Atomic energy mapping of neural network potential

Dongsun Yoo,[1] Kyuhyun Lee,[1] Wonseok Jeong,[1] Dongheon Lee,[1] Satoshi Watanabe,[2] and Seungwu Han[1,*]
[1]*Department of Materials Science and Engineering, Seoul National University, Seoul 08826, Korea*
[2]*Department of Materials Engineering, The University of Tokyo, Bunkyo, Tokyo 113-8656, Japan*

(Received 12 March 2019; revised manuscript received 25 July 2019; published 3 September 2019)

We investigate the atomic energy mapping inferred by machine-learning potentials, in particular neural network potentials. We first show that the transferable atomic energy can be defined within the density functional theory, which means that the core of machine-learning potentials is to deduce a reference atomic-energy function from the given set of total energies. By utilizing invariant points in the feature space at which the atomic energy has a fixed reference value, we examine the atomic energy mapping of neural network potentials. Examples on Si consistently support that NNPs are capable of learning correct atomic energies. However, we also find that the neural network potential is vulnerable to 'ad hoc' mapping in which the total energy appears to be trained accurately while the atomic energy mapping is incorrect in spite of its capability. We show that the energy mapping can be improved by choosing the training set carefully and monitoring the atomic energy at the invariant points during the training procedure. The energy mapping in multicomponent systems is also discussed.

DOI: 10.1103/PhysRevMaterials.3.093802

I. INTRODUCTION

Recently, machine-learning (ML) approaches to developing interatomic potentials are attracting considerable attention because they are poised to overcome the major shortcoming inherent to the classical potential and first-principles method, i.e., difficulty in potential development and huge computational cost, respectively. Favored ML models are the neural network [1–3] and kernel-based models [4,5]. In particular, the high-dimensional neural network potential (NNP) suggested by Behler and Parrinello [1] is attracting wide interest with applications demonstrated over various materials encompassing metals [6–8], insulators [9,10], semiconductors [11,12], and molecular clusters [13].

Based on the original idea, several improvements for NNP have been also put forward. For instance, genetic algorithms and CUR decomposition were applied to optimize feature sets [14–16]. To improve the performance of NNP over complex training sets, various modifications to the model structure were also proposed such as stratified NN [17], implanted NN [18], and a mixture model [19]. Recently, we reported that typical training sets are biased toward specific configurations, which undermines stability and transferability of NNP [20]. We also suggested a weighting scheme that can overcome the sampling bias [20].

While the methodological advances are under rapid progress as above, the conceptual foundation of NNP is still elusive, partly due to the black-box nature of the neural network. Furthermore, most machine-learning potentials including NNP infer atomic energies while it is trained over total energies that are sums of atomic energies, which is an unconventional machine-learning structure. However, the learning quality of NNP is examined by total energies and their derivatives, and less attention has been paid to the atomic energy function that is actually learned and inferred by NNP. Motivated by this, in this paper, we investigate the atomic energy mapping of NNP.

We show that the core of training procedure in NNP is to infer the reference atomic energy grounded on the density functional theory (DFT), from the given relationship between the structure and total energy. With examples on Si, we demonstrate that NNP can learn atomic energies correctly but it is also prone to ad hoc mapping in which the total energy is trained accurately but the atomic energy mapping is incorrect.

The rest of the paper is organized as follows: the computational details are provided in the methods section. In the results and discussion section, we first explicitly show the existence of atomic energy within DFT. Then, we provide different examples of atomic energy mapping in the following subsections and discuss implications to multicomponent systems. We then summarize and conclude.

II. METHODS

Throughout the paper, NNPs are trained by using the in-house code named SIMPLE-NN [21] (https://github.com/MDIL-SNU/SIMPLE-NN). To represent local environment, we use atom-centered symmetry functions [22]. The symmetry function vector $\mathbf{G}$ consists of 8 G_2 and 18 G_4 functions with the cutoff of 6.0–6.5 Å. Neural networks consist of two hidden layers and 30 hidden nodes per layer (26-30-30-1 structure). Errors in both total energy and atomic force are minimized during the training process employing L-BFGS and Adam optimizers.

All DFT calculations for generating training sets are carried out using Vienna *ab initio* simulation package (VASP) [23–25] with the computational setting identical to those in Ref. [20]. The classical MD simulation on Ni nanocluster is performed by the LAMMPS package [26]. More details

*hansw@snu.ac.kr

on the training set are provided in each example and information on the training process such as learning curves and error distributions are available in the Supplemental Material [27]. Although the number of structures in the training set is small compared to other studies (see below), the size is sufficient because model systems in the present study are relatively simple and atomic forces as well as total energies are used in training (see the Supplemental Material [27] for the convergence of errors against the size of the training set). We estimate the prediction uncertainty by training NNP five times with different initial conditions (the initial weights are randomized and the training set is also randomly selected from the total dataset each time). The uncertainty is estimated as one standard deviation in results from the five NNPs.

III. RESULTS AND DISCUSSIONS

A. Existence of transferable atomic energy within DFT

Most ML potentials are based on the representability of the DFT total energy ($E_{\rm tot}^{\rm DFT}$) as a sum of the atomic energy ($E_{\rm at}$) that depends on the local environment within a certain cutoff radius ($R_{\rm c}$):

$$E_{\rm tot}^{\rm DFT} = \sum_i E_{\rm at}(\mathcal{R}_i; R_{\rm c}), \tag{1}$$

where i is the atom index and $\mathcal{R}_i$ is the collection of relative position vectors of atoms lying within $R_{\rm c}$ from the ith atom. For simplicity, we assume a unary system that is large enough that various cutoff spheres in the following discussions do not self-overlap under periodic boundary conditions and wave functions are effectively real valued.

