Adaptive Machine Learning Framework to Accelerate *Ab Initio* Molecular Dynamics

Venkatesh Botu[a] and Rampi Ramprasad*[b]

Quantum mechanics-based *ab initio* molecular dynamics (MD) simulation schemes offer an accurate and direct means to monitor the time evolution of materials. Nevertheless, the expensive and repetitive energy and force computations required in such simulations lead to significant bottlenecks. Here, we lay the foundations for an accelerated *ab initio* MD approach integrated with a machine learning framework. The proposed algorithm learns from previously visited configurations in a continuous and adaptive manner on-the-fly, and predicts (with chemical accuracy) the energies and atomic forces of a new configuration at a minuscule fraction of the time taken by conventional *ab initio* methods. Key elements of this new accelerated *ab initio* MD paradigm include representations of atomic configurations by numerical fingerprints, a learning algorithm to map the fingerprints to the properties, a decision engine that guides the choice of the prediction scheme, and requisite amount of *ab initio* data. The performance of each aspect of the proposed scheme is critically evaluated for Al in several different chemical environments. This work has enormous implications beyond *ab initio* MD acceleration. It can also lead to accelerated structure and property prediction schemes, and accurate force fields. © 2014 Wiley Periodicals, Inc.

DOI: 10.1002/qua.24836

Introduction

Computation-driven rational materials design efforts are rising in popularity and importance.[1,2] This trend is being fueled by systematic improvements in capabilities to compute materials properties accurately and practically. Parameter-free (or *ab initio*) quantum mechanics (QM)-based schemes such as density functional theory (DFT) are central to this unfolding development.[3–5] Although powerful, versatile, and efficient, *ab initio* methods are still too time intensive to adequately handle several important classes of problems. For instance, the explicit dynamical evolution of materials and processes with timescales larger than a nanosecond are still beyond the reaches of DFT computations.

The most direct way to handle and monitor the time evolution of matter is by the molecular dynamics (MD) method.[6] In *ab initio* MD, the ingredients necessary to perform MD, namely, the total potential energies and atomic forces are obtained using QM, but the evolution of the atoms (i.e., determination of the next new configuration, based on the current configuration, velocities, and forces) is performed classically. The repetitive and expensive QM energy and force computations, and the necessity for small time steps (of the order of femtoseconds), lead to the primary bottlenecks of *ab initio* MD. Creative schemes to accelerate MD simulations so that longer timescales can be accessed have indeed been developed in the past.[7–18] These include the use of parameterized force-fields (rather than QM) to evaluate the energies and forces rapidly,[7] and/or speeding the clock using Monte Carlo methods,[8,9] metadynamics,[10,11] temperature accelerated dynamics,[12–15] and hyperdynamics.[13–16] These attempts although are not entirely satisfactory. Force-fields are not transferrable to situations that were not originally used in the parameterization, and altering the clock requires some prior knowledge of the critical features encountered during the evolution process (and involve artificial constraints and some loss of vital dynamical information).

The present contribution provides a pathway for a new solution to the *ab initio* MD acceleration problem that preserves the fidelity of both QM and the clock. First, we make three observations.

1. During a typical MD trajectory, a system is largely exploring similar configurations, and new features or events are encountered rarely, as schematically portrayed in Figure 1a. This observation is quite universal, and applies to many important processes such as defect diffusion in solids or surface chemical reactions. Taking point defect diffusion as an example, the actual site-to-site hopping of the defect is a rare event, while the vibrational motion of the defect (and its surroundings) in its local minimum occupies most of the time and leads to a plethora of similar configurations.
2. It is fair to assume that similar configurations will have similar properties (such as energies, atomic forces, etc.). If a robust numerical representation of the configurations

[a] *V. Botu*
Chemical and Biomolecular Engineering, University of Connecticut, Storrs, Connecticut, 06269
[b] *R. Ramprasad*
Materials Science and Engineering, University of Connecticut, Storrs, Connecticut, 06269
E-mail: rampi@uconn.edu

Contract grant sponsor: Office of Naval Research; contract grant number: N00014-14-1–0098.

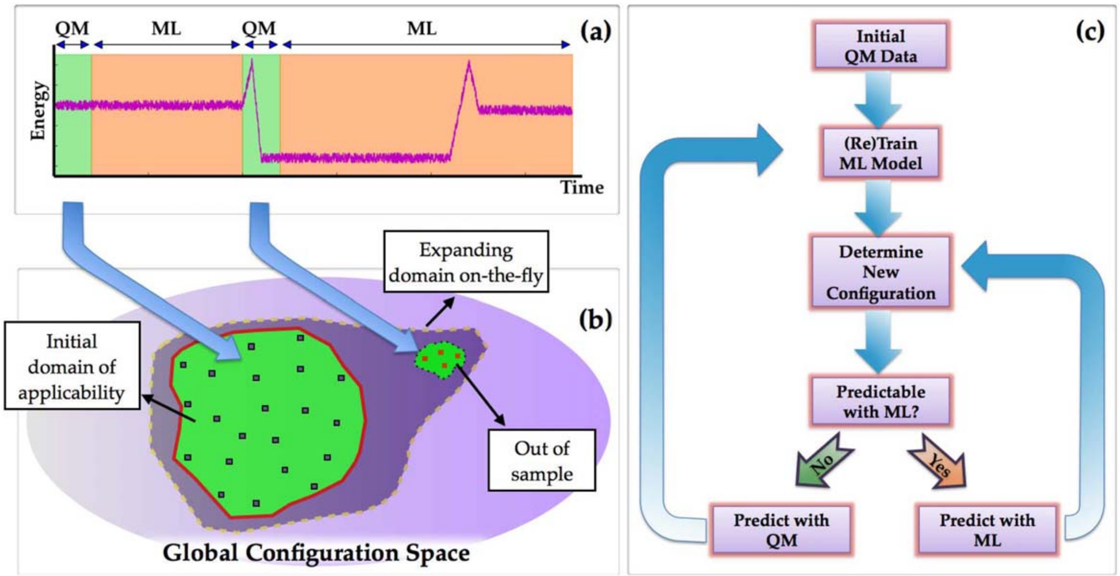

Figure 1. a) A typical MD energy trajectory, with the green and orange regions identifying the quantum mechanical (QM) and machine learning (ML) phases, respectively, of the adaptive learning framework. b) Expansion of the domain of applicability on-the-fly, if and when new configurations are visited. c) A flowchart of the adaptive learning framework. The green and orange arrows indicate the use of QM or ML models.

can be developed, a quantitative measure of (dis)similarity of configurations can be defined, which can then be mapped to (dis)similarities between properties via a learning algorithm. Within the context of accelerated MD, a machine learning (ML) procedure can be used to predict the energies and forces of similar configurations along the MD trajectory rapidly, provided QM training data pertaining to the initial part of the trajectory is available. This is also shown in Figure 1a.

3. When a completely new configuration or event is encountered, a decision has to be made to switch back from ML to QM. Most importantly, the new configurations and properties should be included in the learning framework on-the-fly as illustrated in Figure 1b, making the learning process adaptive, and continuously evolving with progressive improvement in predictive quality. If this can be accomplished, then, the next time a similar rare event is encountered, QM is unnecessary. This aspect is also captured in Figure 1a.

