for tests 1, 4, 5, 6, 7, whereas Atomic Simulation Environment[56] was used for tests 2 and 3. We note that the tests were carried out in the order shown in this table.

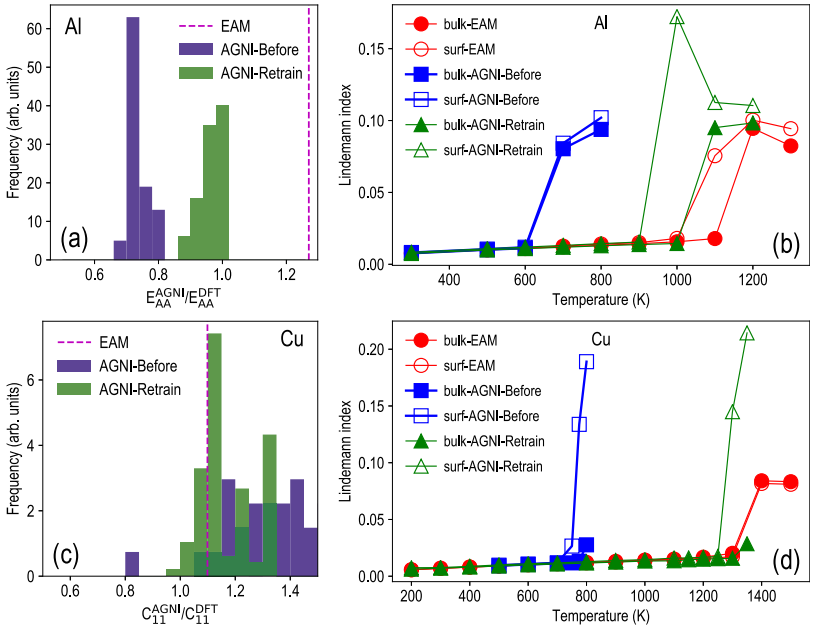


Figure 2. Role of iterative learning in the development of Al (a, b) and Cu force fields (c, d). In each panel, test results of the force fields before and after the failed data augmentation procedure are shown. Two tests for Al force fields are stacking-fault energies (a) and slab melting (b), whereas two tests for Cu force fields are stress/strain (c) and slab melting (d). In (b) and (d), the Lindemann index was computed separately for bulk and surface regions of the slab. For reference, the experimental melting temperatures of Al and Cu are 980 and 1350 K, respectively.

used KRR to construct the mapping between them. This similarity based-learning method has been used to develop numerous force fields,[25−29,35] one of which[27,28] has been distributed (and supported) within the 31Mar17 (or newer) version of Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS).[55] In a KRR-based ML force field, the atomic force F_μ corresponding to a configuration $\mu \equiv (i, \alpha)$ is predicted interpolatively from the reference atomic environments and forces provided in the training set using

$$F_\mu = \sum_{\nu=1}^{N_t} \alpha_\nu \exp\left[-\frac{1}{2}\left(\frac{d_{\mu\nu}}{\sigma}\right)^2\right] \quad (3)$$

Here, the sum runs over N_t reference environments, indexed by ν, of the training set, whereas $d_{\mu\nu}$ is the "distance" between configurations μ and ν, chosen to be the Euclidean norm, in the fingerprint space for which the length scale is specified by σ. During the training phase, the regression weights α_ν and the length scale σ are evaluated by optimizing a regularized objective function via fivefold cross-validation.

2.3. Application-Specific Force Field Tests. Following the reference data and the force field generation steps, a pool of accurate ML force fields with relatively similar (low) RMSE in force prediction was obtained. These force fields were then subjected to a series of application-specific tests. The idea behind these additional tests is that although the low errors (RMSE or other metrics) of a ML force field could advocate its general force prediction accuracy, they are not sufficient to guarantee that the force field indeed captures the physics of the system accurately. The reason for this uncertainty stems from the fact that all error measures, e.g., RMSE, are defined for very large numbers (often $\gtrsim 10^3-10^6$) of configurations, thus some configurations that are critical for a specific simulation can easily be suppressed during the training process when an error measure is minimized. Therefore, any newly developed ML force field should be validated against realistic and stringent