We first scrutinize whether Eq. (1) is justified within DFT. As is well known, the total energy can be expressed by integration of the local energy density, although it is not unique [28]. Then, by partitioning the space into nonoverlapping atomic volumes, one can assign energies to each atom whose sum equals to the total energy [29,30]. However, existence of the atomic energy at the DFT level does not necessarily guarantee that it is transferable to other systems with similar local environments, which is essential for machine-learning potentials based on Eq. (1). As far as we are aware, the transferability of atomic energies has not been explicitly discussed yet. In this paper, by invoking locality or 'nearsightedness' of the electronic structure [31], which empowers the $\mathcal{O}(N)$ approach [32], we formally define $E_{\rm at}(\mathcal{R}_i; R_{\rm c})$ within DFT, which depends only on local environment and so is transferable. Below, we derive $E_{\rm at}(\mathcal{R}_i; R_{\rm c})$ from the total energy with a particular attention to the transferable range.

Within the semilocal density approximation, $E_{\rm tot}^{\rm DFT}$ can be expressed in terms of the one-electron density matrix $\rho(\mathbf{r}, \mathbf{r}')$ and the electron density $\rho(\mathbf{r}) = \rho(\mathbf{r}, \mathbf{r})$:

$$\begin{aligned} E_{\rm tot}^{\rm DFT} &= E_{\rm kin} + E_{\rm XC} + E_{\rm Coul} \\ &= -\frac{1}{2}\int \nabla_{\mathbf{r}}^2 \rho(\mathbf{r}, \mathbf{r}')|_{\mathbf{r}=\mathbf{r}'} d\mathbf{r}' \\ &\quad + \int \rho(\mathbf{r}) \varepsilon_{\rm XC}(\rho(\mathbf{r}), \nabla\rho(\mathbf{r})) d\mathbf{r} \\ &\quad + \frac{1}{2}\int \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r} d\mathbf{r}' - \sum_i \int \frac{q_i \rho(\mathbf{r})}{|\mathbf{r}-\mathbf{r}_i|} d\mathbf{r} \\ &\quad + \sum_{i>j} \frac{q_i q_j}{|\mathbf{r}_i - \mathbf{r}_j|}, \end{aligned} \tag{2}$$

where the atomic unit is used, $\varepsilon_{\rm XC}$ is the exchange-correlation energy density, q_i and $\mathbf{r}_i$ are the ionic charge and position of the ith atom, respectively. Under the assumption that $\mathcal{O}(N)$ methods, in particular the divide-and-conquer (DAC) approach [33,34], work well for given systems, we will explicitly show that (i) each energy term can be split without any loss into atomic contributions that are defined locally around each atomic site, and (ii) the atomic energy depends only on nearby atoms such that it is transferable to other systems as long as local environments are maintained.

We start with partitioning the space into atomic cells without gaps or overlapping (for instance, Voronoi cells). Let V_i be the cell enclosing the ith atom. We define $\rho_i(\mathbf{r})$ as $\rho_i(\mathbf{r}) = \rho(\mathbf{r})[\mathbf{r} \in V_i]$ where $[\ldots]$ is the Iverson bracket whose value is 1 (0) when the logical proposition in the bracket is true (false). Obviously, $\rho(\mathbf{r}) = \sum_i \rho_i(\mathbf{r})$. It is easily seen that $E_{\rm XC}$ is the sum of the atomic exchange-correlation energy ($E_{{\rm XC},i}$) that is obtained by substituting $\rho_i(\mathbf{r})$ for $\rho(\mathbf{r})$ in the integrand of $E_{\rm XC}$. As is assumed in the DAC method [33], the charge density at a certain point is influenced by only nearby atoms if the local chemical potential of electrons is fixed. This means that $\rho_i(\mathbf{r})$, and hence $E_{{\rm XC},i}$, is affected by atomic arrangements within a certain cutoff ($R_{\rm c}^1$) from $\mathbf{r}_i$.

Next, we define the total charge density in V_i: $\rho_{{\rm tot},i}(\mathbf{r}) = q_i\delta(\mathbf{r}-\mathbf{r}_i) - \rho_i(\mathbf{r})$. It is straightforward to show that $E_{\rm Coul}$ can be expressed as a summation of the atomic Coulomb energy, $E_{{\rm Coul},i}$, defined as follows:

$$\begin{aligned} E_{{\rm Coul},i} &= \frac{1}{2}\sum_{j\neq i}\int \frac{\rho_{{\rm tot},i}(\mathbf{r})\rho_{{\rm tot},j}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r} d\mathbf{r}' \\ &\quad + \frac{1}{2}\int \frac{\rho_i(\mathbf{r})\rho_i(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r} d\mathbf{r}' - \int \frac{q_i\rho_i(\mathbf{r})}{|\mathbf{r}-\mathbf{r}_i|} d\mathbf{r}. \end{aligned} \tag{3}$$

The first term on the right-hand side of Eq. (3) is long ranged, which is incompatible with the finite cutoff. Here we assume that the electrostatic interaction is effectively screened or canceled such that it is negligible beyond a certain cutoff ($R_{\rm c}^2$). This would be a reasonable assumption in condensed matters with weak ionic characters. For instance, the short-ranged NNP works well in partly ionic systems such as SiO_2 [21] and GeTe [11], implying that the Coulomb interaction is approximately short ranged in these materials. (Some implementations of NNP explicitly describe the long-range Coulomb potential, separately from short-ranged atomic energies [35,36].) Thus, we omit the Coulomb interaction between $\rho_{{\rm tot},i}$ and $\rho_{{\rm tot},j}$ if $|\mathbf{r}_j - \mathbf{r}_i| > R_{\rm c}^2$. Since $\rho_i(\mathbf{r})$ and $\rho_{{\rm tot},i}(\mathbf{r})$ are influenced by atoms within $R_{\rm c}^1$ (see above), $E_{{\rm Coul},i}$ depends on atoms inside $R_{\rm c}^1 + R_{\rm c}^2$ (neglecting the volume of V_i).