Thus, the basic premise of the proposed strategy is that the significant redundancies implicit in conventional *ab initio* MD schemes can be systematically eliminated. The flowchart shown in Figure 1c summarizes the proposed on-the-fly adaptive ML strategy to accelerate *ab initio* MD. This concept is reminiscent of the adaptive force-field scheme utilized earlier for simulations involving Si,[19] although the present strategy is more general, flexible, and universal in terms of its applicability, learning capability, and representation of atomic interactions.

It is worth noting that ML strategies are making significant inroads into various aspects of materials science,[20] including accelerated and accurate predictions (using past historical data) of phase diagrams,[21,22] crystal structures,[23–26] and material properties,[27–29] mapping complex materials behavior to a set of process variables,[30–32] data analysis of high-throughput experiments,[22,33,34] so forth. Of particular relevance to the present contribution are recent successful efforts that exploit ML methods (neural networks[35] and Gaussian approximation potentials[36]) to develop accurate force-fields (or interatomic potentials) that can allow for significant extension of the time- and length-scales of MD simulations. Nevertheless, the present contribution is one of the first attempts in which the implementation of an adaptive on-the-fly learning scheme to accelerate *ab initio* MD is discussed.

The proposed strategy, as captured in Figure 1c, involves a number of vital ingredients. These include: (1) a rigorous and generalizable scheme to represent atomic configurations by continuous numerical fingerprints that are invariant to translations, rotations, and permutations of like atom types (as such transformations lead to equivalent configurations); (2) a robust learning algorithm that can map the fingerprints to properties; (3) a decision engine that queries whether the properties of a new configuration are predictable using the current learning model; and (4) needless to say, *ab initio* (re)training data from the initial part of the MD trajectory and at points when the decision engine makes *ab initio* calculations mandatory.

A firm understanding of the requirements and the limits of the four ingredients listed above is necessary for the practical realization of a high-fidelity, accelerated MD simulation scheme. To directly address this, in this manuscript, we consider fcc Al, a model elemental metallic system in several chemically distinct environments, including (i) defect-free bulk Al, (ii) bulk Al containing a vacancy, (iii) clean (111) Al surface, and (iv) the (111) surface with an Al adatom. For each of the four cases above, robust numerical configurational fingerprints are created that allow for high-fidelity predictions of energies and forces at chemical accuracy via a similarity-based learning

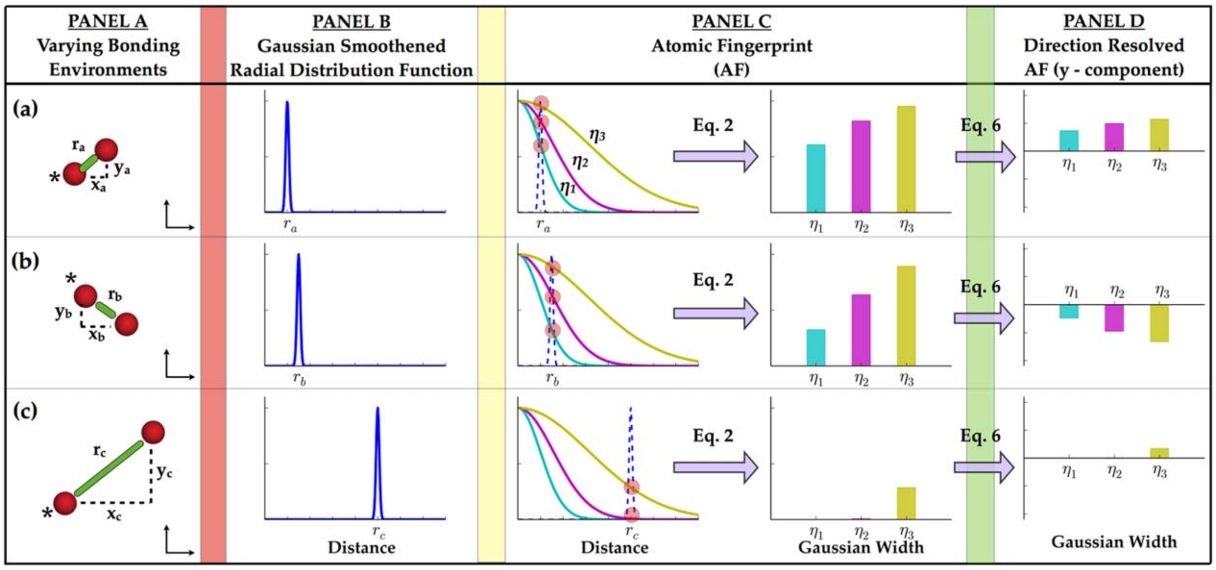

Figure 2. Panel A: A homonuclear diatomic molecule displaying three different bond lengths. Panel B: The corresponding Gaussian smoothened radial distribution function (RDF) for each of the bonding environments. Panel C: Transformation of the RDF using Gaussian functions on an eta-grid as indicated by the colored lines, into an atomic fingerprint. Panel D: The *y*-component of the direction resolved atomic fingerprint of an atom in the three bonding environments. The fingerprints generated are for the atom indicated by * in Panel A.

algorithm. Also, a simple decision engine is presented that detects the occurrence of a new configuration not already in the initial training dataset, thus signaling when a fresh QM calculation is required. The combination of the individual working entities should lead us to the ultimate goal of an adaptive learning framework to significantly accelerate *ab initio* MD simulations on-the-fly.

Methods and Models

Fingerprints: Numerical representations of atomic, molecular, and crystal environments

The first critical step in the proposed learning approach is to represent the chemistry and geometry of our system numerically (hopefully, uniquely), such that a mapping can be established between this numerical representation and the property of interest (namely, the energy or forces). Such a representation is referred to here as a fingerprint (also commonly referred to as the feature vector by the ML community). In what follows, we distinguish between atomic fingerprints and crystal (or molecular) fingerprints. The former captures the coordination environment of a particular atom, while the latter describes the entire ensemble of atoms that are contained within a repeating unit cell (or a molecule). The atomic fingerprint is necessary to predict atomic properties (e.g., forces), while the crystal fingerprint is appropriate to capture global properties (e.g., energy within quantum mechanical schemes, band gap, etc.).

The atomic or crystal fingerprint is required to satisfy certain requirements.[37,38] To adequately capture variations in energy and forces with geometry differences, the fingerprint has to be continuous with respect to slight changes in configuration. Moreover, transformations such as translations, rotations, and permutations of atoms of the same type that lead to equivalent systems should not alter the fingerprint.

