npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

A universal strategy for the creation of machine learning-based atomistic force fields

Tran Doan Huan[1], Rohit Batra[1], James Chapman[1], Sridevi Krishnan[1], Lihua Chen[1] and Rampi Ramprasad[1]

Emerging machine learning (ML)-based approaches provide powerful and novel tools to study a variety of physical and chemical problems. In this contribution, we outline a universal strategy to create ML-based atomistic force fields, which can be used to perform high-fidelity molecular dynamics simulations. This scheme involves (1) preparing a big reference dataset of atomic environments and forces with sufficiently low noise, e.g., using density functional theory or higher-level methods, (2) utilizing a generalizable class of structural fingerprints for representing atomic environments, (3) optimally selecting diverse and non-redundant training datasets from the reference data, and (4) proposing various learning approaches to predict atomic forces directly (and rapidly) from atomic configurations. From the atomistic forces, accurate potential energies can then be obtained by appropriate integration along a reaction coordinate or along a molecular dynamics trajectory. Based on this strategy, we have created model ML force fields for six elemental bulk solids, including Al, Cu, Ti, W, Si, and C, and show that all of them can reach chemical accuracy. The proposed procedure is general and universal, in that it can potentially be used to generate ML force fields for any material using the same unified workflow with little human intervention. Moreover, the force fields can be systematically improved by adding new training data progressively to represent atomic environments not encountered previously.

npj Computational Materials (2017)3:37 ; doi:10.1038/s41524-017-0042-y

INTRODUCTION

The most direct computational strategy to monitor the atomic-level time-evolution of chemical and physical processes is by the molecular dynamics (MD) method. Starting with an initial atomic configuration and atomic velocities, MD simulations require the atomic forces as input to propagate the atoms and their velocities to the next time-step (at which point, the atomic forces are re-evaluated); the cycle continues, thus allowing for an iterative time-evolution of the system. The atomic forces at each time-step may be obtained either using quantum mechanics-based methods, such as density functional theory (DFT), or parameterized classical force fields.

A variety of DFT-based methods have already contributed to the rational design of application-specific materials.[1–3] Nevertheless, DFT-based MD simulations are not practical and routine at the present time, especially to track chemical processes with long time scales ($\gtrsim$1 nanosecond) and large length scales ($\gtrsim$10 nanometers). The repetitive and expensive DFT force computations during MD and the necessary small MD time steps (of the order of femtoseconds), lead to the primary bottlenecks of DFT-based MD simulations. Parameterized classical force fields (which are 6–10 orders of magnitude faster than DFT) may be used to access truly long time scales and large length scales. But, these approaches are not satisfactory either, as such force fields lack accuracy and versatility, i.e., they are not transferable to situations that were not originally used in the parameterization.

Owing to this scenario, monitoring the atomistic details underlying major classes of physical and chemical processes with high fidelity is still beyond the reaches of modern computational methods. Here, we attempt to take a step in remedying this situation via a universal data-driven atomic force prediction recipe that is as fast as classical force fields but as accurate and versatile as quantum mechanics-based methods.[4–9] Simply put, our hypothesis is that the force experienced by an atom is purely a function of the arrangement of the other atoms around it-a notion inspired, and originally suggested, by Feynman within the context of quantum mechanics, which has lead to the celebrated Hellmann–Feynman theorem.[10,11] If we are able to numerically represent this atomic arrangement, and if we have sufficient "atomic arrangement vs. force" examples, we should be able to train a machine to learn the arrangement-force relationship, and make future atomic force predictions based purely on atomic arrangement information.

Machine learning (ML) methods using neural networks, Gaussian processes, and other algorithms have been successful in the development of ML potentials for MD simulations.[12–24] The approach of the present contribution, namely, learning to predict atomic forces directly (from which the total potential energy of the entire system can be determined through appropriate integration), as previously introduced in ref. 25, is far more powerful. This is because the atomic force is a local quantity that can be formally defined for atoms, whereas the potential energy can be formally defined only for the entire system or unit cell. Other force fields that directly predict the potential energy define this quantity as a sum of atomic energies, for which a formal basis does not exist within quantum mechanical treatments.

Figure 1 portrays the essential steps involved in the creation of a ML force field, which are: (1) Generation of reference data, e.g., using DFT; (2) Fingerprinting the atomic environments, so that at

[1]Department of Materials Science & Engineering and Institute of Materials Science, University of Connecticut, 97 North Eagleville Rd., Unit 3136, Storrs, CT 06269-3136, USA
Correspondence: Rampi Ramprasad (rampi.ramprasad@uconn.edu)
Tran Doan Huan and Rohit Batra contributed equally to this work.

Received: 3 March 2017 Revised: 17 August 2017 Accepted: 28 August 2017
Published online: 18 September 2017

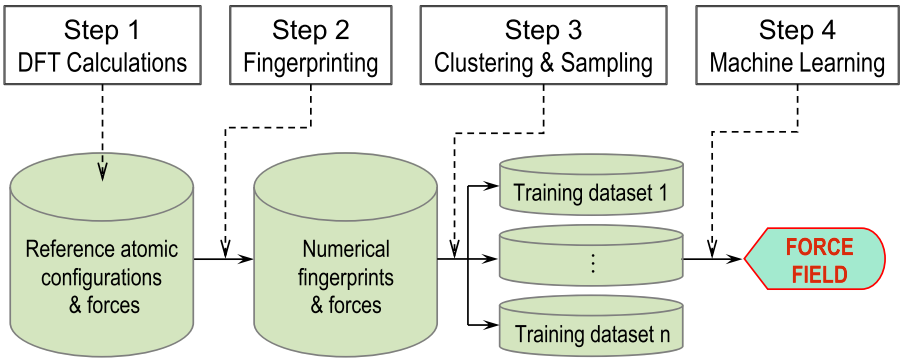

Fig. 1 Strategy for the creation of machine learning force fields

the end of this step, the data has been reduced to a set of numerical atomic fingerprints and the corresponding atomic forces; (3) Creation of compact "training sets" from this data through clustering and sampling; and (4) Learning from these datasets, such that at the end of this step, a robust mapping is established connecting fingerprints and forces. A rudimentary version of this procedure has been recently demonstrated, leading to a class of force fields we refer to as AGNI force fields. For the case of Al, several static and dynamic applications, including stress-strain behavior, melting, proper description of dislocation core regions, point defect diffusion in bulk, adatom organization (island formation) at surfaces, etc., were also explored with the AGNI force field.[4,5,7,8]