As the last step, we discuss the locality of $E_{\rm kin}$. Since $\rho(\mathbf{r}, \mathbf{r}')$ decays exponentially with $|\mathbf{r}-\mathbf{r}'|$ in insulators and metals at finite temperatures [32], one can neglect $\rho(\mathbf{r}, \mathbf{r}')$ when $|\mathbf{r}-\mathbf{r}'|$ is bigger than a cutoff ($R_{\rm c}^3$), which is utilized in the density-matrix-based DAC method [34]. Therefore, for a given position $\mathbf{r}$, $\rho(\mathbf{r}, \mathbf{r}')$ is determined by the atomic

configurations within a cutoff distance (R_c^4) from $\mathbf{r}$, which should be larger than R_c^3. With the projected density matrix $\rho_{ij}(\mathbf{r},\mathbf{r}') = \rho(\mathbf{r},\mathbf{r}')[\mathbf{r}\in V_i][\mathbf{r}'\in V_j]$, we define the atomic density matrix $\rho_{\mathrm{at},i}(\mathbf{r},\mathbf{r}')$ as follows:

$$\rho_{\mathrm{at},i}(\mathbf{r},\mathbf{r}') = \rho_{ii}(\mathbf{r},\mathbf{r}') + \frac{1}{2}\sum_{j\neq i}^{|\mathbf{r}_j-\mathbf{r}_i|<R_c^3} \rho_{ij}(\mathbf{r},\mathbf{r}')\,. \tag{4}$$

It can be shown that $\rho(\mathbf{r},\mathbf{r}') = \sum_i \rho_{\mathrm{at},i}(\mathbf{r},\mathbf{r}')$ and $\rho_{\mathrm{at},i}(\mathbf{r},\mathbf{r}')$ depends only on the atomic arrangements within R_c^4 from the ith atom (neglecting the volume of V_i). The atomic kinetic energy is then given in the following:

$$E_{\mathrm{kin},i} = -\frac{1}{2}\int \nabla_{\mathbf{r}}^2 \rho_{\mathrm{at},i}(\mathbf{r},\mathbf{r}')|_{\mathbf{r}=\mathbf{r}'} d\mathbf{r}'. \tag{5}$$

Since the kinetic-energy operator is linear, the sum of the atomic kinetic energy is equivalent to the total kinetic energy.

Combining the above analyses, the atomic energy of the ith atom formally derives from the DFT calculations:

$$E_{\mathrm{at},i} = E_{\mathrm{kin},i} + E_{\mathrm{XC},i} + E_{\mathrm{Coul},i}, \tag{6}$$

and $E_{\mathrm{tot}}^{\mathrm{DFT}} = \sum_i E_{\mathrm{at},i}$. By evaluating $E_{\mathrm{at},i}$ in various structures, one can obtain in principle the atomic energy as a continuous function of the local environment:

$$E_{\mathrm{at},i} \longrightarrow E_{\mathrm{at}}^{\mathrm{DFT}}(\mathcal{R};R_c)\,, \tag{7}$$

where $R_c = \max(R_c^1 + R_c^2, R_c^4)$. Note that the atomic energy is not unique because it depends on the way to define atomic cells.

The existence of $E_{\mathrm{at}}^{\mathrm{DFT}}$ implies that the objective of the present machine learning is to identify $E_{\mathrm{at}}^{\mathrm{DFT}}$ when only total energies are informed. This is at variance with the conventional view that NNP is merely an interpolation of given total energies [37,38]. Mathematically, the neural network has the capability to infer the underlying function when only sums of function values are provided. (See an example in Supplemental Material [27] on a piecewise cubic spline and the first example in Sec. III B.)

To reduce the huge dimension of $\mathcal{R}$ and obtain E_{at} in a computationally feasible way, two approximations are adopted. First, the cutoff radius is reduced from R_c, which should be fairly large for high accuracy, to r_c that is usually chosen to be 6–7 Å. This is a reasonable range because the chemical influence rapidly diminishes beyond this boundary. (One may have to increase r_c or give a separate treatment to incorporate longer-range interactions such as magnetic interactions or the effect of charged defects.) Second, the local environment is described by feature vectors whose dimension is significantly lower than for $\mathcal{R}$. The popular choices are smooth-overlap-of-atomic-positions (SOAP) [39] or symmetry function vectors ($\mathbf{G}$) [22]. These feature vectors also automatically incorporate rotational and translational invariance inherent to the atomic energy. Here, we employ the symmetry function. Thus,

$$E_{\mathrm{tot}}^{\mathrm{DFT}} = \sum_i E_{\mathrm{at}}^{\mathrm{DFT}}(\mathcal{R}_i;R_c) \simeq \sum_i E_{\mathrm{at}}^{\mathrm{NN}}(\mathbf{G}_i;r_c)\,. \tag{8}$$

The accuracy of NNP therefore hinges on how close $E_{\mathrm{at}}^{\mathrm{NN}}$ obtained through machine learning is to one of reference $E_{\mathrm{at}}^{\mathrm{DFT}}$'s over the configurational space spanned by the given training set. (Note that there are numerous $E_{\mathrm{at}}^{\mathrm{DFT}}$'s that are all valid.) However, since $E_{\mathrm{at}}^{\mathrm{NN}}$ is fitted to the total energies, rather than directly to $E_{\mathrm{at}}^{\mathrm{DFT}}$, the ML procedure does not necessarily guarantee sufficient accuracies in $E_{\mathrm{at}}^{\mathrm{NN}}$. That is to say, $E_{\mathrm{at}}^{\mathrm{NN}}$ can reproduce total energies in the training set precisely but deviate significantly from $E_{\mathrm{at}}^{\mathrm{DFT}}$. Indeed, we will demonstrate that NNP is vulnerable to such ad hoc energy mapping, which leads to incorrect total energies in related configurations and undermines the transferability of NNP. (The atomic force is given by $\mathbf{F}_i = -\sum_j \partial E_{\mathrm{at},j}/\partial \mathbf{r}_i$ where j ranges over the atoms within r_c from the ith atom. This is also a sum of derivatives of atomic energy functions. Therefore, fitting atomic forces does not circumvent ad hoc mapping.)

