Active-learning-based efficient prediction of \textit{ab initio} atomic energy: a case study on a Fe random grain boundary model with millions of atoms

Tomoyuki Tamura\textsuperscript{1,2} and Masayuki Karasuyama\textsuperscript{3,2,4}

\textsuperscript{1} Department of Physical Science and Engineering, Nagoya Institute of Technology, Nagoya, 466-8555, Japan
\textsuperscript{2} Center for Materials research by Information Integration, National Institute for Materials Science (NIMS), Tsukuba 305-0047, Japan
\textsuperscript{3} Department of Computer Science, Nagoya Institute of Technology, Nagoya, 466-8555, Japan
\textsuperscript{4} PRESTO, Japan Science and Technological Agency, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan

E-mail: tamura.tomoyuki@nitech.ac.jp, karasuyama@nitech.ac.jp

Abstract. We have developed a method that can analyze large random grain boundary (GB) models with the accuracy of density functional theory (DFT) calculations using active learning. It is assumed that the atomic energy is represented by the linear regression of the atomic structural descriptor. The atomic energy is obtained through DFT calculations using a small cell extracted from a huge GB model, called replica DFT atomic energy. The uncertainty reduction (UR) approach in active learning is used to efficiently collect the training data for the atomic energy. In this approach, atomic energy is not required to search for candidate points; therefore, sequential DFT calculations are not required. This approach is suitable for massively parallel computers that can execute a large number of jobs.
simultaneously. In this study, we demonstrate the prediction of the atomic energy of a Fe random GB model containing one million atoms using the UR approach and show that the prediction error decreases more rapidly compared with random sampling. We conclude that the UR approach with replica DFT atomic energy is useful for modeling huge GBs and will be essential for modeling other structural defects.
1. Introduction

A grain boundary (GB) is the interface between two grains or crystals in a polycrystalline material. Atomic configurations and chemical bonds near GBs are distinct from those of the bulk crystal. Thus, the electrical properties of materials with GBs can greatly differ from those of a single crystal, and GBs govern a wide range of material properties [1]. A majority of the GB research is based on the coincidence site lattice (CSL) theory; a CSL GB is a simplified model with regularity and is usually characterized by the $\Sigma$ value, which is defined as the reciprocal of the density of the coincidence sites. To understand the atomic and electronic structures of CSL GBs, experiments using high-resolution electron microscopies and computer simulations using empirical potentials and first-principles calculations have been utilized. However, most GBs in actual materials are random, with no regularity, and have local amorphous structures. As it is difficult to identify atomic arrangements in amorphous phases using experimental observations, computer simulations play a major role. To study random GBs based on simulations, a large supercell containing a large number of atoms is required, which results in a high computational cost. Although a classical molecular dynamics (MD) simulation using a model that contains billions of atoms can be performed using the latest supercomputers, results obtained from the simulation depend on the parameters associated with the empirical potentials. Therefore, first-principles calculations are required, but modeling random GBs is not realistic, as it incurs huge computational costs, which include computational time and memory size. First-principles plane-wave density functional theory (DFT) calculations are widely used to identify defects in materials. Various $O(N)$ DFT methods have been developed [2, 3, 4], but it remains impossible to model a huge GB.

Iron and its alloys, because of their high strength and toughness, play important roles as structural materials in industries, infrastructures, and our daily lives. The high strength and toughness of these polycrystalline materials are strongly affected by their GBs [1]. Therefore, it is necessary to clarify the correlation between GB microstructures and mechanical properties. The GBs in iron and its alloys have been studied extensively for a long time, being one
of the most studied GBs to date. Recently, Shibuta et al. performed classical MD simulations using a model containing one million and one billion atoms to investigate the nucleation process of pure Fe [5, 6]. As a result, huge random GB models have been obtained.