In the present contribution, we establish the true power and universality of this force field by applying it to a set of elemental solids, displaying a diversity of chemical bonding, including Al, Cu, Ti, W, Si, and C, encompassing metals and insulators occurring in a variety of different crystal structures. We also show that predictions of atomic forces and total potential energies (obtained through appropriate integration of the forces along an MD trajectory) reach unprecedented chemical accuracy levels for all cases considered, provided fundamental improvements to each of the four steps of Fig. 1 are made. For instance, the reference data should have sufficiently low intrinsic errors, and the fingerprinting scheme should have sufficient "resolving power" to represent the atomic configurations with high fidelity, carrying as much information about the atomic configuration as possible. Moreover, an important innovation of the present development is the clustering of the parent dataset so that different atomic environments are separated into different clusters, and each of them is sampled separately so that compact, non-redundant, and diverse "training sets" can be collected. The ML force field thus created can be improved progressively by including new configurations (i.e., new clusters), as required.

The proposed class of enhanced AGNI force fields can potentially (1) reach an arbitrary level of accuracy approaching that of the reference data, (2) be about 6–8 orders of magnitude faster than DFT calculations, (3) be systematically improved in quality, transferability and versatility by progressively adding new training set configurations, and (4) be generalized to any element (or combination of elements) of the periodic table for which reference (one-time) quantum mechanics-based calculations can be done. Below, we describe the key steps of this strategy and the results in detail.

RESULTS

Strategy

All four steps of the strategy, shown in Fig. 1, are crucial for the fidelity of enhanced AGNI force fields. In what follows, the critical aspects of these steps are discussed while the technical details can be found in section Methods.

Reference data preparation

The reference data used for creating AGNI force fields must be prepared accurately and consistently, ensuring sufficiently low intrinsic errors.[26–33] Except for the aforementioned computational cost, first-principles-based methods for computing atomic forces are considered to meet these requirements. Our reference data was prepared using a two-step procedure. First, atomic configurations and forces were obtained from several DFT-based MD simulations performed for bulk Al, Cu, Ti, W, Si, and C at a sufficient level of accuracy. Then, by rotating the collected atomic configurations and forces, new data was accumulated, making the reference data truly large in volume, and providing access to a lot more force components than in the original dataset.

Fingerprinting

A fingerprint is needed to numerically represent atomic configurations, especially for force field development.[5–8,13,15,19,20] Typically, a fingerprint must be invariant with respect to translations and rotations of the whole system, and to permutations of like atoms. To be useful for a ML force field, it must also be directionally resolved and continuous, i.e., proportionately change with small changes of the atomic arrangement. Our proposal is a d-dimensional vector $\mathbf{V}_{i,\alpha}$, representing the atomic environment of atom i viewed along the Cartesian α direction. For elemental materials, the k^{th} component ($k \in \{1, d\}$) of this vector is defined as[5,6,13]

$$V_{i,\alpha;k} = \sum_{j\neq i} \frac{r_{ij}^{\alpha}}{r_{ij}} \frac{1}{\sqrt{2\pi}w} \exp\left[-\frac{1}{2}\left(\frac{r_{ij}-a_k}{w}\right)^2\right] f_c(r_{ij}). \quad (1)$$

Here, $\mathbf{r}_i$ and $\mathbf{r}_j$ are the positions of atoms i and j, $r_{ij} = |\mathbf{r}_j - \mathbf{r}_i|$, r_{ij}^{α} is the projection of $\mathbf{r}_j - \mathbf{r}_i$ onto the α direction, and the sum runs over the neighbor list $\{j\}$ of atom i. In our strategy, $\mathbf{V}_{i,\alpha}$, schematically illustrated in Fig. 2, is used to learn (and to predict) the Cartesian α component of the atomic force exerted on atom i. The definition (1) can readily be generalized to multi-element materials.

By construction, $\mathbf{V}_{i,\alpha}$ satisfies the translation and permutation invariants. Although this fingerprint is directionally specific, the reference dataset is symmetrized by rotations, as described in the previous section of "Reference data preparation", thus the force field, i.e., the ML model established in the training set, can handle any orientation of the simulated system. Moreover, three factors included in each term of (1) are designed to handle those directly related to the ML force fields. First, r_{ij}^{α}/r_{ij} characterizes the contribution of atom j to the α component of the force on atom i. Next, the Gaussian factor $G(a_k, w; r) \equiv \frac{1}{\sqrt{2\pi}w}\exp\left[(r-a_k)^2/(2w^2)\right]$ captures, in a continuous manner, the