B. Examples of atomic energy mapping

Even though the existence of $E_{\mathrm{at}}^{\mathrm{DFT}}$ was shown formally in the previous section, the actual calculation of $E_{\mathrm{at}}^{\mathrm{DFT}}$ would be highly expensive. (We note a recent effort to directly train NNP over atomic energies from DFT [40].) Furthermore, a direct comparison between $E_{\mathrm{at}}^{\mathrm{NN}}$ and $E_{\mathrm{at}}^{\mathrm{DFT}}$ is unfeasible because there exist an infinite number of valid $E_{\mathrm{at}}^{\mathrm{DFT}}$ and it is difficult to know which one is chosen by NNP. This makes it hard to grade the energy mapping of $E_{\mathrm{at}}^{\mathrm{NN}}$. However, there are invariant points in the $\mathbf{G}$ space at which $E_{\mathrm{at}}^{\mathrm{DFT}}$ is uniquely defined without any degree of freedom. For instance, in the crystalline Si, all the atoms are equivalent, and so the total energy per atom is simply equal to $E_{\mathrm{at}}^{\mathrm{DFT}}$ for the corresponding $\mathbf{G}$. Transforming lattice vectors of the unit cell also results in similar conditions. The invariant $\mathbf{G}$ points allow for examining accuracy of NNPs in mapping the atomic energy.

Below, we investigate the atomic energy mapping of NNP with four examples. In the first example on a Ni nanocluster, we employ the classical embedded-atom method (EAM) in which the atomic energy can be explicitly defined and compared with the NNP results. This demonstrates the ability of NNP for mapping the atomic energy when only total energies are provided. The other three examples on Si examine the atomic energy mapping of NNPs when they are trained over total energies and forces from DFT calculations, by utilizing the invariant $\mathbf{G}$ points. To be specific, the second and third examples are about crystalline and surface models, respectively. These model systems are simple but confirm the capability of NNP to map the atomic energy at the DFT level. They also demonstrate the ad hoc mapping that arises from limitations in the training set. The last example on the Si cluster corresponds to a more practical situation, showing that the accuracy of atomic energy mapping depends on the training procedure.

1. *Classical potential*

First, we examine whether NNP can identify underlying classical potential when only total energies are provided. Specifically, we train NNP on the total energy of EAM potential [41]. The training set consists of MD snapshots of Ni_{85} nanocluster at 500 K, sampled in 10-fs interval. [See Fig. 1(a)]. After training, the root-mean-squared error (RMSE) is 0.3 meV/atom for the total energy and 0.01 eV/Å for the atomic force. The atomic energy within EAM is

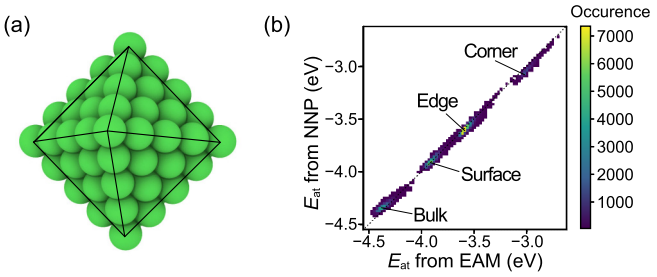

FIG. 1. (a) Structure of the Ni_{85} octahedron that consist of 6 corner, 36 edge, 24 surface, and 19 bulk atoms. (b) Correlation between atomic energy of EAM and NNP.

defined as follows:

$$E_i = F\left(\sum_{j\neq i}\rho(r_{ij})\right) + \frac{1}{2}\sum_{j\neq i}\phi(r_{ij}), \tag{9}$$

where E_i is the atomic energy of atom i, and r_{ij} is the distance between atoms i and j. In Eq. (9), F is the embedding function that depends on the sum of pairwise electron densities ρ, and ϕ is the pairwise potential. Figure 1(b) compares atomic energies between EAM and NNP. Good correlations are found and RMSE for the atomic energy is 25 meV. In particular, NNP successfully resolves different configurations, namely corner, edge, surface, and bulk, although only total energies and atomic forces are informed. This example supports that NNP can identify the atomic energy function that underlies the total energy.

2. Si crystal

Next, we train NNP over DFT energies and forces in crystalline Si. The training (validation) set consists of 350 (150) MD snapshots of the 64-atom cubic supercell under the NVT condition of 1000 K and the equilibrium volume at 0 K (sampled in 20-fs interval). After training, RMSE in the total energy and atomic force is 1 (1) meV/atom and 0.10 (0.11) eV/Å for the training (validation) set, respectively. In addition, total energies and atomic forces of NNP and reference DFT are highly correlated (see Supplemental Material [27]). Atomic vibrations during MD give rise to local expansion or compression. As a result, atomic configurations around certain Si atoms resemble those in the crystalline phase under hydrostatic pressures, which forms the equation of state (EOS) and corresponds to invariant **G** points explained above.