We first consider atomic fingerprints with the expectation that crystal fingerprints can be built from the constituent atomic fingerprints. A natural first choice for the atomic fingerprint of an elemental system could be the radial distribution function (RDF) defined as follows for a particular atom *i*

$$R_i(r)=\sum_{j\neq i}\delta(r-r_{ij}) \tag{1}$$

where $\delta(r)$ is the Dirac delta function and $r_{ij}=|\boldsymbol{r}_i-\boldsymbol{r}_j|$, with $\boldsymbol{r}_i$ being the vectorial position of atom *i*. The sum runs over all the neighboring atoms within an arbitrarily large cutoff distance from atom *i*. Clearly, the RDF, $R_i(r)$, satisfies both the fingerprint requirements mentioned above, and has recently been used to establish structure-property mappings in materials.[39] The values of R_i in a radial grid can thus be viewed as a numerical fingerprint (or feature vector) describing the coordination environment. Moreover, $R_i(r)$ also captures the geometry in a visually appealing manner. This is demonstrated in Figure 2. Panel A contains three homonuclear diatomic molecules (labeled a, b, and c) used here to illustrate our fingerprint choices, and Panel B shows the corresponding Gaussian smoothened RDFs. Clearly, the similarity between the bond distances of molecules a and b, and their dissimilarity with that of molecule c is reflected by the corresponding RDFs. Nevertheless, while these (dis)similarities are apparent to a human, it may not be so for a machine. Typical measures of (dis)similarity utilize the Euclidean norm of the difference between the fingerprint vectors or the dot product between the fingerprint vectors. Clearly, such measures will fail to capture the similarity between molecules a and b, and their dissimilarity with respect to molecule c (as the Euclidean norms of the difference between any pair of the three fingerprint vectors is the same constant value, and the dot products between any pair is zero).

Extending the RDF in a particular way can circumvent the above problem. Rather than using the RDF itself, a transformed quantity defined as the integral of the product of $R_i(r)$ and a Gaussian window function

$$G_i(\eta)=\int R_i(r)\mathrm{e}^{-\left(\frac{r}{\eta}\right)^2}dr=\sum_{j\neq i}\mathrm{e}^{-\left(\frac{r_{ij}}{\eta}\right)^2} \tag{2}$$

can be used, where η is a parameter that describes the extent of the window function. $G_i(\eta)$ is essentially a "cumulative" version of $R_i(r)$. This is visually demonstrated in Panel C of Figure 2, for three η values. Although $R_i(r)$ is defined in a radial grid, $G_i(\eta)$ is defined in a η-grid. To account for the diminishing importance of atoms far away from the reference atom i, we multiply the summand of $G_i(\eta)$ by a cutoff function $f(r_{ij})$ that smoothly vanishes for large r_{ij} values, resulting in our choice of the atomic fingerprint (AF) function, $A_i(\eta)$, given by

$$A_i(\eta)=\sum_{j\neq i}\mathrm{e}^{-\left(\frac{r_{ij}}{\eta}\right)^2}f(r_{ij}). \tag{3}$$

We note that $A_i(\eta)$ is essentially the radial symmetry function proposed earlier by Behler et al.[40] Following that previous work we define $f(r_{ij})$ as

$$f(r_{ij})=\begin{cases}0.5\left[\cos\left(\frac{\pi r_{ij}}{R_c}\right)+1\right] & \text{if } r_{ij}\leq R_c\\ 0 & \text{if } r_{ij}>R_c\end{cases} \tag{4}$$

where R_c is the cutoff radius, chosen here to be 8 Å. Interestingly, the η-grid does not have to be as fine as the radial grid. More importantly, $A_i(\eta)$ does not have the issues that $R_i(r)$ has, with respect to capturing the (dis)similarity between actual physical situations as defined by Euclidean norms. This can be ascertained by inspecting Panel C of Figure 2.

For the molecular or crystal fingerprint (i.e., the fingerprint of the entire molecule or unit cell, $C(\eta)$, also defined on a η-grid) to be used for mapping the total potential energy of a configuration, we use the average of the atomic fingerprint $A_i(\eta)$ over the constituent atoms, as given by

$$C(\eta)=\frac{1}{N}\sum_{i}^{N}A_i(\eta) \tag{5}$$

where N is the total number of atoms in the molecule or unit cell.

Finally, we consider the extension of the $A_i(\eta)$ definition so that it becomes applicable to represent vectorial atomic quantities such as forces. This can be simply done by resolving each term in the summation of $A_i(\eta)$ into its Cartesian components, leading to the direction-resolved atomic fingerprints, $\boldsymbol{V}_i(\eta)=\{V_i^x(\eta),V_i^y(\eta),V_i^z(\eta)\}$ as follows

$$V_i^k(\eta)=\sum_{j\neq i}\frac{r_{ij}^k}{r_{ij}}\mathrm{e}^{-\left(\frac{r_{ij}}{\eta}\right)^2}f(r_{ij}),\quad k\in\{x,y,z\} \tag{6}$$

where r_{ij}^k is the k-th component of $(\boldsymbol{r}_i-\boldsymbol{r}_j)$. Panel D of Figure 2 visually demonstrates the $V_i^y(\eta)$ function for the homonuclear diatomic molecular systems of Panel A.

To extend the atomic fingerprint (be it $A_i(\eta)$ or $\boldsymbol{V}_i(\eta)$), to nonelemental systems, one could follow a similar approach as above, whereby the atomic fingerprint contains components, one for each atom type. For example, given a binary system with elements m and n, the possible atomic neighbor pair distribution types are: mm, mn, nm, and nn. Thus, by considering each interaction separately, we propose a new multielement atomic fingerprint generated by concatenating the independent atomic pair fingerprints, that is, $A_i(\eta)=\left[A_i^{mm},A_i^{mn},A_i^{nm},A_i^{nn}\right]$. The crystal fingerprint may be generated by averaging over the individual $A_i(\eta)$ within a given supercell. Similarly, the direction resolved atomic fingerprint could be generated by concatenation of the individual pair components.

Learning method: Kernel ridge regression

The second critical step is the choice of the learning method. In this work, we have chosen the kernel ridge regression (KRR) technique, which has been used successfully in the recent past within the materials and chemical sciences.[27,28,34] KRR transforms the input fingerprint into a higher dimensional space whereby a linear relation between the transformed fingerprint and the property of interest can be established.[41–43] To be precise, the mapping process between the fingerprint and property involves the "distances" between fingerprints rather than the fingerprints themselves. KRR may thus be viewed as a similarity-based learning method, that is, similar fingerprints will lead to similar properties.

Within KRR, the property of a system u is given by a sum of weighted Gaussians,

$$P_u=\sum_{v}\alpha_v\mathrm{e}^{-\frac{1}{2}\left(\frac{|d_{uv}|}{\sigma}\right)^2} \tag{7}$$

where v runs over all the cases in the training dataset. d_{uv} is the Euclidean distance between the fingerprint vectors of systems u and v. The coefficients α_vs and the parameter σ are determined during the training phase, whence an objective function that includes a regularization term is minimized.[27,28] The σ and regularization parameters are determined by k-fold cross-validation (in this work $k=5$) on the training dataset. In this method, the training dataset is split into k bins. Each bin acts as a new test dataset, whilst the remaining $k-1$ bins are combined into a new training dataset. The process is repeated for every bin in the k bins, and for every σ and regularization parameters on a preselected logarithmically scaled fine grid. The optimal parameters (i.e., ones that lead to the lowest k-fold cross validation error) are then used in the final model development stage to determine the α_v values for the entire training dataset. In this work, model error is measured by the mean absolute error (MAE). At this point, the machinery is in place to predict the property value using Eq. (7).