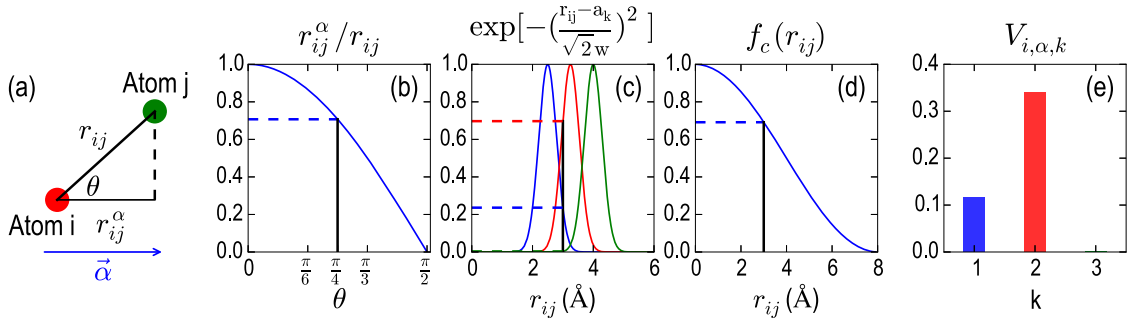

Fig. 2 The *k*-dimensional fingerprint vector, $\mathbf{V}_{i,\alpha}$, of reference atom *i*, (determined by the positions $\mathbf{r}_{ij}$ of other atoms *j* with respect to atom *i*), to be mapped to its α Cartesian force component. As shown in **a**, $\mathbf{r}_{ij}$ makes a θ angle with the α direction. In **b**, the first term of Eq. (1), i.e., $r^{\alpha}_{ij}/r_{ij} = \cos\boldsymbol{\theta}$, is represented. The distance $r_{ij} \equiv |\mathbf{r}_{ij}|$ is projected onto a series of Gaussian functions, yielding $G(a_k, w; r_{ij})$, as depicted in **c**. The damping factor $f_c(r_{ij})$ is represented in **d** and a schematic illustration of $\mathbf{V}_{i,\alpha}$ is given in **e**

possibility of atom *j* falling within the *k*th spherical shell centered at atom *i*. Finally, $f_c(r_{ij}) \equiv \frac{1}{2}\left[\cos\left(\frac{\pi r_{ij}}{R_c}\right)+1\right]$ gradually diminishes contributions from distant atoms, and setting them to zero when $r_{ij} > R_c$, a cutoff radius. The last two terms, namely $G(a_k, w; r)$ and $f_c(r_{ij})$, have been used for radial symmetry function, a similar fingerprint introduced to construct a neural-network potential.[13]

Starting from the same intuitive idea, a related fingerprint was recently defined,[5,6] leading to an early version of our AGNI ML force fields.[5–9] Instead of $G(a_k, w; r)$, $G(0, \eta_k; r) \equiv \exp(r^2/\eta_k^2)$ was used in the previo'us work to handle atoms *j* residing within the *k*th sphere centered at atom *i*. Consequently, the earlier fingerprint, denoted by $\mathbf{V}'_{i,\alpha}$, was defined as[5,6]

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

Two principal parameters of $\mathbf{V}_{i,\alpha}$ (the new fingerprint defined in (1)) are the dimensionality *d* and the width *w* of the Gaussian factors (although the $G(a_k, w; r)$s are not required to have the same *w*, we use this simplification). The particular distribution of $\{a_k\}_{k=1}^{d}$ is not important, especially for large *d*, given that the shells occupied by $G(a_k,w;r)$s cover the sphere of radius R_c from atom *i*. Similarly, *d* is the main parameter of $\mathbf{V}'_{i,\alpha}$ while the manner by which $\{\eta_k\}_{k=1}^{d}$ is distributed is of minor importance when *d* is large. The optimal parameters of $V_{i,\alpha}$ and $\mathbf{V}'_{i,\alpha}$ can be determined using the well-defined procedure described below. The fundamental difference between the two fingerprint schemes is explained by using a principal component analysis discussed in detail in the Supporting Information.

Sampling and clustering

An AGNI force field is a ML model established on a training dataset selected from the reference data. The training data should be compact and non-redundant, allowing the development and applications of the force fields to be time-efficient. However, the main assumption in training data selection is that the diversity of the reference data could be properly captured. In the context of generic ML, this work is non-trivial, especially when the original dataset is big and diverse.[34,35] Within the most common approach, the training data is selected randomly from the reference data.[5–9,27–29] Such training data is likely dominated by the configurations from the highly-populated domains while other domains are under-represented. This scenario is visualized in Fig. 3a, where the training data randomly selected from the reference data contains essentially no configuration with large-amplitude forces.

We propose several training data selection methods for better capturing the diversity of the reference data we prepared. In the *force-binning* method, we arrange the reference data into a number of force amplitude intervals and select training data from all the intervals, properly capturing the force amplitude profile of the reference data. In the clustering approach, the reference data is split into a given number of clusters in fingerprint space, from each of which a part of the training data is selected. Figure 3b, c visualize the improved diversity of the training set prepared by these two methods.

In fact, the clustering technique can potentially lead to a far more powerful concept in creating ML force fields. The current idea used for creating baseline AGNI ML force fields[5–9] and predicting material properties[27–31] is to use the predictive model (mapping) created on a single training set for the whole domain occupied by the reference data. The error of such a baseline ML force field grows unavoidably when the reference data becomes highly diverse, regardless of how big the training set size N_t is. Our approach, as sketched in Fig. 1, is that for each data cluster (domain), a separate training set is selected and then, a fingerprint-force mapping is established. By assembling these mappings, we obtain a ML force field, which is *domain-specific* in the sense that the atomic force of an atomic configuration is evaluated using the mapping created for the closest domain. Because any domain is significantly less diverse than the reference data, the domain-specific force fields can readily surpass the aforementioned intrinsic limit of the baseline ML force fields. A caveat though is that care must be taken to ensure that discontinuities are not introduced in the predicted atomic forces, say, during the course of an MD simulation, if the atomic environment shifts from one cluster to another. This aspect has not been explored in the present work. In the discussion below, all of the approaches will be critically examined.

Machine learning

Given a set of training data, a learning algorithm is needed to establish the fingerprint-force mapping. Here, we use kernel ridge regression (KRR),[36,37] a powerful method that has widely been used in materials informatics.[5–9,27–29,38] KRR predicts the atomic force F_i corresponding to the configuration *i* as

$$F_i = \sum_{j=1}^{N_t} a_j \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{\sigma}\right)^2\right]. \quad (3)$$

Here, the sum runs over N_t configurations, indexed by *j*, of the training set while d_{ij} is the "distance" between configurations *i* and *j*, chosen to be the Euclidean norm, in fingerprint space. The "length" scale in this space is specified by σ. During the training phase, the regression weights a_j and σ are evaluated by optimizing a regularized objective function via 5-fold cross-validation.