To show this clearly, we define $d_{NN}(\mathbf{G})$ as the Euclidean distance in the feature space from a certain **G** to the nearest point in the training set (excluding self). Figure 2(a) shows the distribution of d_{NN} for **G**'s in the training set, and Fig. 2(b) displays the training points on the two axis from the principle component analysis (PCA). The solid disks in Fig. 2(b) correspond to **G** vectors for fcc Si along the equation of states (EOS). As can be seen in Fig. 2(a), most of d_{NN} for **G**'s in the training set is lower than 0.2. In Figs. 2(b) and 2(c), the square bracket indicates the range along EOS where d_{NN} is as low as that of **G**'s in the training set (lower than 0.2), implying that **G**'s inside the square bracket would be learnable although they do not belong to the training set. Therefore, if atomic

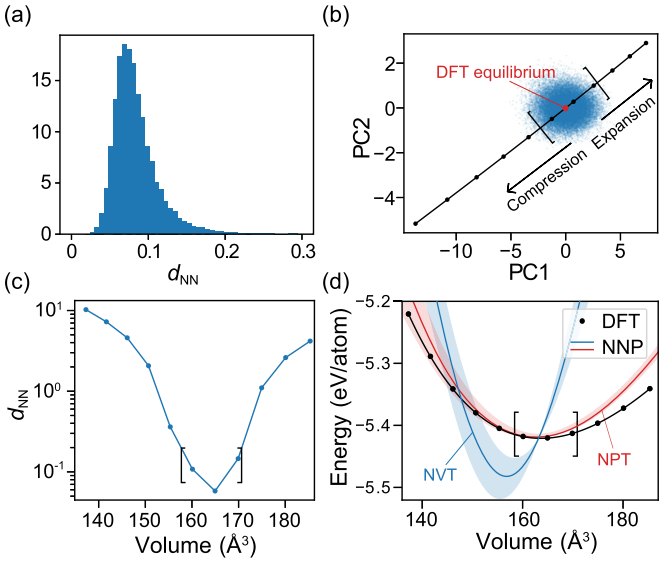

FIG. 2. (a) The distribution of d_{NN} for the training set. (b) The distribution of **G** in the training set (dots) and the equation of state (EOS) (solid disks), projected onto principal component (PC) axes. (c) The d_{NN} for each point in EOS. In both (b) and (c), the square bracket indicates the same range for EOS where the d_{NN} is lower than 0.2. (d) The EOS for Si crystal compared between DFT and NNP. The blue and red solid lines are the average EOS over five NNPs that are trained with NVT- and NPT-MD snapshots, respectively. The shades are one standard deviation from the average, corresponding to the prediction uncertainty. The squared bracket indicates the volume range where corresponding **G**'s lie in the proximity of the training set.

energies are properly mapped, NNP should be able to predict correctly the energy-volume relation at 0 K.

Figure 2(d) compares EOS inferred by the as-trained NNP (blue line) with DFT results (black dots). The light shade means prediction uncertainty evaluated by ensembles of NNP [42]. The squared bracket indicates the range of the volume whose **G** is in close proximity to the training set. Interestingly, NNP predicts correctly the energy at the equilibrium volume of DFT, but energies at other volumes significantly deviate from the DFT curve with errors far bigger than RMSE in the total energy. That is to say, NNP predicts the total energy correctly but the atomic energy is markedly wrong, which corresponds to the ad hoc energy mapping. From the continuity in E_{at}, the incorrect energy mapping should affect other training points neighboring invariant **G** points, implying that the ad hoc mapping extends over a significant portion of the training set.

The ad hoc mapping in this case happens because the training set consists of structures with a fixed volume. This condition constrains the local expansion and contraction to occur concurrently within the same structure. Consequently, any additional term in the atomic energy that varies linearly with the volume does not affect the total energy, and so the slope of EOS at the equilibrium volume of DFT becomes an arbitrary number. The ad hoc mapping in this case can be resolved by considering structures with different volumes or including virial stress in the loss function. For instance,

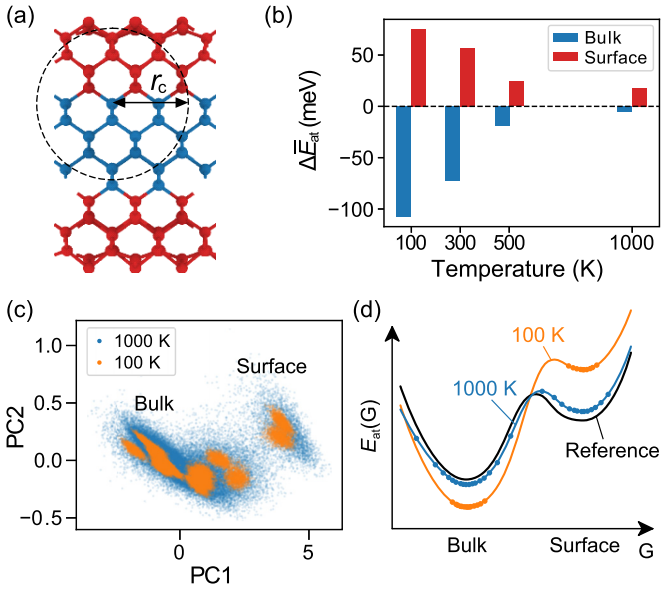

FIG. 3. (a) The structure of Si(100)-(2 × 2) slab. The atoms in bulk and surface regions are marked in blue and red, respectively. r_c is the cutoff radius of symmetry functions. (b) The average of atomic-energy difference between DFT and NNPs for bulk and surface groups, plotted against the temperature of the training set. (c) Scatter plot along principal components (PC) of **G** vectors in the training set. (d) Schematic illustration of ad hoc mapping due to separate groups of training points.

the red line in Fig. 2(d) shows EOS predicted with NNPs that are trained with MD snapshots from NPT ensembles at 1000 K and zero pressure. During MD, the supercell expands or shrinks, avoiding the exact cancellation among the local volume changes. As a result, it is seen that NNPs can predict the slope and curvature of EOS reasonably.