DOI: 10.1021/acs.jpcc.9b04207
J. Phys. Chem. C 2019, 123, 20715−20722

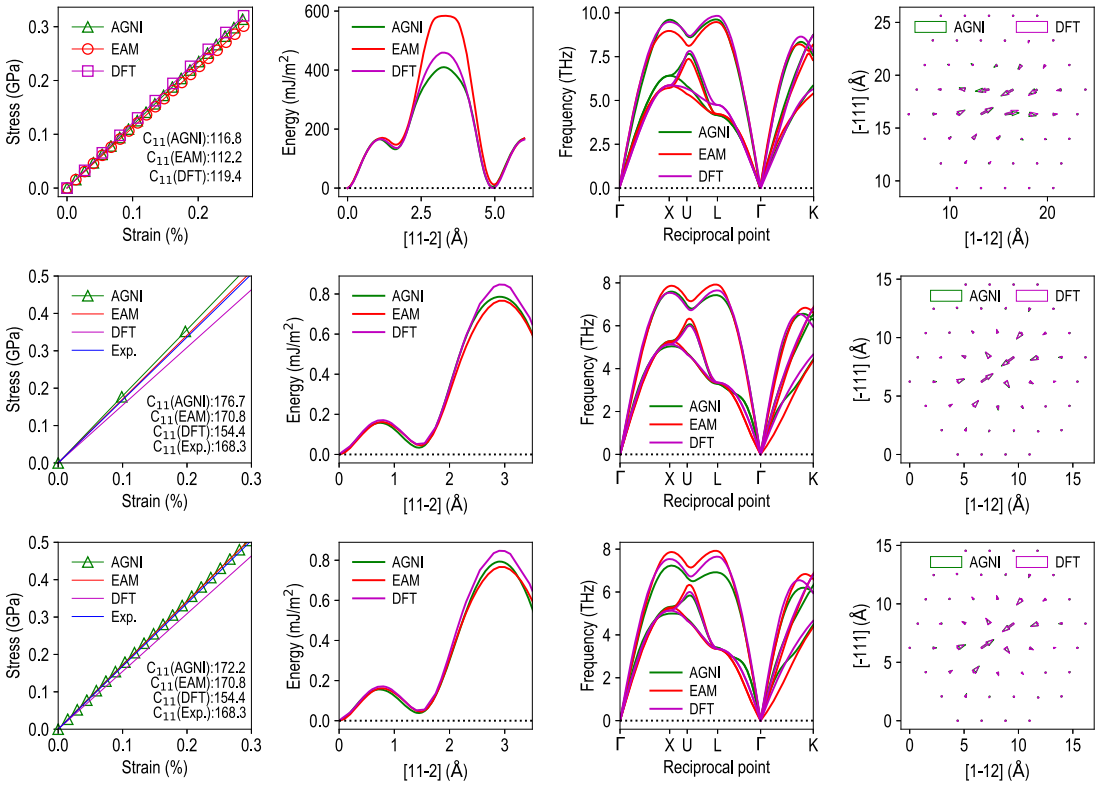


Figure 3. Validations of three force fields developed for Al (top row) and Cu (middle and bottom rows) against four tests, including (from left to right) stress/strain behavior, stacking-fault energy, phonon structure, and screw dislocation core structure. Technical details of the calculations can be found in the Supporting Information.

static and/or dynamic tests, in addition to having a small training error.[25−27,38,39]

In the present work, we compiled a list of seven carefully tailored application-specific tests, enumerated in Table 2. To pass a test, the ML force field must reasonably reproduce DFT and/or experimental results. The tests are conducted in an orderly manner based on their computational cost, i.e., time-intensive tests are generally conducted after the inexpensive tests, and only the force fields that pass a given test are advanced to the next test. For the purpose of quickly screening force fields during the development process, a short-time version of the (relatively stringent) "adatom diffusion" test, i.e., "adatom stability", was used to quickly discard those that cannot describe the (assumed) stability of an adatom on the slab surface. Table 2 summarizes all of the seven tests that were used for the force field development, whereas their technical details are provided in the Supporting Information.[57] It should be noted that the list of proposed tests can be easily expanded to make accurate and robust ML force fields targeted at new applications. This can be achieved using failed data augmentation process described in the next section.