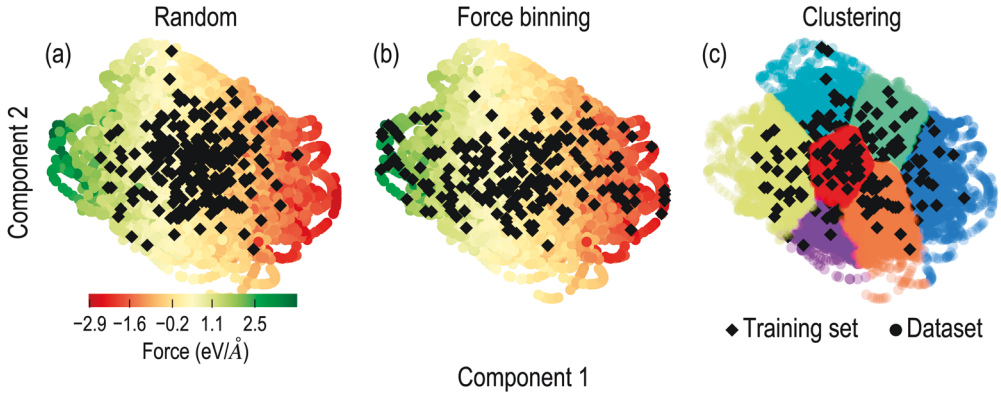

Fig. 3 An illustration of three methods for selecting a training set, including random **a**, force amplitude sampling **b**, and fingerprint space clustering **c**. In this Figure, the reference dataset is projected onto a 2-dimensional manifold spanned by the first two principal components (Component 1 and Component 2), identified by a principal component analysis. Different colors are used in **a**, **b** for representing force magnitude while in **c**, they are used to label different data clusters

Critical assessment of force fields

Fingerprint optimization. We consider five classes of ML models categorized in terms of fingerprint type and sampling method. These include using (I) $\mathbf{V}'_{i,\alpha}$ with random sampling, (II) $\mathbf{V}_{i,\alpha}$ with random sampling, (III) $\mathbf{V}_{i,\alpha}$ with fingerprint space clustering, (IV) $\mathbf{V}_{i,\alpha}$ with force binning sampling, and (V) $V_{i,\alpha}$ with the domain-specific learning approach. In the last class, each force field includes five fingerprint-force mappings established on five data clusters. When a new atomic configuration is encountered, the cluster whose centroid is closest to this point in fingerprint space is identified, and the ML model established for this cluster is used to evaluate the force desired. Since (I) and (II) differ only by the fingerprint choice, we use these two classes to critically compare fingerprints $\mathbf{V}_{i,\alpha}$ and $\mathbf{V}'_{i,\alpha}$, and also to determine the optimal choices of the fingerprint parameters.

$\mathbf{V}_{i,\alpha}$ and $\mathbf{V}'_{i,\alpha}$ were optimized by determining the primary parameters that minimize the error of the ML force field predictions. Although d can be arbitrarily large, we examined $d = 4$, 8, 16, 32, or 48. For $\mathbf{V}_{i,\alpha}$, w was chosen to be 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, or 0.60 Å while a_k were uniformly distributed between a half of the nearest-neighbor distance of the material considered to R_c, chosen to be 8 Å following refs. 5–8. In the case of $\mathbf{V}'_{i,\alpha}$, η_k were uniformly distributed on the log scale between a half of the nearest-neighbor distance to $\eta^{\max}$, ranging up to R_c. A training set of $N_t = 1{,}000$ atomic configurations was prepared using the random sampling method, leaving the complement as the test set. Baseline ML force fields were then created, utilizing $\mathbf{V}_{i,\alpha}$ and $\mathbf{V}'_{i,\alpha}$. The root-mean-square error δ_{RMS} of the force predicted on the test set was used to identify the optimal parameters. Details of this procedure can be found in Fig. S1 and related discussions.

We now summarize our findings for these baseline ML force fields. First, the general trend is that the higher the dimensionality d, the better the fingerprint, i.e., $d = 48$ is optimal for both $\mathbf{V}_{i,\alpha}$ and $\mathbf{V}'_{i,\alpha}$. Second, the optimized $\mathbf{V}_{i,\alpha}$ is superior to the optimized $\mathbf{V}'_{i,\alpha}$ in terms of δ_{RMS}. Finally, for a given d, the optimal w is of the order of $a_{k+1} - a_k$, the spacing between the centers of two adjacent Gaussian functions. Such a value allows for some overlap between these radial basis functions, maximizing the sensitivity of $\mathbf{V}_{i,\alpha}$ with respect to changes of the atomic environment (for more information, see Fig. S2 and related discussions for a principal component analysis of $\mathbf{V}_{i,\alpha}$ and $\mathbf{V}'_{i,\alpha}$).

Figure 4 shows the learning curves, constructed from δ_{RMS}, for the baseline ML force fields (i.e., those of classes (I) and (II), which are based on random data sampling) created for Al, Cu, Ti, W, Si, and C using the optimized $\mathbf{V}_{i,\alpha}$ and $\mathbf{V}'_{i,\alpha}$. For both fingerprints, δ_{RMS} scales inversely with N_t initially[39] but reaches a limit at $N_t \gtrsim 500$. By using $\mathbf{V}_{i,\alpha}$ instead of $\mathbf{V}'_{i,\alpha}$, δ_{RMS} drops from ≃0.16 eV/Å to ≃0.09 eV/Å in case of C. Significant improvement was also obtained for W, Ti, and Cu. For Al and Si, whose δ_{RMS} obtained with $\mathbf{V}'_{i,\alpha}$ is consistent with that reported in refs. 6,9, the reduction of δ_{RMS} is smaller but remains noticeable.