3. Si slab

In the third example, the training set consists of MD trajectories of Si(100)-(2 × 2) symmetric slab (128 atoms) in the NVT condition at a certain temperature between 100 and 1000 K, sampled in 20-fs interval [see Fig. 3(a)]. To assess the learning quality, we compare atomic energies for the geometry relaxed at 0 K with DFT. Unlike crystalline Si in the previous example, the reference E_{at}^{DFT} is not available directly. Nevertheless, Si atoms inside the slab (blue atoms) have neighborhood similar to that in the crystal (see a dashed circle). Therefore, E_{at} in this region should be close to the crystalline E_{at}^{DFT} at the equilibrium volume [E_{at}^{DFT}(bulk)]. Since E_{tot}^{DFT} is available for the whole structure, the average E_{at}^{DFT} for the surface region (red atoms) can be obtained as $[E_{tot}^{DFT} - N_b \cdot E_{at}^{DFT}(\text{bulk})]/N_s$, where N_b and N_s are the number of atoms in the bulk and surface regions, respectively. By taking the difference in averaged values of E_{at}^{NN} and E_{at}^{DFT} in each region, one can quantify average mapping errors, $\Delta\bar{E}_{at}$(bulk) and $\Delta\bar{E}_{at}$(surface), respectively.

Figure 3(b) presents $\Delta\bar{E}_{at}$(bulk) and $\Delta\bar{E}_{at}$(surface) for NNPs trained over MD trajectories at different temperatures. At a low temperature of 100 K, the mapping error is −108 and 76 meV/atom for bulk and surface regions, respectively,

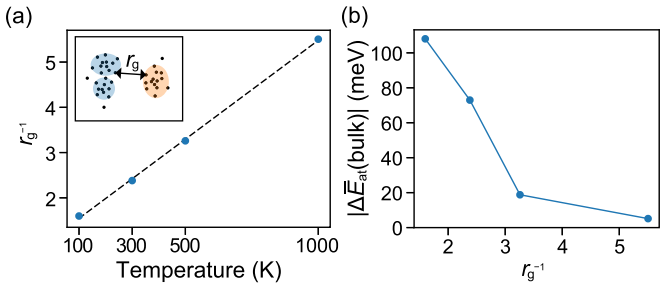

FIG. 4. (a) r_g^{-1} against the temperature of training set and (b) absolute atomic-energy error (bulk) versus r_g^{-1} for the Si slab model.

which is far bigger than RMSE (0.3 meV/atom). This is another example of ad hoc energy mapping; NNP correctly predicts the total energy because errors in the atomic energy mapping cancel with each other. In Fig. 3(b), it is intriguing that the mapping error gradually decreases as the temperature in the training set increases, and at the high temperature of 1000 K, the magnitude of mapping errors becomes comparable to RMSE in the total energy (3 meV/atom).

To understand the temperature-dependent mapping error, we examine in Fig. 3(c) the distribution of training points in the **G** space using PCA on the training sets at 100 and 1000 K. It is seen that at 100 K, the training points corresponding to the bulk and surface region are well separated. In contrast, energetic vibrations at 1000 K result in much broader distribution of training points such that bulk and surface regions are slightly connected. (Other combinations of principal axes show similar behaviors.) As schematically drawn in Fig. 3(d), if clusters of training points are separate as in 100 K, the machine learning is prone to ad hoc mapping because any canceling offsets give almost the same total energy and atomic forces. On the other hand, at higher temperatures with every region connected to some degrees, E_{at} at intermediate configurations helps adjust the energy offset between basins.

Following the above example, training points sampled on a certain type of structure should be appropriately connected to avoid the ad hoc mapping [43]. For a systematic analysis, it would be useful to develop a metric that measures the connectivity in the **G** space. To this end, we iteratively carry out single-linkage clustering of training points, a kind of hierarchical clustering used in the statistical analysis. (See Supplemental Material [27] for schematic illustration of the procedure.) At each step, two clusters with the shortest distance merge into one. (Initially, every training point represents an independent cluster.) The intercluster distance, reflecting dissimilarity between clusters, is set to the minimum Euclidean distance between two points from each cluster. We define r_g as the distance between the two lastly-linked large clusters whose size is larger than a threshold. Here, the threshold is set to 0.5 times the number of structures in the training set, but the result is largely insensitive to this value because the cluster size highly polarizes near the end of iterations. Since r_g approximates the maximum distance among cluster groups, r_g^{-1} can be regarded as the connectivity of the training set. Figure 4(a) shows r_g^{-1} against the temperature of training sets in the slab model. It is seen that r_g^{-1} increases linearly with the temperature, supporting that the training set is more

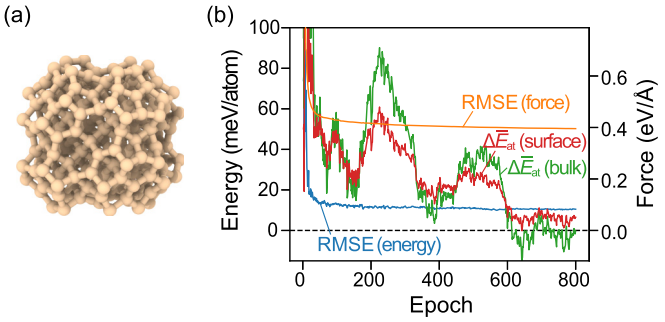

FIG. 5. (a) Si_{239} nanocluster relaxed at 0 K. (b) Change of RMSE for energy and force, and mapping errors for surface and bulk regions of Si(100)-(2 × 2) slab in Fig. 3(a), with respect to the training epoch.