Decision engine: Fingerprint range

The third critical step is the decision engine that guides prediction machinery choice (either QM or ML) for energy and force evaluations. If a simulation spends a majority of the time using the *ab initio* engine, it nullifies any speedup. This raises an important question, namely, how do we judge whether the property of a new configuration can be predicted with the ML approach? One way to classify a new structure is to compare its fingerprint with those in the training dataset (once sufficient initial training data has been accumulated). If every component of a new fingerprint lies within the range of components of fingerprints already in the training dataset, then we decide that we are in the predictable domain. If not, a fresh QM calculation is mandatory. The new results should then be included in the training set and retraining must be performed to improve the predictive capability. Certainly, a more complex decision engine can be developed by taking inspiration from the field of domain applicability as used within drug prediction,[44–46] but this is not attempted here.

Data generation: Quantum Mechanics

Data for the four cases (i) defect-free bulk Al, (ii) bulk Al containing a vacancy, (iii) a clean (111) Al surface, and (iv) the (111) surface with an Al adatom was generated from *ab initio* (DFT) MD runs in a microcanonical ensemble using a timestep of 0.5 fs, with the Vienna *ab initio* Simulation Package.[47,48] The bulk cases (i and ii) consisted of a 32 (or 31 with the vacancy) atom model. The surface cases (iii and iv) consisted of a 16 (or 17 with adatom) atom surface model. The generalized gradient approximation functional parameterized by Perdew, Burke, and Ernzerhof to treat the electronic exchange-correlation interaction, the projector augmented wave potentials, and plane-wave basis functions up to a kinetic energy cutoff of 520 eV were used.[47–50] A Γ-centered *k*-point mesh of $7 \times 7 \times 7$ and $7 \times 7 \times 1$ were used for the bulk and surface calculations, respectively.

Training and test datasets

As a reminder, we note that two types of fingerprints are used in this work: $C(\eta)$ to map to total potential energies and $\boldsymbol{V}_i(\eta)$ to map to atomic forces. Using *ab initio* MD, a total of 2000 configurations was generated for each of the four material systems considered. For the energy prediction assessment, various amounts of training data was randomly selected from the above sampled configurations, while the remaining was considered as the test dataset, used to gauge model performance. Similarly for the force prediction assessment, various amounts of training data was randomly selected from all the atomic environments sampled (i.e., 64,000 for case i and ii, and 32,000 for case iii and iv), with the remaining considered as the test dataset.

Desired Parameter Choices

The specific choice of model parameters is critical to performance.[51] To establish fidelity in predictions, we extensively tested two key quantities: (1) the length of the fingerprint vector (i.e., number of points in the η grid), and (2) the training dataset size.

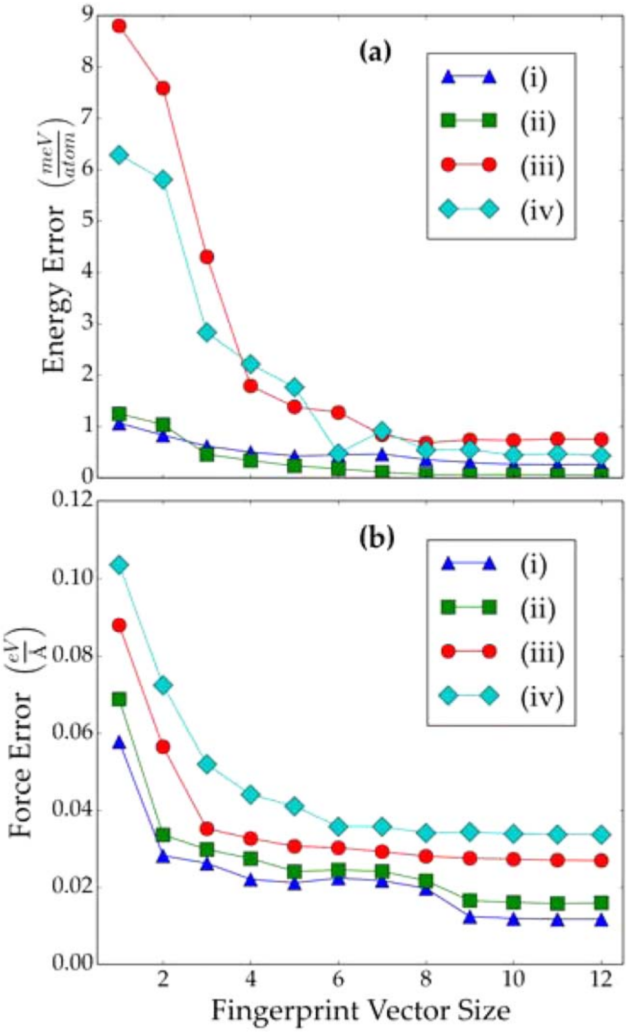

Figure 3. Energy a) and force b) error versus length of fingerprint size for (i) defect-free bulk Al, (ii) bulk Al containing a vacancy, (iii) a clean (111) Al surface, and (iv) the (111) surface with an Al adatom.

Fingerprint vector size

A natural question that arises is, how dense should the η-grid defined in Eqs. (3) and (6) be, to adequately describe the varying atomic and crystal environments encountered. To critically address this question, we systematically increased the number of η values from 1 to 12 (thus increasing fingerprint complexity). Starting with small η values, as this captures the dominant nearest neighbor shell contributions, we added more components to the fingerprint based on a logarithmic η grid between 10^{-1} and 10^{2} Å. For each case, we used a training dataset size of 100 and 500 for the energy and force models, respectively (these sizes are shown in the next subsection to be sufficient to ensure convergence of the predictions). The model error as shown in Figure 3, decreases with increasing fingerprint complexity for all four cases, suggesting that convergence has been achieved.

Interestingly, with the energy model the fingerprint complexity is also dependent on the type of structure being studied. As seen in Figure 3a, to achieve chemical accuracy in energy (MAE < 1 $\frac{\text{meV}}{\text{atom}}$), the bulk cases (i and ii) required a 3-component fingerprint, whereas the surface cases (iii and iv) required an 8-component fingerprint. The above observation is not entirely surprising. A surface model, unlike the bulk, is nonperiodic along the surface normal whereby atoms of varying coordination exist, depending on the atom position (surface or below). The learning algorithm maps the energy to a crystal fingerprint (which is averaged across all atoms), and hence the resolution of each individual atom is smeared out. Only on increasing fingerprint complexity can we achieve an accurate model. Such a concern does not exist for the force model, since a one-to-one mapping between the atomic environment and the force is undertaken. It is for this reason that the force error, as seen in Figure 3b, for all four cases starts high (MAE > 0.05 $\frac{\text{eV}}{\text{Å}}$) and decreases systematically, with error levels converging well below numerical DFT noise.

Training dataset size

Another factor affecting the performance of the learning algorithm is the size and choice of the training data used. With KRR, increasing the training dataset size generally improves model performance, so long as accurate data and cross-validation methods are used during model training. The size of the dataset, n, has to be carefully chosen as computational overhead scales as $\mathcal{O}(n^3)$.[52] To determine the optimal training size that balances computational expense with accuracy, model error versus training dataset size was studied as shown in Figure 4, using an 8-component crystal and direction-resolved atomic fingerprint. Clearly, a systematic decrease in error with increased training once again signifies convergence.