In plane-wave DFT calculations, the total energy is obtained as the average value of the whole system. Conversely, an attempt has been made to analyze the local physical properties using the local energy obtained by dividing the total energy into local regions [7]. Using this local-energy analysis, a large amount of local energy can be extracted from the DFT calculation of one system. However, the computational cost of the DFT calculation for a GB model is high; thus, exhaustive investigation is not realistic for large Σ CSL GBs. We have developed an efficient scheme to predict the GB energy, where the correlation between the local environment and the DFT atomic energy for GB atoms is estimated from a few small Σ CSL GBs with a linear regression model in machine learning, and the atomic energy and its sum, namely the GB energy, are predicted using the local environment information as a descriptor and learned parameters for any CSL GB [8]. We applied this scheme to the fcc-Al [110] tilt CSL GB and obtained good prediction results. However, as the DFT local-energy analysis requires a supercell calculation under periodic conditions, it cannot be directly applied to random GB models.

In this study, we have developed a machine-learning-based method to predict the atomic energy of a huge random GB model using the DFT local-energy analysis. A tractable replica supercell that includes the surrounding atomic arrangements is constructed for each atom in a random GB model, and the training data of the DFT atomic energy are collected. Similar to our previous scheme [8], the atomic energy is predicted using local environment information and learned parameters. Although this strategy is effective, the selection of the training data set can have a significant effect on the prediction accuracy. Therefore, we use an active-learning approach [9], which has been widely studied in the machine-learning community, to select an appropriate training data set. In particular, we construct the training data set so that the uncertainty of the prediction over the entire huge GB model can be minimized. Figure [1] shows a
schematic of our proposed procedure, in which a machine-learning model is built based on an atomic descriptor space. Based on the huge Fe GB model, we show that our strategy can rapidly decrease the prediction error compared with simple random sampling approaches. An important point is that the DFT calculation is not necessary to select candidate points because the uncertainty criterion does not depend on the atomic energy. In other words, sequential DFT calculations are not necessary, and it is possible to execute many calculations simultaneously using a massively parallel computer.

2. Method

2.1. GB model

In this paper, we use a model containing 1,037,880 atoms at 1,400 K [5]. The atomic configuration is relaxed with the embedded atom method (EAM) [10]. The obtained atomic configuration is visualized using the Open Visualization Tool (OVITO) [11], as shown in Fig. 2. A common neighbor analysis (CNA) [12] is then performed to identify atomic configurations. The adaptive CNA, which employs variable cutoff distances, distinguishes atomic configurations precisely as face-centered cubic (FCC), hexagonal closed pack (HCP), body-centered cubic (BCC), icosahedron (ICO), and unknown (OTH) coordination structures. Most atoms are BCC and OTH at GBs. Only a few atoms are ICO, HCP, or FCC.

2.2. Calculation of DFT-based atomic energy

In plane-wave DFT calculations, the total energy is obtained as the average value of the whole system. The supercell is divided into Bader regions around the atom [13], and the integral value in that region denotes the atomic energy $E^{\text{atom}}_i$.

$$E^{\text{tot}} = \sum_i \int_{V_i^{\text{Bader}}} \epsilon(r) \, dr = \sum_i E^{\text{atom}}_i. \tag{1}$$

This local-energy analysis scheme, incorporated in the computational code QMAS [14], has already been applied to defect systems such as fcc-Al (111) surfaces [7], fcc-Al and fcc-Cu [110] GBs [15, 16, 17], and bcc-Fe [110] GBs [18, 19]. We use the generalized gradient approximation [20] for the exchange–correlation functional and a cut-off energy of 544 eV for the valence wave
The cubic cell centered on the target atom is extracted from the huge GB model. We fixed the cell size to $10 \times 10 \times 10 \text{Å}^3$. As a very close atomic pair occurs near the edge of the cell, those with interatomic distances of $<2.2 \text{Å}$ have been removed. The atomic energy is calculated using the DFT by fixing the atomic configuration, and the bulk energy is subtracted per atom. As metallic bonding has a large screening effect, the atomic energy of the target atom can be obtained with accuracy using a small cell. We call this atomic energy the replica DFT atomic energy.