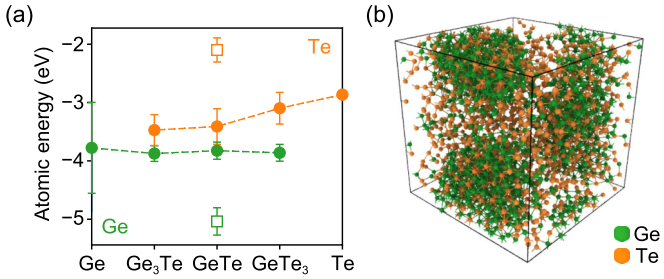

FIG. 6. (a) Average atomic energies of Ge and Te when the training set encompasses the whole composition range (Ge, Ge_3Te, GeTe, $GeTe_3$, and Te; filled circles), or when the training set includes only compositions near 1:1 (empty squares). Error bars indicate one standard deviation in atomic energies for MD snapshots. (b) The unphysical phase separation results with NNP that is trained over only compositions near 1:1.

connected at high temperatures. Figure 4(b) plots $\Delta\bar{E}_{at}$(bulk) with respect to r_g^{-1}. It is seen that the mapping error is sufficiently low when r_g^{-1} is larger than a certain cutoff. While the cutoff r_g may vary with system or training set size, it can serve as a parameter to examine the connectivity quantitatively.

4. Si nanocluster

Albeit simple, the above two cases substantiate the ad hoc mapping that originates from limitations in the training set. In practice, a single training set usually encompasses diverse structures such as bulk, surfaces, and defects, and chances are that the ad hoc mapping can be avoided in principle. Nevertheless, the error-cancelling energy offsets as in Fig. 3(d) still exist, which can go unnoticed if the training procedure is monitored by RMSE only. To show this, we generate a training set from MD simulations of a 239-atom Si nanocluster with Wulff-constructed {100}, {110}, and {111} facets at 1000–1700 K. [See Fig. 5(a) for the structure relaxed at 0 K]. The training (validation) set consists of 832 (208) MD snapshots that are sampled in the interval of 10–20 fs. The analysis on the connectivity (see above) confirms that training points are well connected.

In Fig. 5(b), we plot RMSE for the total energy and force with respect to the training epoch. It also shows $\Delta\bar{E}_{at}$(bulk) and $\Delta\bar{E}_{at}$(surface) for the (100)-(2 × 2) surface model in Fig. 3(a). The analysis similar to Fig. 2 shows that the **G** points in the (100)-(2 × 2) slab model are in the vicinity of training points, and hence they are learnable. Therefore, NNP is expected to predict surface and bulk energies in reasonable agreement with DFT results. In Fig. 5(b), it is seen that RMSE remains almost constant after about 100 epochs while $\Delta\bar{E}_{at}$(bulk) and $\Delta\bar{E}_{at}$(surface) converge at much slower rates. This indicates a risk in concluding the training convergence in terms of RMSE and supports E_{at} at invariant **G** points as alternative convergence parameters. Obviously, if the crystalline structures are included in the training set, $\Delta\bar{E}_{at}$(bulk) would converge as fast as RMSE, but this does not guarantee the proper energy mapping at other training points. Therefore, we suggest collecting invariant **G** points as a separate test set for monitoring the atomic energy mapping, rather than including them in the training set, at least in the initial stage of training.

After a sufficient number of epochs, the surface energies for (100)-(2 × 2), (110)-(2 × 1), and (111)-(2 × 1) slab models that are fully relaxed by NNP agree with DFT results within 8%. (The corresponding errors by NNP trained up to 200 epochs are within 20%.) It is intriguing that just one type of structure (nanocluster) can train NNP over such a wide range of configurations when the energy mapping is correct. This implies that NNPs with proper mapping are more transferable than those with ad hoc mapping, which may contribute to improving the stability of MD simulations [20]. It will be also useful in developing general-purpose NNPs [44]. Finally, we find that monitoring the energy mapping is helpful in selecting training parameters such as the regularization parameter of the neural network.

C. Implications for multicomponent systems

In multicomponent systems, the sum of E_{at} is uniquely defined within DFT in the high-symmetry structure (for instance, E_{at}(Ge)+E_{at}(Te) in fcc GeTe) such that it can be used as reference values to test atomic energy mapping of NNP. However, this also means that relative offsets in E_{at} among different chemical species are not uniquely defined. Nevertheless, the offset is not entirely arbitrary because physically rational distribution of the total energy should limit the atomic energy to within a certain range. If the training set consists of only structures with single stoichiometry, the energy offset among chemical types becomes arbitrary. In the example of GeTe, E_{at}(Ge) + Δ and E_{at}(Te) − Δ produce exactly the same total energies and atomic forces even for unreasonable values of Δ. This corresponds to the ad hoc mapping in multicomponent systems, which can undermine the stability of MD simulations.

We demonstrate this with the example of GeTe. (The full description including details on the training set will be published elsewhere [45].) In Fig. 6(a), the filled circles are average atomic energies of Ge and Te when the training set encompasses the whole composition range (Ge, Ge_3Te, GeTe, $GeTe_3$, and Te in solid and liquid phases). It is seen that the atomic energy changes gradually with the compositional variation. (Nevertheless, this does not mean that atomic energies are uniquely defined at mixed compositions.) In

addition, we do not encounter any particular problem during test simulations on liquid GeTe. On the other hand, the empty squares are atomic energies when the training set consists of GeTe with compositions only near 1:1. In comparison with the atomic energies of the previous NNP, the atomic energies are assigned with rather unphysical values (Ge energies are too low while Te energies are too high), which corresponds to the ad hoc mapping in multicomponent systems. When MD simulations on liquid GeTe are carried out using this NNP, we always observe unphysical phase separations as shown in Fig. 6(b), which is likely to be caused by the ad hoc mapping.

IV. CONCLUSION

In conclusion, we investigated the atomic energy mapping inferred by neural network potentials. We first showed that the transferable atomic energy can be defined within DFT. This implies that the aim of training NNP is to learn the atomic energy function defined at the DFT level from total energies, and the transferability of NNP lies in the accuracy of atomic energy mapping. The invariant **G** points with the unique $E_{\mathrm{at}}^{\mathrm{DFT}}$ provided ways to examine the atomic energy mapping. Although we were able to compare atomic energies between DFT and NNP only for spatial configurations, the examples in Sec. III B consistently support that NNPs are capable of learning correct atomic energies even for structures with nonunique energy decomposition. It was also observed that NNP is vulnerable to ad hoc mapping due to limitations in the training set and/or certain choices of computational procedure. (Classical force fields may not suffer from the ad hoc mapping because they assume pre-defined mathematical functions.) The energy mapping can be improved by choosing the training set carefully and monitoring the atomic energy at the invariant points during the training procedure. Although we focused our discussion on NNP, the conclusion is generally applicable to any machine-learning potentials that are based on Eq. (1). By clarifying what NNP actually learns, the present work will contribute to constructing accurate and transferable machine-learning potentials.