Models with small training dataset sizes (<25 for energy and <50 for force) leads to poor learning, resulting in high errors. For the energy model, bulk cases (i and ii) require 25 configurations or more, while the surface cases (iii and iv) require 50 configurations or more to achieve error convergence. Conversely for the force model, the bulk cases converge to the desired accuracy with <50 training configurations, while the surface cases require >200 configurations. Similar to the observations with fingerprint complexity, as the configurational expanse increases from the bulk to surface owing to the nonperiodicity, the training size required increases accordingly.

Prediction of Energy and Forces

Based on the convergence studies of model parameters in Desired Parameter Choices section, we chose eight components for both the crystal fingerprint [$C(\eta)$] and direction-resolved atomic fingerprint [$\boldsymbol{V}_i(\eta)$]. Second, 100 training configurations for energy and 100 [for (i) and (ii)] or 750 [for (iii) and (iv)] training configurations for the force model, were randomly selected. Using the above parameters as input to the learning algorithm, we predict energy and forces for the four test cases of elemental Al, as shown in Figure 5. Each prediction takes roughly a millisecond (for comparison, the 32 atom bulk Al case with DFT takes ≈ 45 min on a 16 core machine, a speed up on the order of 10^6). Our ML predictions agree well with the QM data, with the observed errors (< 1 $\frac{\text{meV}}{\text{atom}}$ and < 0.05 $\frac{\text{eV}}{\text{Å}}$) reported in Table 1. This suggests well learned models in all the cases. Errors of this magnitude are comparable to errors arising within the approximations made within DFT itself. It is accuracy at this level that allows us to bypass expensive QM methods and rely on the proposed learning approach for quick energy and force predictions. However, to build a self evolving learning method (as we propose in Fig. 1c),

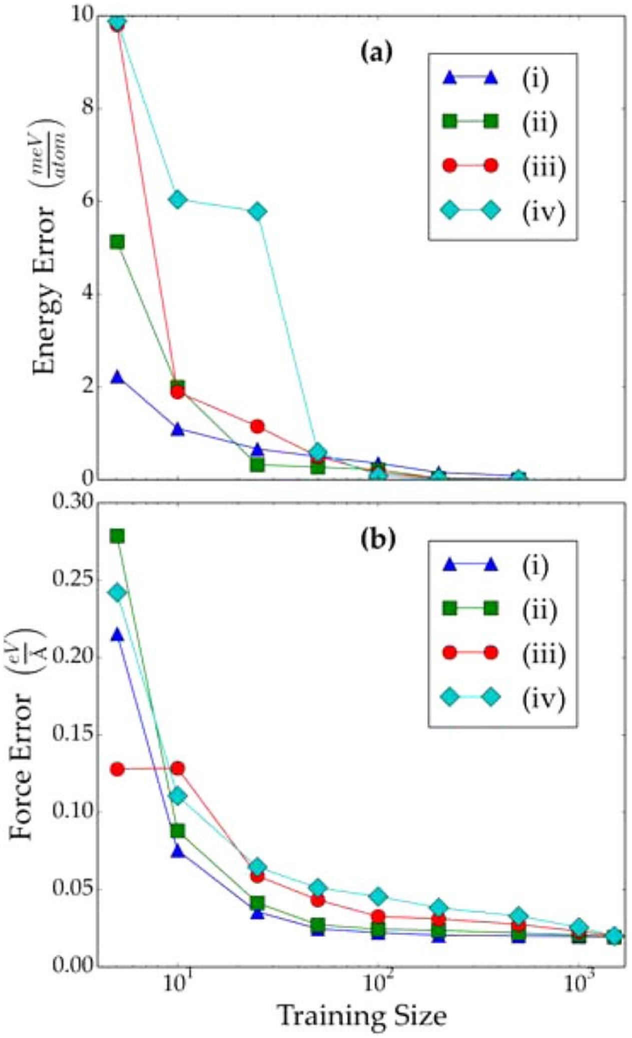

Figure 4. Energy a) and force b) error versus training size for (i) defect-free bulk Al, (ii) bulk Al containing a vacancy, (iii) a clean (111) Al surface, and (iv) the (111) surface with an Al adatom.

Table 1. Mean absolute error in energy and force predictions of the four cases.

Case	Energy ($\frac{\text{meV}}{\text{atom}}$)	Force ($\frac{\text{eV}}{\text{Å}}$)
(i) Defect-free bulk Al	**0.04** (0.03)	**0.02** (0.02)
(ii) Bulk Al w. vacancy	**0.06** (0.02)	**0.02** (0.02)
(iii) Clean (111) Al surface	**0.16** (0.08)	**0.03** (0.02)
(iv) (111) Surface w. adatom	**0.22** (0.07)	**0.03** (0.03)

Test error in bold and training error in brackets.

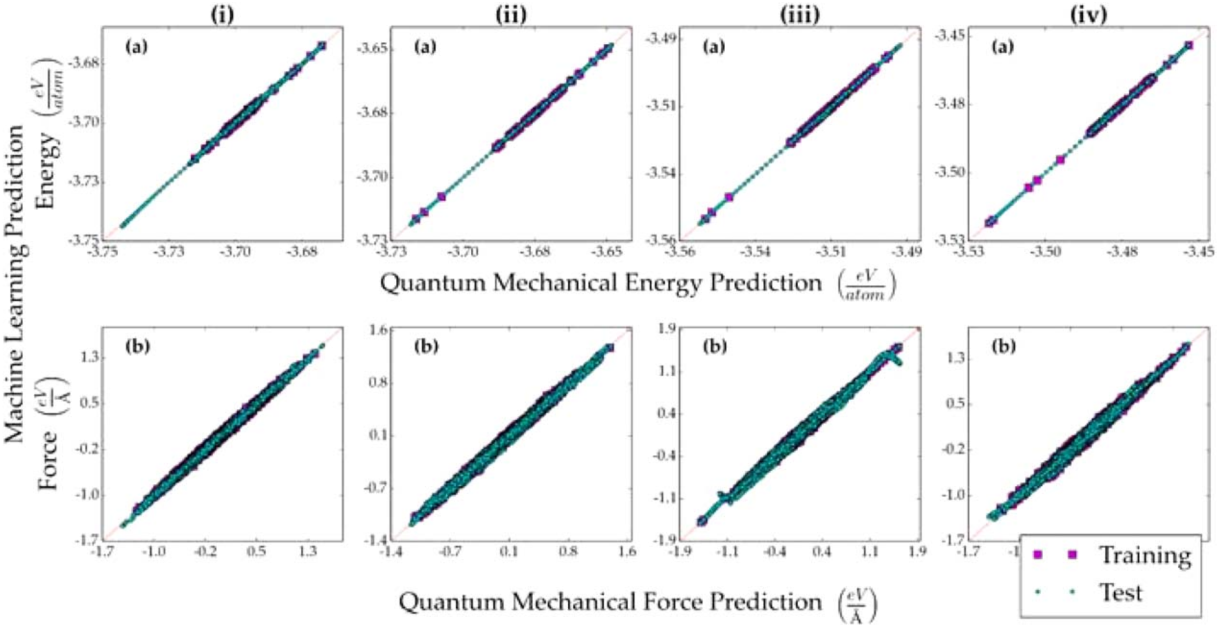

Figure 5. Parity plot for (i) defect-free bulk Al, (ii) bulk Al containing a vacancy, (iii) a clean (111) Al surface, and (iv) the (111) surface with an Al adatom, with energy a) and force b) predictions in the top and bottom rows, respectively. An eight component fingerprint, with 100 training configurations for the energy models and 100 [for (i) and (ii)] and 750 [for (iii) and (iv)] training configurations for the force models were used.

that adapts during the course of a simulation requires a scheme able to recognize situations that are outside the original training domain.