2.4. Failed Data Augmentation. The stochastic nature of the sampling methods (discussed in Section 2.2.2) and the deficiency of the error measures (discussed in Section 2.3) may produce a training dataset in which critical atomic configurations for a particular application (simulation) of interest are under-represented. Consequently, the ML force fields created by training the KRR model on such data may perform poorly for that specific application since the force predictions are unreliable (and perhaps inaccurate) and can result in error accumulation during MD simulations. The iterative-learning strategy, of which the failed data augmentation is the most critical component, can be used to overcome this limitation and allow systematic improvements of ML force fields. This procedure involves (1) preparing new data corresponding to out-of-domain configurations extracted from the unsuccessful test, (2) sampling this newly generated data and augmenting (appending) the sampled data directly to the training set of the selected force field followed by a retraining step, and (3) testing the new generation of force fields to identify those that show actual improvement and pass the tests. In this work, we use an augmented data volume that is approximately 10−15% of the current training data, although the optimum volume should be determined by a more critical analysis. It can clearly be seen from Figure S2 (of the Supporting Information) that the augmented data, taken from the new data generated using the configurations of the unsuccessful test, significantly enlarges the domain covered by the training data of the failed force field. The performance of this procedure for the cases of Al and Cu is given in Figure 2.

Figure 2a shows the AA stacking-fault energy E^{AA} computed using DFT, EAM,[58] and two generations of ML force fields developed for Al, the second is generated by applying the failed data augmentation procedure for an "original" force field from the first generation. Since we have a pool of ML force fields from each generation, we look at the statistical nature of the improvement introduced by the failed data augmentation. Using the DFT-computed E^{AA}_{DFT} result as the reference, it can be seen that the majority of the "pre-augmentation" ML force fields underestimate E^{AA} by $\gtrsim$25% owing to the absence of corresponding environments during training, whereas the EAM potential used in this work overestimates E^{AA} by $\gtrsim$25%. However, when one of the best preaugmentation ML force fields was retrained on the new data augmented by atomic

DOI: 10.1021/acs.jpcc.9b04207
J. Phys. Chem. C 2019, 123, 20715−20722

Table 3. Melting Temperature T_m (Estimated from the Lindemann Index Computation) and the Surface Energy Barrier (Estimated from the Arrhenius Plot) Predicted Using the ML Force Field Developed[a]

		T_m (K)			ΔE (eV)		
materials	fingerprint	expr.	EAM	ML	DFT	EAM	ML
Al	$\mathbf{v}^{(2)}$	933	≃1100	≃1000	0.040	0.040	0.050
Cu	$\mathbf{v}^{(1)}$	1350	≃1300	≃1180	0.032	0.014	0.012
Cu	$\mathbf{v}^{(2)}$	1350	≃1300	≃1130	0.032	0.014	0.036

[a]The DFT value of ΔE was computed using the DFT-based nudged elastic band (NEB) method.

environments close to the AA stacking fault, significant improvements in the E^{AA} predictions can be observed clearly, i.e., more than 60% of "post-augmentation" force fields display less than 5% error in E^{AA}.

Similarly, c_{11} computed by DFT, EAM,[59] and two generations of ML force fields developed for Cu is shown in Figure 2c. Compared to DFT, the EAM potential used in this work overestimates c_{11} by roughly 10%, whereas c_{11} computed by pre-augmentation ML force fields ranges from 0.8 to 1.5 times the DFT result. The failed data augmentation procedure significantly improves the ML c_{11} results, as a majority of the post-augmentation force fields produce c_{11} within an error of 5%.