2.3. Regression model for atomic energy of the GB model

We assume that the atomic energy of the $i$-th atom $y_i^{\text{true}} = E_i^{\text{atom}}$ can be represented as

$$y_i^{\text{true}} = x_i^\top w^{\text{true}},$$

(2)

where $x_i \in \mathbb{R}^d$ is the $d$-dimensional structural descriptor vector and $w^{\text{true}} \in \mathbb{R}^d$ is the unknown parameter vector. The actual observation of the atomic energy based on the DFT calculations $y_i^{\text{DFT}}$ is assumed to contain independent Gaussian noise

$$y_i^{\text{DFT}} = x_i^\top w^{\text{true}} + \epsilon,$$

(3)

where $\epsilon \sim N(0, \sigma^2)$ and $\sigma^2$ is the variance. Suppose that $X \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^n$ are the training data set consisting of $n$ instances. The $i$-th row of $X$ is the descriptor vector $x_i^\top$, and the $i$-th element of $y$ is the calculated atomic energy $y_i^{\text{DFT}}$. Let $\hat{w}$ be the parameter vector estimated by ridge regression. Ridge regression minimizes the following objective function with a regularization parameter $\lambda$:

$$L = \|y - Xw\|_2^2 + \lambda\|w\|_2^2,$$

(4)

for which the minimizer is written as

$$\hat{w} = M^{-1}X^\top y,$$

(5)

where $M \equiv X^\top X + \lambda I$ with the identity matrix $I \in \mathbb{R}^{d \times d}$. Using the estimated $\hat{w}$, a prediction for the $j$-th atomic energy in the GB model can be obtained as

$$y_j^{\text{true}} \approx x_j^\top \hat{w}. $$

(6)
For the structural descriptor $\mathbf{x}$ of each atom, we employed the smooth overlap of atomic positions (SOAP) \[21, 22\]. The computation of the SOAP is easier than DFT calculations; thus, the SOAP for all atoms in the GB model can be computed. In our previous study \[8\], we verified that the SOAP can accurately predict the atomic energy of the fcc-Al GB, in which there occurred a remarkable charge redistribution and a bond reconstruction between interface atoms with reduced coordination numbers.

### 2.4. Sampling training data with active learning

To estimate $\hat{\mathbf{w}}$, we assume that the $n$ atomic energy values $\mathbf{y}$ are already calculated as the training data. As the computational cost of DFT calculations is expensive, the possible numbers of $n$ are usually much smaller than the number of atoms in the GB model. The prediction accuracy of the resulting model depends on the selection of the $n$ training points. *Active learning* \[9\] is a framework that provides sampling schemes of training data for machine-learning algorithms. Here, we introduce an active learning strategy that reduces the uncertainty of prediction for the GB model.

Let $\mathcal{X}$ be the set of $n$ training inputs $\mathbf{x}_i$, and $\bar{\mathcal{X}}$ be the set of all $\mathbf{x}_i$ in the entire GB model. The prediction for the $j$-th atom $\mathbf{x}_j \in \bar{\mathcal{X}}$ is given as

$$
\hat{y}_j = \mathbf{x}_j^\top \hat{\mathbf{w}} = \mathbf{x}_j^\top \left( \mathbf{M}^{-1} \mathbf{X}^\top \mathbf{y} \right).
$$

(7)

To determine effective training samples, we evaluate the uncertainty of the current regression prediction. Let $\mathbb{V}[\mathbf{a}] = \mathbb{E}[(\mathbf{a} - \mathbb{E}[\mathbf{a}]) (\mathbf{a} - \mathbb{E}[\mathbf{a}])^\top]$ be the variance–covariance matrix of a random vector $\mathbf{a} \in \mathbb{R}^n$, where $\mathbb{E}$ is the expectation. Using $\mathbb{V}[\mathbf{c}^\top \mathbf{a}] = \mathbf{c}^\top \mathbb{V}[\mathbf{a}] \mathbf{c}$ for a constant vector $\mathbf{c} \in \mathbb{R}^n$, the variance of the prediction for the $j$-th atom is

$$
\mathbb{V}[\hat{y}_j] = \mathbb{V}[\mathbf{x}_j^\top \mathbf{M}^{-1} \mathbf{X}^\top \mathbf{y}]
= \mathbf{x}_j^\top \mathbf{M}^{-1} \mathbf{X}^\top \mathbb{V}[\mathbf{y}] \mathbf{X} \mathbf{M}^{-1} \mathbf{x}_j.
$$