Force field performance

In this section, we critically examine the five classes of AGNI ML force fields considered via their prediction error. Because the forces to be evaluated may span over several orders of magnitude, δ_{RMS} may not be sufficient for describing its performance. We propose a few other error metrics. They are (1) the absolute error δ_{ABS}, (2) the standard deviation of the error distribution (assumed to be normal) δ_{STD} and (3) the average of the top 1% of the force errors δ_{MAX}. While δ_{ABS} should be examined on the whole range of force amplitude (see Fig. 5), δ_{STD} and δ_{MAX} capture the regimes of small and large errors, respectively.

We first focus on the ML force fields created by four recipes that employ different fingerprints and sampling methods, i.e., (I), (II), (III), and (IV), whose δ_{RMS}, δ_{STD}, and δ_{MAX} are summarized in Table 1. The evolution of the error measures from (I) to (II) demonstrates that $\mathbf{V}_{i,\alpha}$ is clearly better than $\mathbf{V}'_{i,\alpha}$, especially for W and C. The errors associated with force binning and clustering methods are comparable, being significantly smaller than those of the force fields created using the random sampling method, especially for Ti. As shown in Fig. 5, the absolute error δ_{ABS} of these ML force fields falls within 0.02 – 0.09 eV/Å in the whole range of force magnitude. This respectable level of error is equivalent or better than those of other contemporary ML force fields, e.g., ≃0.1 eV/Å for bulk Si,[25] ≃0.2 eV/Å for Si_n clusters,[19] and ≳0.04 eV/Å for W.[18]

Next, we discuss the applicability of the AGNI force fields within MD simulations, especially pertaining to our ability to obtain total potential energies by appropriate integration of the atomic forces along a MD trajectory, and the stability of such simulations with respect to total energy conservation. In fact, the force field of class (I) created for Al (employing $\mathbf{V}'_{i,\alpha}$) has been successfully used for a variety of MD simulations.[5,7,8] Herein, we tested the ML force fields created using (IV), combining $\mathbf{V}_{i,\alpha}$ and the force binning sampling method [we note that (III) and (IV) offer essentially similar performance]. For this purpose, we carried our several micro-canonical (NVE) and canonical (NVT) MD simulations, employing the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS).[40] The simulations were performed at 200 and 400 K over 500 ps, using $\Delta t = 0.5$ fs for the time step. Along the MD trajectories, the total potential energy E_{pot} at a time (t) is

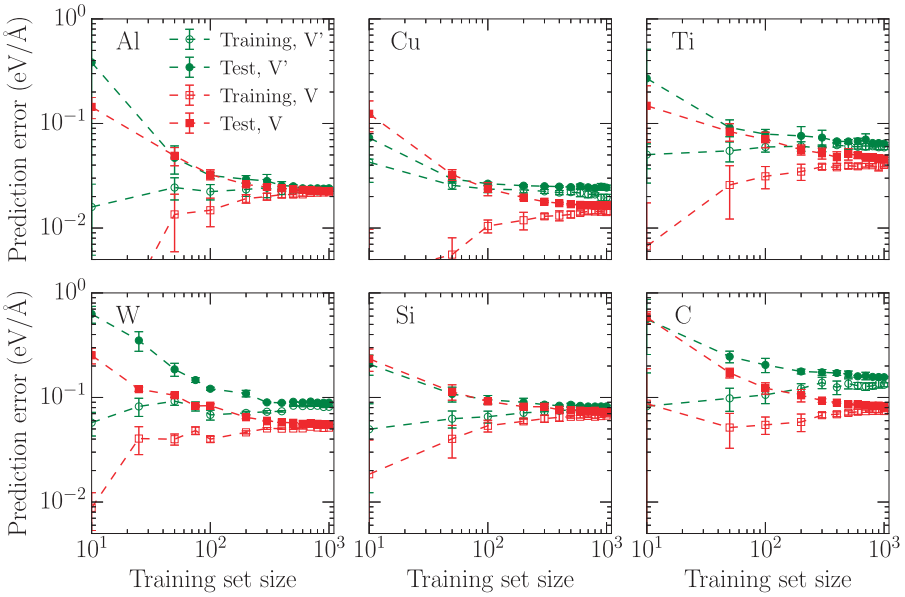

Fig. 4 Learning curves constructed from δ_{RMS}, the root-mean-square error of the AGNI force field created for Al, Cu, Ti, W, Si, and C using the random sampling method and either optimized $\mathbf{V}_{i,a}$ (with $d = 48$ and $w = 0.1$ Å) or optimized $\mathbf{V}'_{i,a}$ (with $d = 48$). Each data point (and its error bar) is obtained from 10 force fields created on 10 independently selected training sets

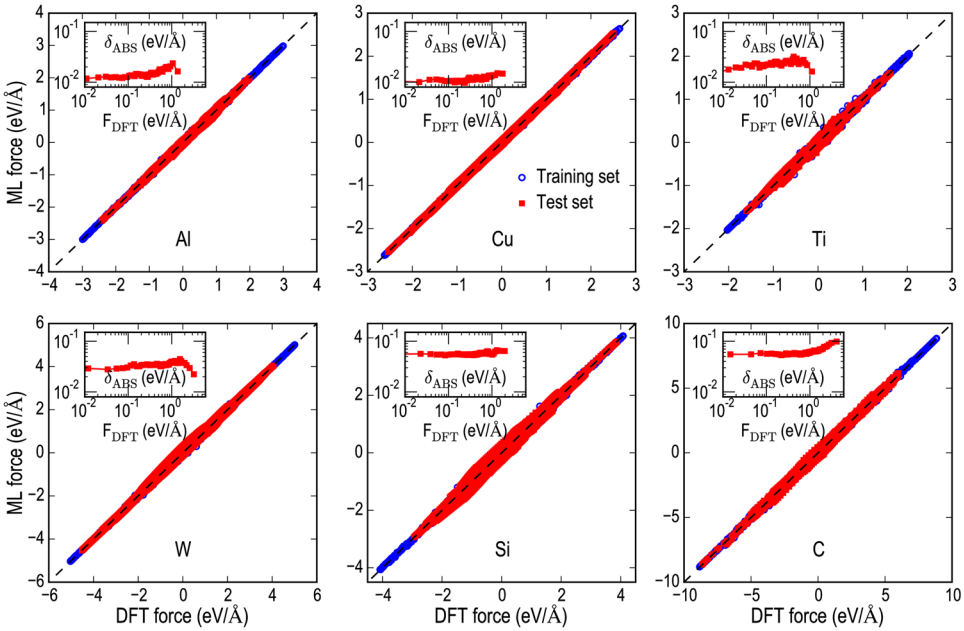

Fig. 5 Atomistic forces evaluated by the domain-specific force fields developed for Al, Cu, Ti, W, Si, and C, shown against the reference DFT forces. The absolute errors of the force predictions are shown in insets as a function of reference force magnitudes. These force fields are created with the training set size $N_t = 1000$