For the slab melting test, the Lindemann index q_i was computed to estimate the melting point of (111) slabs of fcc Al and Cu. The Lindemann index q_i, defined as the relative root-mean-square interatomic distance (r_{ij}) fluctuation[60]

$$q_i = \frac{1}{N-1}\sum_{i<j}\frac{\sqrt{\langle r_{ij}^2\rangle - \langle r_{ij}\rangle^2}}{\langle r_{ij}\rangle} \tag{4}$$

measures the thermal fluctuations of N atoms in the system. Near the melting point, q_i departs abruptly from a small and weakly linear regime for solids ($\sim 10^{-2}$ in order), jumping to $\sim 10^{-1}$ or higher for liquids. Since q_i is accessible within a MD simulation, this index has been widely used to estimate the melting point.[61−63] This test is computationally too expensive for DFT, thus only data from MD simulations using EAM and the developed ML force fields that pass the other tests are available. Computed q_i, shown in Figure 2b,d, clearly indicates that the EAM potential used herein captures the experimental melting temperature T_m of Al (933 K) and Cu (1350 K) quite well. The pre-augmentation ML force fields predict much lower melting temperatures, i.e., ≃600 K for Al and ≃800 K for Cu. However, by augmenting the training data with new environments generated for liquid Al and Cu (obtained during failed tests), the melting temperature predicted by the ML force fields was significantly improved, approaching the experimental values. We note that all our force fields are tested sequentially, and the improved force fields discussed next have passed all of the tests listed in Table 2, including the c_{11}, stacking fault, and slab melting test for which no initial version of ML force field successfully pass.

3. RESULTS

Using the demonstrated iterative-learning strategy, we have developed three new ML force fields, one for Al (using $\mathbf{v}^{(2)}$) and two for Cu (one uses $\mathbf{v}^{(1)}$ and the other uses $\mathbf{v}^{(2)}$). The parameters of these fingerprints were determined by minimizing the RMSE of the force predictions. For $\mathbf{v}^{(1)}$, we used 32 η_ks uniformly distributed in the log scale between 1 Å and R_c, whereas for $\mathbf{v}^{(2)}$, we used w = 0.3 Å and 32 a_ks uniformly distributed between 1 Å and R_c. Similar to the $\mathbf{v}^{(1)}$-based Al force field previously developed,[27] the $\mathbf{v}^{(1)}$-based Cu force field can also be used for MD simulations using the 31Mar17 (or newer) version of LAMMPS and can be downloaded freely from https://khazana.gatech.edu/.

These force fields have been validated against the set of realistic and stringent seven tests (Setion 2.3), with the results reported in Figure 3 and Table 3. Figure 3 presents the stress/strain curves, the stacking-fault energy, the phonon band structure, and the screw dislocation core structure computed using DFT, the EAM potentials, and the developed ML force fields. For the stress/strain curve and phonon band structure, all three ML force fields are better than the EAM potentials used in this work in reproducing the experimental/DFT results. In the other two tests, the ML force fields are comparable with EAM.

The melting temperature T_m (estimated by monitoring the Lindemann index) and the adatom diffusion barrier ΔE extracted from MD simulations, employing EAM and the ML force fields developed, are shown in Table 3. For reference, the experimental value of T_m and the DFT-based NEB computed ΔE are also provided. The melting temperature computed by using the ML force fields is comparable with that computed using the EAM potentials, whereas the adatom diffusion barrier ΔE is closer than the reference values.

Because the developed ML force fields learn and predict atomic forces directly but not as the derivatives of an energy, the energy conservation of an MD simulation is not obvious. The general trend is that the smaller the force prediction errors, the better the energy conservation. In fact, the achieved accuracy of the ML force fields in this class is sufficient for the energy, calculated as an integration along the MD trajectory, to be conserved, as demonstrated in refs 27 and 29 for NVT MD simulations using LAMMPS.