Decisions on Predictability

ML methods are, in general, interpolative and are unable to handle situations outside the training domain. To demonstrate such a situation within the context of this work, a series of configurations that mimic the migration trajectory of a vacancy in bulk Al were generated as shown in Figure 6. The energy and forces for each configuration along the migration trajectory were predicted using QM and our ML model. Given the short time span explored while generating the training data (case ii), no such migration event was actually observed. Thus, configurations close the transitions state (TS) should be inaccurately predicted by ML.

Figure 7a plots the true (QM) and predicted (ML) energy of each configuration along the migration trajectory, with the TS at the apex. Clearly, the starting and ending configurations are predicted well (as they resemble those in the training dataset). However, the error increases significantly as we move toward the TS, as these configurations were never sampled during training. On adding just the TS configuration to the training database and retraining, the error along the entire trajectory drops within acceptable accuracy (Fig. 7b). Adding more configurations along the migration pathway to the training dataset and retraining further refines the energy predictions even more (Fig. 7c). The configurations added for retraining are indicated by ⋆ in Figures 7b and 7c. Interestingly, as can be seen in Figure 8a, the atomic forces of all configurations along the trajectory are accurately predicted with error $<0.05\ \frac{\text{eV}}{\text{Å}}$, without any retraining.

To illustrate how one can detect whether the properties of a structure are predictable or not, we used the method discussed in Decision Engine: Fingerprint range section. A plot of the relative location of each crystal fingerprint component compared with the training dataset bounds (maximum and minimum value given by the red and blue dotted lines is shown in Fig. 7d). In the retrained models (only including the TS configuration, Fig. 7e, and including the TS with other configurations, Fig. 7f) the crystal fingerprint components approach the training dataset bounds, and the error drops as a result. With the forces however, all the atomic

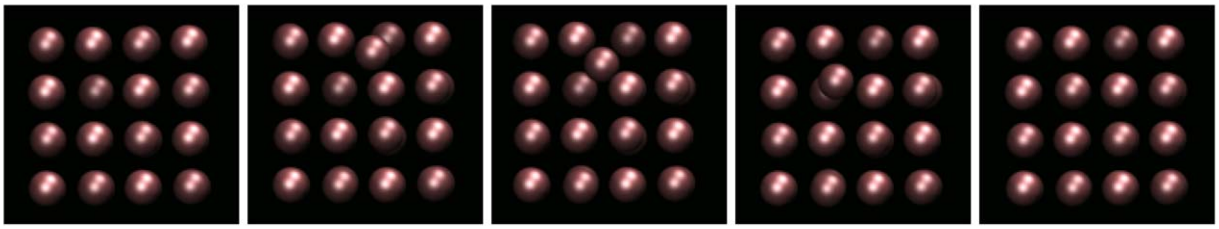

Figure 6. Vacancy migration within bulk Al. The structures shown correspond to steps 1, 5, 10, 15, and 20 along the 20-step trajectory.

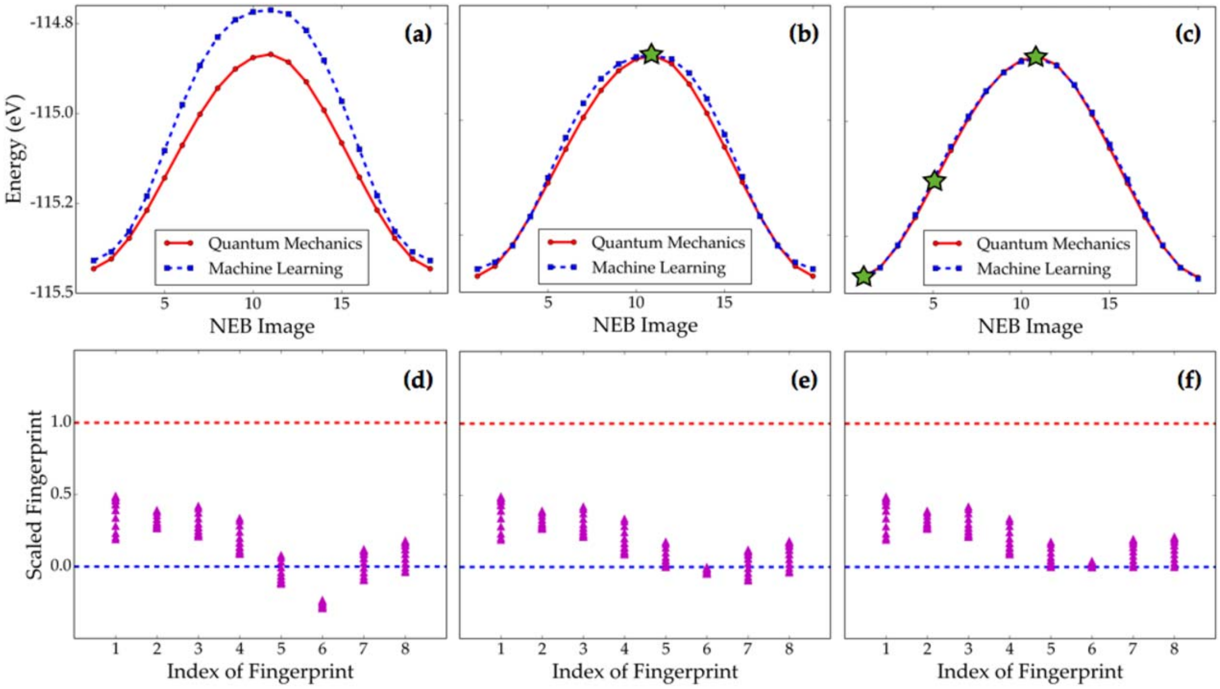

Figure 7. QM and ML energy, a)–c), and the range of crystal fingerprint components with respect to the training dataset, d)–f), of each image along the vacancy migration trajectory. a) and d) with no retraining, b) and e) with the TS added to training and c) and f) with TS and image 1 and 5 added to the training. ⋆ indicates the configurations added during retraining.

fingerprint components in the migration trajectory fall within training dataset bounds even before training, as shown in Figure 8b. Therefore, the predicted force errors are negligible as seen in the parity plot of Figure 8a. The proposed decision engine is a rudimentary but an effective approach to recognize structures which may fall outside the original training domain.

Implications of this Work

Thus far, we have demonstrated that energies and atomic forces may be predicted with chemical accuracy using a ML algorithm trained on QM data. Critical to this capability is the representation of atomic configurations and environments using continuous numerical fingerprints. Here, we have presented a class of simple, intuitive, efficient, and elegant fingerprints that can capture scalar (e.g., energy) and vector (e.g., force) quantities. We also presented a scheme that can recognize new cases not already in the training domain, which can subsequently be included in the training process thus making the prediction scheme adaptive and the predictive power monotonically increasing in quality.