computed from that of the previous step $(t - \Delta t)$ as[7]

$$E_{pot}(t) = E_{pot}(t - \Delta t) - \Delta t \sum_{i,a} F_i^a v_i^a, \quad (4)$$

where the sum runs over the atom index i and the Cartesian index a (with $E_{pot}(t = 0) = 0$). The kinetic energy E_{kin}, the potential energy E_{pot}, and the total energy E_{tot} of the NVE and NVT simulations performed on a 256 atom supercell of Cu are shown in Fig. 6a, b, respectively, while those for other materials are given in the Supporting Information. For the NVE simulations, E_{tot} is clearly conserved, presumably due to the intrinsic error cancellation of the ML force fields used, evidenced by the symmetric distribution of the prediction errors frequently reported.[5,7] The NVT simulations suggest that E_{pot} constructed from Eq. (4) works well with the thermostats implemented in LAMMPS. The agreement between E_{pot} computed using Eq. (4) and those computed using DFT for several snapshots taken at 100, 200, 300, 400, and 500 ps, is shown in Fig. 6c, strongly supporting the accuracy of our force fields (with respect to both atomic force and total potential energy predictions) and their applicability in MD simulations.

Table 1. Force evaluation errors of the force fields developed, obtained for a training set size $N_t = 1000$

FF	Al			Cu		
	δ_{RMS}	δ_{MAX}	δ_{STD}	δ_{RMS}	δ_{MAX}	δ_{STD}
(I)	0.025	0.100	0.025	0.024	0.093	0.024
(II)	0.023	0.096	0.023	0.017	0.071	0.017
(III)	0.023	0.092	0.023	0.017	0.074	0.017
(IV)	0.025	0.097	0.025	0.018	0.076	0.017
(V)	0.021	0.082	0.021	0.016	0.056	0.016
FF	Ti			W		
	δ_{RMS}	δ_{MAX}	δ_{STD}	δ_{RMS}	δ_{MAX}	δ_{STD}
(I)	0.065	0.290	0.065	0.094	0.398	0.094
(II)	0.054	0.306	0.054	0.063	0.268	0.063
(III)	0.045	0.173	0.045	0.065	0.244	0.065
(IV)	0.047	0.162	0.047	0.068	0.253	0.068
(V)	0.035	0.149	0.034	0.049	0.200	0.049
FF	Si			C		
	δ_{RMS}	δ_{MAX}	δ_{STD}	δ_{RMS}	δ_{MAX}	δ_{STD}
(I)	0.081	0.296	0.081	0.161	0.778	0.161
(II)	0.074	0.251	0.074	0.088	0.373	0.088
(III)	0.074	0.260	0.074	0.083	0.322	0.083
(IV)	0.074	0.263	0.074	0.087	0.339	0.087
(V)	0.074	0.253	0.074	0.085	0.326	0.085

For each material, three error measures, i.e., δ_{RMS}, δ_{MAX}, and δ_{STD} are reported in eV/Å for (I), (II), (III), (IV), and (V), five recipes of force field creation (described in the text)

Table 1 and Fig. 5 also show that the domain-specific force fields created by (V) are superior to (I), (II), (III), and (IV) in terms of the considered error measurements, approaching chemical accuracy for all the materials considered. Learning curves for the domain-specific force field recipe (V) (analogous to those shown in Fig. 4 for the (I) and (II) recipes) are provided in the Supporting Information. The applicability of the domain-specific force fields within MD simulations has not been tested extensively at this point. As alluded to earlier, as configurations evolve during the course of an MD simulation, atomic environments may shift from one domain to another, and it is imperative that the force prediction during such shifts do not go through discontinuities. These aspects have to be addressed appropriately before recipe (V) can be reliably used within MD simulations. Nevertheless, as the reference dataset grows in size, and the scaling of training and prediction times become intensive, domain-specific learning approaches may become a necessary alternative in the future, and may be a powerful pathway for the creation of adaptive and high-fidelity force fields.

DISCUSSION

ML has emerged as a powerful approach for developing a new generation of highly accurate force fields.[5–9,12–16,23] Accuracy and transferability are not the only advantages of ML force fields. Their most appealing aspect is that they can be created and improved in a highly automatic manner with minimal human intervention within one unified framework. This contribution discusses one such force field, using which atomic forces can be efficiently and accurately computed given just the configuration of atoms. The total potential energy during the course of an MD simulation may then be computed through appropriate integration of the atomic forces along the trajectory. The true versatility of the present single unified strategy is rigorously demonstrated in this work for a variety of elemental solids, displaying a diversity of chemical bonding types, namely Al, Cu, W, Ti, Si, and C. For all of them, the ML force fields reach chemical accuracy while being several orders of magnitude faster than corresponding quantum mechanics-based computations. The accuracy and applicability of the ML force fields can be systematically improved without adversely affecting their efficiency. This maybe accomplished by (1) preparing the reference data at a higher level of theory for computing the force more accurately and/or (2) including new training datasets selected from new domains that have not been considered yet. We hope that after being fully developed, such ML force fields may provide a pathway for high-fidelity MD simulations to enter the regimes of time scales ($\gtrsim$ nanoseconds) and/or length scales ($\gtrsim$10 nanometers) presently difficult to reach using purely quantum mechanical simulations.