4. CONCLUSIONS

ML force fields are generally interpolative in the nature. Therefore, they are expected to behave correctly within the trained configuration domain, whereas for applications involving configurations outside this domain, they may be inadequate. To resolve this deficiency, we have used an iterative-learning strategy, systematically expanding the applicability of AGNI ML force fields, and making them suitable for targeted applications. This strategy was based upon a failed data augmentation procedure and a series of application-specific tests. Given a ML force field that is unsuitable for a specific simulation (application), relevant atomic configurations and forces are purposely augmented, improving the behavior of this force field while keeping its performance on other applications uncompromised. Using this strategy, we have developed three ML force fields, one for Al and two for Cu, that can simultaneously be used for various applications,

DOI: 10.1021/acs.jpcc.9b04207
J. Phys. Chem. C 2019, 123, 20715−20722

including (elastic) stress/strain analysis, stacking fault energy, and melting behavior simulations. We believe that this strategy is generalizable, using which ML force fields for other materials can also be developed. The reference data and force fields generated within this work are available at https://khazana.gatech.edu/.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpcc.9b04207.

All of the technical details of the force field application-specific tests we designed to validate the ML force fields developed (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rampi.ramprasad@mse.gatech.edu.

ORCID

Tran Doan Huan: 0000-0002-8093-9426
Rohit Batra: 0000-0002-1098-7035
James Chapman: 0000-0001-8451-0275
Chiho Kim: 0000-0002-1814-4980
Rampi Ramprasad: 0000-0003-4630-1565

Author Contributions

†T.D.H. and R.B. contributed equally to this paper.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The present work was financially supported by the National Science Foundation (grant number 1821992) and the Office of Naval Research (grant number N000141712148).

■ REFERENCES

(1) Alder, B. J.; Wainwright, T. E. Studies in Molecular Dynamics. I. General Method. *J. Chem. Phys.* **1959**, *31*, 459−466.
(2) Rahman, A. Correlations in the Motion of Atoms in Liquid Argon. *Phys. Rev.* **1964**, *136*, A405−A411.
(3) Hohenberg, P.; Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **1964**, *136*, B864−B871.
(4) Kohn, W.; Sham, L. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **1965**, *140*, A1133−A1138.
(5) Jones, J. E. On the determination of molecular fields. *Proc. R. Soc. A* **1924**, *106*, 463−477.
(6) Daw, M. S.; Baskes, M. I. Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals. *Phys. Rev. B* **1984**, *29*, 6443−6453.
(7) Daw, M. S.; Foiles, S. M.; Baskes, M. I. The embedded-atom method: a review of theory and applications. *Mater. Sci. Rep.* **1993**, *9*, 251−310.
(8) Tersoff, J. New empirical approach for the structure and energy of covalent systems. *Phys. Rev. B* **1988**, *37*, 6991−7000.
(9) Bazant, M. Z.; Kaxiras, E.; Justo, J. F. Environment-dependent interatomic potential for bulk silicon. *Phys. Rev. B* **1997**, *56*, 8542−8552.
(10) van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A. ReaxFF: A Reactive Force Field for Hydrocarbons. *J. Phys. Chem. A* **2001**, *105*, 9396−9409.