ACKNOWLEDGMENTS

This work was supported by Technology Innovation Program (10052925) by Ministry of Trade, Industry & Energy, and Creative Materials Discovery Program by the National Research Foundation of Korea (2017M3D1A1040689). The computations were carried out at Korea Institute of Science and Technology Information (KISTI) National Supercomputing Center (KSC-2018-CHA-0038).

D.Y., K.L., and W.J. contributed equally to this work.

[1] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[2] L. Zhang, J. Han, H. Wang, R. Car, and W. E, Phys. Rev. Lett. **120**, 143001 (2018).
[3] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, J. Chem. Phys. **148**, 241722 (2018).
[4] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[5] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, Nat. Commun. **9**, 3887 (2018).
[6] N. Artrith and A. M. Kolpak, Nano Lett. **14**, 2670 (2014).
[7] J. R. Boes and J. R. Kitchin, J. Phys. Chem. C **121**, 3479 (2017).
[8] H. Eshet, R. Z. Khaliullin, T. D. Kühne, J. Behler, and M. Parrinello, Phys. Rev. Lett. **108**, 115701 (2012).
[9] N. Artrith and A. Urban, Comput. Mater. Sci. **114**, 135 (2016).
[10] W. Li, Y. Ando, E. Minamitani, and S. Watanabe, J. Chem. Phys. **147**, 214106 (2017).
[11] G. C. Sosso, G. Miceli, S. Caravati, F. Giberti, J. Behler, and M. Bernasconi, J. Phys. Chem. Lett. **4**, 4241 (2013).
[12] J. Behler, R. Martoňák, D. Donadio, and M. Parrinello, Phys. Rev. Lett. **100**, 185501 (2008).
[13] B. Kolb, B. Zhao, J. Li, B. Jiang, and H. Guo, J. Chem. Phys. **144**, 224103 (2016).
[14] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, J. Chem. Phys. **148**, 241709 (2018).
[15] W. Li and Y. Ando, Phys. Chem. Chem. Phys. **20**, 30006 (2018).
[16] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, J. Chem. Phys. **148**, 241730 (2018).
[17] S. Hajinazar, J. Shao, and A. N. Kolmogorov, Phys. Rev. B **95**, 014114 (2017).
[18] B. Onat, E. D. Cubuk, B. D. Malone, and E. Kaxiras, Phys. Rev. B **97**, 094106 (2018).
[19] T. L. Pham, H. Kino, K. Terakura, T. Miyake, and H. C. Dam, J. Chem. Phys. **145**, 154103 (2016).
[20] W. Jeong, K. Lee, D. Yoo, D. Lee, and S. Han, J. Phys. Chem. C **122**, 22790 (2018).
[21] K. Lee, D. Yoo, W. Jeong, and S. Han, Comput. Phys. Commun. **242**, 95 (2019).
[22] J. Behler, J. Chem. Phys. **134**, 074106 (2011).
[23] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).
[24] G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 (1996).
[25] G. Kresse and J. Hafner, Phys. Rev. B **48**, 13115 (1993).
[26] S. Plimpton, J. Comput. Phys. **117**, 1 (1995).
[27] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevMaterials.3.093802 for the detailed information about training process, an example of piecewise cubic spline, and the illustration of single-linkage clustering to calculate r_g.
[28] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods* (Cambridge University Press, Cambridge, UK, 2004), Appendix H and references therein.
[29] M. Yu, D. R. Trinkle, and R. M. Martin, Phys. Rev. B **83**, 115113 (2011).
[30] P. L. A. Popelier, Int. J. Quantum Chem. **115**, 1005 (2015).
[31] W. Kohn, Phys. Rev. Lett. **76**, 3168 (1996).
[32] S. Goedecker, Rev. Mod. Phys. **71**, 1085 (1999).
[33] W. Yang, Phys. Rev. Lett. **66**, 1438 (1991).
[34] W. Yang and T. Lee, J. Chem. Phys. **103**, 5674 (1995).
[35] N. Artrith, T. Morawietz, and J. Behler, Phys. Rev. B **83**, 153101 (2011).
[36] T. Morawietz, V. Sharma, and J. Behler, J. Chem. Phys. **136**, 064103 (2012).
[37] J. Behler, Angew. Chem. Int. Ed. **56**, 12828 (2017).

[38] A. Khorshidi and A. A. Peterson, Comput. Phys. Commun. **207**, 310 (2016).

[39] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).

[40] Y. Huang, J. Kang, W. A. Goddard, and L.-W. Wang, Phys. Rev. B **99**, 064103 (2019).

[41] S. M. Foiles, M. I. Baskes, and M. S. Daw, Phys. Rev. B **33**, 7983 (1986).

[42] A. A. Peterson, R. Christensen, and A. Khorshidi, Phys. Chem. Chem. Phys. **19**, 10978 (2017).

[43] Alternatively, one can lift the ad-hoc mapping by adding structures that sample particular regions in the **G** space. For instance, if the slab model is trained together with the crystalline structure, the mapping error disappears regardless of temperatures in the training set. However, such a deliberate choice of training sets would be difficult in general.

[44] A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, Phys. Rev. X **8**, 041048 (2018).

[45] D. Lee, K. Lee, D. Yoo, W. Jeong, K. Lee, and S. Han (unpublished).