METHODS

Reference data preparation with DFT

Our reference data was prepared by first-principles computations performed using the DFT,[41,42] a plane-wave basis set, and the projector augmented wave (PAW) method[43] as implemented in *the Vienna Ab initio simulation package* [44–48] The Perdew, Burke, and Ernzerhof functional[49] was used for the exchange-correlation energies.

The elemental (bulk) materials considered are face-centered cubic (fcc) aluminum Al, fcc copper Cu, hexagonal close packed Ti, body-centered cubic, tungsten (W), diamond silicon (Si), and diamond carbon (C). For each of them, we constructed a suitable supercell of n_{at}^{sc} atoms, and then performed DFT-based MD simulations at two temperatures, $T = 300$ K and $T = 600$ K (for the purpose of data accumulation, the temperature is a parameter for spanning diverse environments). Atomic configurations and forces were then extracted from the MD trajectories obtained. We note that to create a reliable force field for a specific application, e.g., for systems with defects, vacancies, and grain boundaries, the reference data should be more comprehensive, including the atomic environments expected to be encountered during the MD simulations.[7,8]

The convergence of force calculations by DFT is generally slower than that of energy calculations. To ensure the convergence of our data, the kinetic energy cutoff E_c was taken to be 130% of the default value while the projection operators (involved in the calculations of the non-local part of the PAW pseudopotentials) were evaluated in the reciprocal space. The convergence of the forces with respect to the *k*-point mesh, generated without considering possible symmetries of the crystals, was carefully tested. We note that the accuracy level adopted for the reference data generation is significantly higher than used for typical DFT-based MD simulations (see, for example, refs. 50,51). All the parameters used for the data generation step are given in Table 2.

Details of sampling and clustering

The key idea underlying the force binning method is that the training set should reflect the force-amplitude profile of the reference dataset while ensuring sparsely populated domains have their own representatives. Therefore, this procedure involves splitting the reference dataset into a given number of equal-size bins, covering the entire range of force amplitude. Then, 30% of the desired training set was equally selected from the prepared bins. The contribution of each bin to the remaining 70% of the training set was proportional to the population of this bin. Similarly, the clustering method makes sure that any domain in fingerprint space that is occupied by the reference dataset has their own representatives in the training set. In this work, we used the k-means method to cluster the reference dataset into five clusters in fingerprint space based on the distance used in KRR. Then, 20% of the desired training set was selected from each data cluster. The number of clusters was specified beforehand.

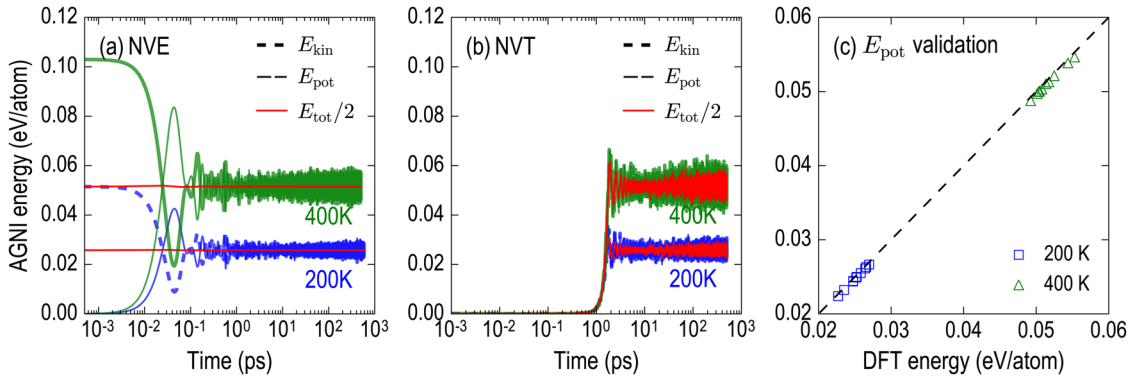

Fig. 6 Kinetic energy E_{kin}, potential energy E_{pot}, and total energy E_{tot} obtained for **a** two NVE and **b** two NVT MD simulations of bulk Cu, performed at 200 and 400 K. Our simulations employ the force field created using $\mathbf{V}_{i,q}$ and force binning sampling method. In **c**, the potential energies computed during the MD simulations using Eq. (4) and DFT for the snapshots taken at 100, 200, 300, 400, and 500 ps from these simulations are shown

Table 2. Parameters used to generate the dataset for force field generation

Materials	n_{at}^{sc}	E_c (eV)	$k-$mesh	PAW set	T (K)
Al	32	315	$7\times7\times7$	Al	300, 600
Cu	32	480	$5\times5\times5$	Cu	300, 600
Ti	16	360	$9\times9\times9$	Ti_sv	300, 600
W	16	290	$9\times9\times9$	W	300, 600
Si	64	320	$5\times5\times5$	Si	300, 600
C	64	520	$7\times7\times7$	C	300, 600

Data availability

All the data used for this work, i.e., the MD trajectories of the six materials, is freely available in http://khazana.uconn.edu with record identification number ranging from 3,278 to 3,283.

ACKNOWLEDGEMENTS

This work was supported financially by the Office of Naval Research (Grant No. N00014-14-1-0098) and by the National Science Foundation (Grant No. 1600218). The authors thank Chiho Kim, Venkatesh Botu, and Erik Nykwest for discussions and technical helps.

AUTHOR CONTRIBUTIONS

R.R. designed and supervised the research. T.D.H. and R.B. implemented the method. T.D.H., R.B., J.C., S.K., and L.C. prepared data. T.D.H. and R.R. wrote the manuscript with inputs from all the authors.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Computational Materials* website (https://doi.org/10.1038/s41524-017-0042-y).