(11) Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem., Int. Ed.* **1993**, *32*, 503−527.
(12) Sumpter, B. G.; Getino, C.; Noid, D. W. Theory and Applications of Neural Computing in Chemical Science. *Annu. Rev. Phys. Chem.* **1994**, *45*, 439−481.
(13) Rajan, K. Materials informatics. *Mater. Today* **2005**, *8*, 38−45.
(14) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning and Materials Informatics: Recent Applications and Prospects. *npj Comput. Mater.* **2017**, *3*, No. 54.
(15) Mannodi-Kanakkithodi, A.; Chandrasekaran, A.; Kim, C.; Huan, T. D.; Pilania, G.; Botu, V.; Ramprasad, R. Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond. *Mater. Today* **2018**, *21*, 785−796.
(16) Kim, C.; Chandrasekaran, A.; Huan, T. D.; Das, D.; Ramprasad, R. Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. *J. Phys. Chem. C* **2018**, *122*, 17575−17585.
(17) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **2013**, *3*, No. 2810.
(18) Huan, T. D.; Mannodi-Kanakkithodi, A.; Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B* **2015**, *92*, No. 014106.
(19) Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine learning strategy for the accelerated design of polymer dielectrics. *Sci. Rep.* **2016**, *6*, No. 20952.
(20) Kim, C.; Pilania, G.; Ramprasad, R. From Organized High-Throughput Data to Phenomenological Theory using Machine Learning: The Example of Dielectric Breakdown. *Chem. Mater.* **2016**, *28*, 1304−1311.
(21) Behler, J. Perspective Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **2016**, *145*, No. 170901.
(22) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, No. 146401.
(23) Bartók, A. P.; Csányi, G. Gaussian Approximation Potentials: A Brief Tutorial Introduction. *Int. J. Quantum Chem.* **2015**, *115*, 1051−1057.
(24) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **2017**, *3*, No. e1603015.
(25) Botu, V.; Ramprasad, R. Learning scheme to predict atomic forces and accelerate materials simulations. *Phys. Rev. B* **2015**, *92*, No. 094306.
(26) Botu, V.; Ramprasad, R. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *Int. J. Quantum Chem.* **2015**, *115*, 1074−1083.
(27) Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine learning force fields: Construction, validation, and outlook. *J. Phys. Chem. C* **2017**, *121*, 511−522.
(28) Botu, V.; Chapman, J.; Ramprasad, R. A study of adatom ripening on an Al (111) surface with machine learning force fields. *Comput. Mater. Sci.* **2017**, *129*, 332−335.
(29) Huan, T. D.; Batra, R.; Chapman, J.; Krishnan, S.; Chen, L.; Ramprasad, R. A universal strategy for the creation of machine learning based atomistic force fields. *npj Comput. Mater.* **2017**, *3*, No. 37.
(30) Rowe, P.; Csányi, G.; Alfè, D.; Michaelides, A. Development of a machine learning potential for graphene. *Phys. Rev. B* **2018**, *97*, No. 054303.
(31) Podryabinkin, E. V.; Tikhonov, E. V.; Shapeev, A. V.; Oganov, A. R. Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning. *Phys. Rev. B* **2019**, *99*, No. 064114.
(32) Podryabinkin, E. V.; Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Comput. Mater. Sci.* **2017**, *140*, 171−180.
(33) Deringer, V. L.; Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B* **2017**, *95*, No. 094203.
(34) Kuritz, N.; Gordon, G.; Natan, A. Size and temperature transferability of direct and local deep neural networks for atomic forces. *Phys. Rev. B* **2018**, *98*, No. 094109.