All the ingredients required to eliminate (expensive) redundancies that plague *ab initio* MD simulations and hence accelerate them significantly are thus in place. The scheme proposed here, shown in Figure 1c, closely integrates with an existing DFT code; this will allow the learning scheme to become adaptive on-the-fly, and significantly mitigate the time-scale challenge that *ab initio* MD schemes currently face (although care must be taken to insure that the scheme preserves ergodicity and that ensemble averages are properly represented). As several such simulations are performed for a particular system, the accumulated information (i.e., fingerprints, forces and energies), if diverse, can lead to the creation of a force-field, using which subsequent simulations can be performed without the need for an explicit DFT engine (this is in the spirit of recent ML-based force-field development efforts[35,37]). Indeed, this is particularly true with the forces and the force fingerprints, $\boldsymbol{V}_i(\eta)$, which are purely functions of the atomic environment, unlike the total potential energy and the crystal fingerprint, which are functions of the supercell as a whole. Thus, a scheme purely based on the forces (which is conceivable as energies can be obtained from the forces through integration) does not have to be linked to a particular supercell. Such a development can mitigate the length-scale challenge faced by *ab initio* MD.

The present work may also impact non-MD simulations. For instance, structure prediction schemes require either total potential energies or total potential energies and forces.[26,53,54] A scheme analogous to the flowchart of Figure 1c can be conceived for an adaptive on-the-fly ML scheme to accelerate structure prediction calculations (or even standalone schemes once sufficient history is accumulated, as discussed above). Going further, the same paradigm can be applied to map the fingerprints to other local and global properties of interest, such as effective charges, dipoles, polarization, band gap, dielectric constant, so forth. Finally, we note that, although the QM training data discussed here came from one flavor of DFT calculations, the present scheme is applicable to any class of data, including beyond-DFT and other more

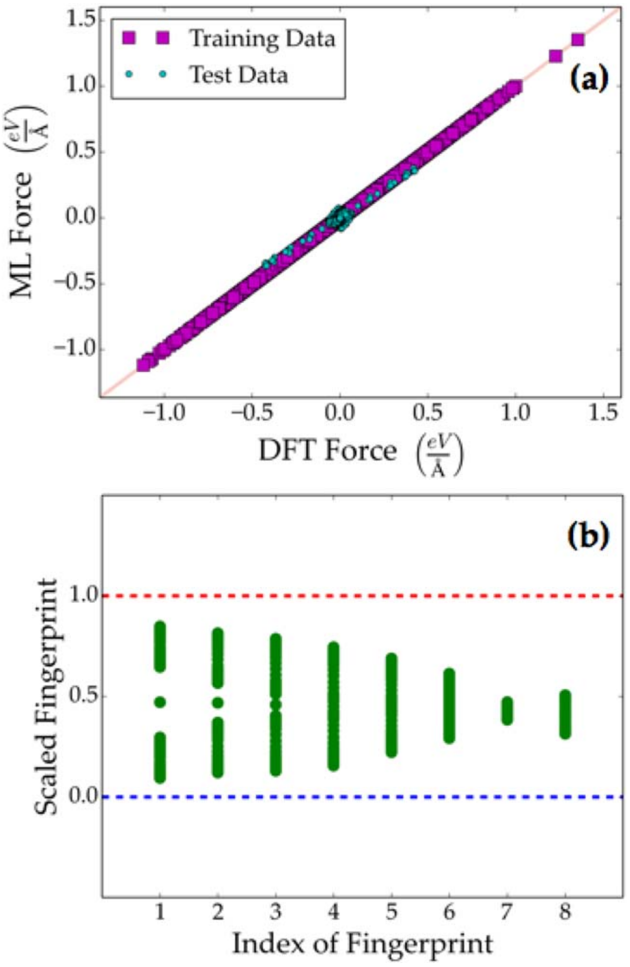

Figure 8. a) Parity plot showing accurate force prediction without any retraining, and b) Direction resolved atomic fingerprint range compared to the training dataset of the force model.

sophisticated QM methods, thus improving of the predictive power further at no extra cost (other than that incurred at the training phase). The implications of the present development are expected to be far reaching.

Summary

A detailed understanding of the dynamical evolution of materials and processes involves timescales that are beyond the reaches of present day quantum mechanical or *ab initio* MD methods. The primary causes of the bottlenecks in such approaches are the expensive and repetitive energy and force computations required, and the small timesteps involved. Acceleration schemes proffered thus far either do not preserve the fidelity of the time evolution, or have very limited domains of applicability.

In this contribution, we presented a scheme that can enormously accelerate MD simulations while still preserving the fidelity of the time-evolution, and allow us to span timescales previously inaccessible at the *ab initio* level of accuracy. The basic premise of this work is that similar configurations are constantly visited during the course of an MD simulation, and that the redundancies implicit in conventional *ab initio* MD schemes can be systematically eliminated. The foundations for such an accelerated *ab initio* MD scheme is laid out here. A ML scheme is proposed which learns from previously visited configurations in a continuous and adaptive manner on-the-fly, and predicts the energies and forces of a new configuration at a minuscule fraction of the time taken by conventional *ab initio* methods. Key elements of this new accelerated *ab initio* MD paradigm include representations of atomic configurations by numerical fingerprints, the learning algorithm, a decision engine that guides the choice of the prediction scheme, and, of course, the requisite amount of *ab initio* (re)training data.

The performance of each aspect of the proposed *ab initio* MD acceleration scheme is critically evaluated for Al, a model elemental system, in several different chemical environments, including defect-free bulk, bulk with a vacancy, clean (111) surface, and the (111) surface with an adatom. The robust configurational fingerprints utilized, and the learning algorithm adopted lead to energy and force predictions at chemical accuracy, provided sufficient fingerprint components and *ab initio* training data are used. The simple and intuitive decision engine that guides whether ML or QM needs to be used to predict the energies and forces of a new configuration is also shown to be robust. When QM is mandated, the new results are to be used in a ML retraining step; this makes the scheme adaptive on-the-fly. With the above critical pieces in place, we have a complete prescription for a new accelerated *ab initio* MD paradigm.

The ideas contained within this manuscript, although demonstrated for just an elemental metallic system, is readily extendable and applicable to nonmetallic as well as nonelemental systems. Even though the focus of the present work is to accelerate *ab initio* MD simulations, the same adaptive strategy can be applied for the learning and prediction of other properties as well.

Acknowledgments

The authors would like to acknowledge useful discussions with Kenny Lipkowitz and Avinash M. Dongare, and a critical reading of the manuscript by Ghanshyam Pilania, Tran D. Huan, and Arun Mannodi-Kanakkithodi. Partial computational support through a Extreme Science and Engineering Discovery Environment (XSEDE) allocation is also gratefully acknowledged.

Keywords: ab initio molecular dynamics · accelerate · machine learning · fingerprint · adaptive

How to cite this article: V. Botu, R. Ramprasad. *Int. J. Quantum Chem.* **2015**, *115*, 1075–18083. DOI: 10.1002/qua.24836

[1] G. Ceder, K. Persson, *Sci. Am.* **2013**, *309*, 36.
[2] V. Sharma, C. Wang, R. G. Lorenzini, R. Ma, Q. Zhu, D. W. Sinkovits, G. Pilania, A. R. Oganov, S. Kumar, G. A. Sotzing, S. A. Boggs, R. Ramprasad, *Nat. Commun.* **2014**, *5*, 4845.
[3] J. Neugebauer, T. Hickel, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3*, 438.