Competing interests: The authors declare that thay have no competing financial interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Jain, A., Shin, Y. & Persson, K. A. Computational predictions of energy materials using density functional theory. *Nat. Rev. Mater* **1**, 15004 (2016).
2. Mannodi-Kanakkithodi, A. et al. Rational co-design of polymer dielectrics for energy storage. *Adv. Mater.* **28**, 6277–6291 (2016).
3. Sharma, V. et al. Rational design of all organic polymer dielectrics. *Nat. Commun.* **5**, 4845 (2014).
4. Mueller, T., Kusne, A. G. & Ramprasad, R. in *Reviews in Computational Chemistry*, (eds Parrill, A. L. & Lipkowitz, K. B.) Ch. 4 (Wiley, 2016).
5. Botu, V. & Ramprasad, R. Learning scheme to predict atomic forces and accelerate materials simulations. *Phys. Rev. B* **92**, 094306 (2015).
6. Botu, V. & Ramprasad, R. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *Int. J. Quant. Chem.* **115**, 1074–1083 (2015).
7. Botu, V., Batra, R., Chapman, J. & Ramprasad, R. Machine learning force fields: Construction, validation, and outlook. *J. Phys. Chem. C* **121**, 511–522 (2017).
8. Botu, V., Chapman, J. & Ramprasad, R. A study of adatom ripening on an al (111) surface with machine learning force fields. *Comput. Mater. Sci.* **129**, 332–335 (2016).
9. Suzuki, T., Tamura, R. & Miyazaki, T. Machine learning for atomic forces in a crystalline solid: Transferability to various temperatures. *Int. J. Quant. Chem.* **117**, 33–39 (2017).
10. Hellmann, H. in *Einführung in die Quantenchemie* (eds Andrae, D.) Ch. 1 (Franz Deuticke, 1937).
11. Feynman, R. P. Forces in molecules. *Phys. Rev.* **56**, 340–343 (1939).
12. Blank, T. B., Brown, S. D., Calhoun, A. W. & Doren, D. J. Neural network models of potential energy surfaces. *J. Chem. Phys.* **103**, 4129–4137 (1995).
13. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
14. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
15. Behler, J. Perspective: machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, 170901 (2016).
16. Deringer, V. L. & Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B* **95**, 094203 (2017).
17. Bartók, A. P. & Csányi, G. Gaussian approximation potentials: a brief tutorial introduction. *Int. J. Quant. Chem.* **115**, 1051–1057 (2015).
18. Shapeev, A. Moment tensor potentials: a class of systematically improvable interatomic potentials. *Multiscale Model. Simul.* **14**, 1153–1173 (2016).
19. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
20. Szlachta, W. J., Bartók, A. P. & Csányi, G. Accuracy and transferability of Gaussian approximation potential models for tungsten. *Phys. Rev. B* **90**, 104108 (2014).
21. Thompson, A. P., Swiler, L. P., Trott, C. R., Foiles, S. M. & Tucker, G. K. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **285**, 316–330 (2015).
22. Glielmo, A., Sollich, P. & De Vita, A. Accurate interatomic force fields via machine learning with covariant kernels. *Phys. Rev. B* **95**, 214302 (2017).
23. Faraji, S. et al. High accuracy and transferability of a neural network potential through charge equilibration for calcium fluoride. *Phys. Rev. B* **95**, 104105 (2017).
24. Chmiela, S. et al. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
25. Li, Z., Kermode, J. R. & De Vita, A. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.* **114**, 096405 (2015).
26. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Dat* **1**, 140022 (2014).
27. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. & Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **3**, 2810 (2013).

28. Huan, T. D., Mannodi-Kanakkithodi, A. & Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B* **92**, 014106 (2015).
29. Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. Machine learning strategy for the accelerated design of polymer dielectrics. *Sci. Rep.* **6**, 20952 (2016).
30. Kim, C., Pilania, G. & Ramprasad, R. From organized high-throughput data to phenomenological theory using machine learning: the example of dielectric breakdown. *Chem. Mater.* **28**, 1304–1311 (2016).
31. Kim, C., Pilania, G. & Ramprasad, R. Machine learning assisted predictions of intrinsic dielectric breakdown strength of abx3 perovskites. *J. Phys. Chem. C* **120**, 14575–14580 (2016).
32. Huan, T. D. et al. A polymer dataset for accelerated property prediction and design. *Sci. Data* **3**, 160012 (2016).
33. Kim, C., Huan, T. D., Krishnan, S. & Ramprasad, R. A hybrid organic-inorganic perovskite dataset. *Sci. Data* **4**, 170057 (2017).
34. Wang, J., Neskovic, P. & Cooper, L. N. in Advances in Natural Computation: First International Conference, ICNC 2005, Changsha, China, August 27–29, 2005, Proceedings, Part I, (eds Wang, L., Chen, K. & Ong, Y. S.) 554–564 (Springer, 2005).
35. Lessmann, S., Stahlbock, R. & Crone, S. F. Genetic algorithms for support vector machine model selection. The 2006 IEEE International Joint Conference on Neural Network Proceedings, 3063–3069 (2006).
36. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd edn (Springer, 2009).
37. Hofmann, T., Schölkopf, B. & Smola, A. J. Kernel methods in machine learning. *Ann. Stat.* **36**, 1171–1220 (2008).
38. Rupp, M., Ramakrishnan, R. & von Lilienfeld, O. A. Machine learning for quantum mechanical properties of atoms in molecules. *J. Phys. Chem. Lett.* **6**, 3309–3313 (2015).
39. Huang, B. & von Lilienfeld, O. A. Communication: understanding molecular representations in machine learning: the role of uniqueness and target similarity. *J. Chem. Phys.* **145**, 161102 (2016).
40. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).
41. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).
42. Kohn, W. & Sham, L. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
43. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
44. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993).
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.* **6**, 15–50 (1996).
47. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
48. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
49. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
50. Sjostrom, T., Crockett, S. & Rudin, S. Multiphase aluminum equations of state via density functional theory. *Phys. Rev. B* **94**, 144101 (2016).
51. Karin, T., Dunham, S. & Fu, K.-M. Alignment of the diamond nitrogen vacancy center by strain engineering. *Appl. Phys. Lett.* **105**, 053106 (2014).