DOI: 10.1021/acs.jpcc.9b04207
J. Phys. Chem. C 2019, 123, 20715−20722

(35) Suzuki, T.; Tamura, R.; Miyazaki, T. Machine learning for atomic forces in a crystalline solid: Transferability to various temperatures. *Int. J. Quantum Chem.* **2017**, *117*, 33−39.
(36) Bartók, A. P.; Kermode, J.; Bernstein, N.; Csányi, G. Machine Learning a General-Purpose Interatomic Potential for Silicon. *Phys. Rev. X* **2018**, *8*, No. 041048.
(37) Szlachta, W. J.; Bartók, A. P.; Csányi, G. Accuracy and transferability of Gaussian approximation potential models for tungsten. *Phys. Rev. B* **2014**, *90*, No. 104108.
(38) Chen, C.; Deng, Z.; Tran, R.; Tang, H.; Chu, I.-H.; Ong, S. P. Accurate force field for molybdenum by machine learning large materials data. *Phys. Rev. Mater.* **2017**, *1*, No. 043603.
(39) Dragoni, D.; Daff, T. D.; Csányi, G.; Marzari, N. Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron. *Phys. Rev. Mater.* **2018**, *2*, No. 013808.
(40) Zong, H.; Pilania, G.; Ding, X.; Ackland, G. J.; Lookman, T. Developing an interatomic potential for martensitic phase transformations in zirconium by machine learning. *npj Comput. Mater.* **2018**, *4*, No. 48.
(41) Li, W.; Ando, Y. Construction of Accurate Machine Learning Force Fields for Copper and Silicon Dioxide. 2018, arXiv:1807.02042. arXiv.org e-Print archive. http://arxiv.org/abs/1807.02042.
(42) Chapman, J.; Batra, R.; Uberuaga, B.; Pilania, G.; Ramprasad, R. A comprehensive computational study of adatom diffusion on the aluminum (100) surface. *Comput. Mater. Sci.* **2019**, *158*, 353−358.
(43) Wang, H.; Zhang, L.; Han, J.; Weinan, E. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **2018**, *228*, 178−184.
(44) Kresse, G.; Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **1993**, *47*, 558−561.
(45) Kresse, G. Ab initio Molekular Dynamik für flüssige Metalle. Ph.D. Thesis, Technische Universität Wien, 1993.
(46) Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **1996**, *6*, 15−50.
(47) Kresse, G.; Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **1996**, *54*, 11169−11186.
(48) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.
(49) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, No. 184115.
(50) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, 2009.
(51) Hofmann, T.; Schölkopf, B.; Smola, A. J. Kernel methods in machine learning. *Ann. Stat.* **2008**, *36*, 1171−1220.
(52) Glielmo, A.; Zeni, C.; De Vita, A. Efficient nonparametric *n*-body force fields from machine learning. *Phys. Rev. B* **2018**, *97*, No. 184307.
(53) Grisafi, A.; Wilkins, D. M.; Csányi, G.; Ceriotti, M. Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys. Rev. Lett.* **2018**, *120*, No. 036002.
(54) Jeong, W.; Lee, K.; Yoo, D.; Lee, D.; Han, S. Toward Reliable and Transferable Machine Learning Potentials: Uniform Training by Overcoming Sampling Bias. *J. Phys. Chem. C* **2018**, *122*, 22790−22795.
(55) Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **1995**, *117*, 1−19.
(56) Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; Hermes, E. D.; Jennings, P. C.; Jensen, P. B.; Kermode, J.; Kitchin, J. R.; Kolsbjerg, E. L.; Kubal, J.; Kaasbjerg, K.; Lysgaard, S.; Maronsson, J. B.; Maxson, T.; Olsen, T.; Pastewka, L.; Peterson, A.; Rostgaard, C.; Schiótz, J.; Schütt, O.; Strange, M.; Thygesen, K. S.; Vegge, T.; Vilhelmsen, L.; Walter, M.; Zeng, Z.; Jacobsen, K. W. The atomic simulation environment-a Python library for working with atoms. *J. Phys.: Condens. Matter* **2017**, *29*, No. 273002.
(57) See Supporting Information for more information reported in this paper.
(58) Zhou, X.; Wadley, H.; Johnson, R.; Larson, D.; Tabat, N.; Cerezo, A.; Petford-Long, A.; Smith, G.; Clifton, P.; Martens, R.; Kelly, T. Atomic scale structure of sputtered metal multilayers. *Acta Mater.* **2001**, *49*, 4005−4015.
(59) Mishin, Y.; Mehl, M. J.; Papaconstantopoulos, D. A.; Voter, A. F.; Kress, J. D. Structural stability and lattice defects in copper: Ab initio, tight-binding, and embedded-atom calculations. *Phys. Rev. B* **2001**, *63*, No. 224106.
(60) Stillinger, F. H.; Weber, T. A. Lindemann melting criterion and the Gaussian core model. *Phys. Rev. B* **1980**, *22*, 3790−3794.
(61) Alavi, S.; Thompson, D. L. Molecular Dynamics Simulations of the Melting of Aluminum Nanoparticles. *J. Phys. Chem. A* **2006**, *110*, 1518−1523.
(62) Neyts, E. C.; Bogaerts, A. Numerical Study of the Size-Dependent Melting Mechanisms of Nickel Nanoclusters. *J. Phys. Chem. C* **2009**, *113*, 2771−2776.
(63) Chavoshi, S. Z.; Xu, S.; Goel, S. Addressing the discrepancy of finding the equilibrium melting point of silicon using molecular dynamics simulations. *Proc. R. Soc. A* **2017**, *473*, No. 20170084.

DOI: 10.1021/acs.jpcc.9b04207
J. Phys. Chem. C 2019, 123, 20715−20722