[4] G. Hautier, A. Jain, S. P. Ong, *J. Mater. Sci.* **2012**, *47*, 7317.
[5] A. D. Becke, *J. Chem. Phys.* **2014**, *140*, 18A301.
[6] R. Petrenko, J. Meller, Encyclopedia of Life Sciences, Chap. Molecular Dynamics; Wiley, **2010**.
[7] I. M. Torrens, Interatomic Potentials; Academic Press, **1972**.
[8] G. Henkelman, H. Jonsson, *J. Chem. Phys.* **2001**, *115*, 9657.
[9] A. Chatterjee, D. G. Vlachos, *J. Comput?Aided Mater.* **2007**, *14*, 253.
[10] A. Laio, M. Parrinello, *Proc. Natl. Acad. Sci.* **2002**, *99*, 12562.
[11] A. Laio, F. L. Gervasio, *Rep. Prog. Phys.* **2008**, *71*, 126601.
[12] M. R. Sorensen, A. F. Voter, *J. Chem. Phys.* **2000**, *112*, 9599.
[13] F. Voter, *J. Chem. Phys.* **1997**, *106*, 4665.
[14] F. Voter, F. Montalenti, T. C. Germann, *Annu. Rev. Mater. Res.* **2002**, *32*, 321.
[15] F. Voter, M. R. Sorensen, *Mater. Res. Soc. Symp. Proc.* **1999**, *538*, 427.
[16] F. Voter, *Phys. Rev. Lett.* **1997**, *78*, 3908.
[17] D. Hamelberg, J. Mongan, J. A. McCammon, *J. Chem. Phys.* **2004**, *120*, 11919.
[18] J. A. Elliott, *Int. Mat. Rev.* **2011**, *56*, 207.
[19] G. Csanyi, T. Albaret, M. C. Payne, A. D. Vita, *New J. Phys.* **2013**, *15*, 095003.
[20] T. Mueller, A. G. Kusne, R. Ramprasad, Machine Learning in Materials Science: Recent Progress and Emerging Applications, in *Reviews in Computational Chemistry*; K. B. Lipkowitz, A. L. Parrill-Baker, Eds., Wiley, **2015**.
[21] S. Srinivas, K. Rajan, *Materials* **2013**, *6*, 279.
[22] C. J. Long, J. Hattrick?Simpers, M. Murakami, R. C. Srivastava, I. Takeuchi, V. L. Karen, X. Li, *Rev. Sci. Instrum.* **2007**, *78*, 072217.
[23] G. Hautier, C. C. Fisher, A. Jain, T. Mueller, G. Ceder, *Chem. Mater.* **2010**, *22*, 3762.
[24] C. C. Fischer, K. J. Tibbetts, D. Morgan, G. Ceder, *Nat. Mater.* **2006**, *5*, 641.
[25] X. Zhang, L. Yu, A. Zakutayev, A. Zunger, *Adv. Funct. Mater.* **2012**, *22*, 1425.
[26] A. R. Oganov, Y. Ma, A. O. Lyakhov, M. Valle, C. Gatti, *Rev. Mineral. Geochem.* **2010**, *71*, 271.
[27] M. Rupp, A. Tkatchenko, K. R. Muller, O. A. von Lilienfeld, *Phys. Rev. Lett.* **2012**, *108*, 058301.
[28] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, *Sci. Rep.* **2013**, *3*, 2810.
[29] G. Montavon, M. Rupp, V. Gobre, A. Vazquez?Mayagoitia, K. Hansen, A. Tkatchenko, K. R. Muller, O. A. von Lilienfeld, *New J. Phys.* **2013**, *15*, 095003.
[30] P. V. Balachandran, S. R. Broderick, K. Rajan, *Proc. R. Soc. A* **2011**, *467*, 2271.
[31] E. W. Bucholtz, C. S. Kong, K. R. Marchman, W. G. Sawyer, S. R. Phillpot, S. B. Sinnot, K. Rajan, *Tribol. Lett.* **2012**, *47*, 211.
[32] I. E. Castelli, K. W. Jacobsen, *Model. Simul. Mater. Sci. Eng.* **2014**, *22*, 055007.
[33] D. Morgan, G. Ceder, S. Curtarolo, *Meas. Sci. Technol.* **2005**, *16*, 296.
[34] A. G. Kusne, T. Gao, A. Mehta, L. Ke, M. C. Nguyen, K. M. Ho, V. Antropov, C. Z. Wang, M. J. Kramer, C. Long, I. Takeuchi, *Sci. Rep.* **2014**, *4*, 6367.
[35] J. Behler, *Phys. Chem. Chem. Phys.* **2011**, *13*, 17930.
[36] A. P. Bartok, M. C. Payne, R. Kondor, G. Csanyi, *Phys. Rev. Lett.* **2010**, *104*, 136403.
[37] A. P. Bartok, R. Kondor, G. Csanyi, *Phys. Rev. B* **2013**, *87*, 184115.
[38] L. Yang, S. Dacek, G. Ceder, *Phys. Rev. B* **2014**, *90*, 054102.
[39] K. T. Schutt, H. Glawe, F. Brockherde, A. Sanna, K. R. Muller, E. K. U. Gross, *Phys. Rev. B.* **2014**, *89*, 205118.
[40] J. Behler, *J. Chem. Phys.* **2011**, *134*, 074106.
[41] T. Hofmann, B. Scholkopf, A. J. Smola, *Ann. Statist.* **2008**, *36*, 1171.
[42] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, B. Scholkopf, *IEEE Trans Neural Netw* **2001**, *12*, 181.
[43] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed.; Springer: New York, **2009**.
[44] H. Kaneko, K. Funatsu, *J. Chem. Inf. Model.* **2014**, *54*, 2469.
[45] P. Carrio, M. Pinto, G. Ecker, F. Sanz, M. Pastor, *J. Chem. Inf. Model.* **2014**, *54*, 1500.
[46] R. P. Sheridan, *J. Chem. Inf. Model.* **2013**, *53*, 2837.
[47] G. Kresse, J. Furthmuller, *Phys. Rev. B* **1996**, *54*, 11169.
[48] G. Kresse, D. Joubert, *Phys. Rev. B* **1999**, *59*, 1758.
[49] P. E. Blochl, *Phys. Rev. B* **1994**, *50*, 17953.
[50] J. P. Perdew, K. Burke, Y. Wang, *Phys. Rev. B* **1996**, *54*, 16533.
[51] K. Hansen, G. Montavon, F. Biegler, S. Fazil, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, K. Muller, *J. Chem. Theory Comput.* **2013**, *9*, 3404.
[52] I. H. Witten, E. Frank, M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques; Elsevier, **2011**.
[53] S. Goedecker, *J. Chem. Phys.* **2004**, *120*, 9911.
[54] D. J. Wales, J. P. K. Doye, *J. Phys. Chem. A* **1997**, *191*, 5111.

Received: 10 October 2014
Revised: 13 November 2014
Accepted: 17 November 2014
Published online 23 December 2014