As the noise term $\epsilon$ is assumed to be independent for each $i$ in $\mathbf{y}_i^{\text{DFT}} = \mathbf{x}_i^\top \mathbf{w}^{\text{true}} + \epsilon$, we see $\mathbb{V}[\mathbf{y}] = \sigma^2 \mathbf{I}$. Then, we obtain

$$
\mathbb{V}[\hat{y}_j] = \sigma^2 \mathbf{x}_j^\top \left( \mathbf{M}^{-1} \mathbf{X}^\top \mathbf{X} \mathbf{M}^{-1} \right) \mathbf{x}_j.
$$

(8)
Note that the right-hand side does not contain \( y \), which means that the prediction variance does not depend on the calculated atomic energy. Suppose that we add a new candidate \( x_i \in \bar{X} - X \) into the training data, and \( \hat{w}^{(+i)} \) is the regression coefficient vector ”after” adding \( x_i \) into the training data.

As \( M \) and \( X \) are changed by the addition of \( x_i \), the variance of prediction with the updated coefficient vector is given as

\[
\mathbb{V}[x_j^T \hat{w}^{(+i)}] = \sigma^2 x_j^T (M + x_i x_i^T)^{-1} (X^T X + x_i x_i^T) (M + x_i x_i^T)^{-1} x_j. \tag{9}
\]

Then, by summing the updated variance values of all the atoms, we obtain the total uncertainty in the prediction analysis for the GB model after the addition of \( x_i \) into the training data.

\[
\sum_{x_j \in \bar{X}} \mathbb{V}[x_j^T \hat{w}^{(+i)}]. \tag{10}
\]

We iteratively add \( i \) which minimizes this score to the training data, so that the resulting regression model has smaller prediction uncertainty for the entire GB model (Note that for this purpose, \( \sigma \) is not necessary because it is common for all \( i \)). This method is called uncertainty reduction (UR). An important property of UR is that it does not require \( y \) because the variance of Eq. (9) does not depend on \( y \). We can determine a set of candidates before performing DFT calculations. Therefore, DFT calculations for the training data set can be performed in parallel.

3. Results

3.1. CNA analysis with principle component analysis

We generated a 386-dimensional SOAP vector for each atom in the GB model. We applied principal component analysis (PCA) to the original SOAP vector, by which the dimensions were reduced to 39, keeping 99.99% of the original variance. Figure 3 shows the first two principal components (PCs) with the CNA structure. This two-dimensional plot contains 92% of the variance of the original 386-dimensional space (the first PC contains 68% of the variance, and the second PC contains 24% of the variance). For each structure type, we plot at most 1000 points randomly chosen from the GB model (if a specific type has less
than 1000 points, all the points in that type are plotted). We can see that BCC, ICO, FCC, and HCP are concentrated around different centers (FCC and HCP are distributed around a similar location because of their structural similarity). OTH spreads out entirely and partially overlaps with the other known types, although it is also distributed at locations where no known structures exist.

3.2. Active learning results

3.2.1. Training and test data settings

We evaluated three sampling strategies to create a training data set. The first set contains 150 training instances, selected by UR. The second set, called Random 1, contains 150 instances randomly selected from all the atoms except for BCC (which we call Non-BCC). The third set, called Random 2, contains 10 randomly selected instances from BCC and 140 instances from Non-BCC. To create the test data set for performance evaluation regarding unseen atoms, we first define the coordination number as the number of atoms less than $r_{\text{cut}}$. For the BCC lattice, the 8 first- and 6 second-nearest neighbors must be considered, and a local cutoff is set halfway between the second and third BCC coordination shells.

$$r_{\text{cut}} = \frac{1 + \sqrt{2}}{2} a_{\text{bcc}},$$

(11)

where local $a_{\text{bcc}}$ is computed using the 14 nearest neighbors as

$$a_{\text{bcc}}^{\text{local}} = \frac{1}{2} \left[ \frac{2}{\sqrt{3}} \sum_{j=1}^{8} |r_j| + \sum_{j=9}^{14} |r_j| \right].$$

(12)

We randomly selected 10 atoms from different coordination numbers $11 \sim 18$, which resulted in 80 atoms in total. We define the set of the test atoms as $\mathcal{X}_{\text{Test}}$.

3.2.2. Comparison of atomic energy with DFT

Figure 4 shows the prediction of Random 1 and UR. We measure the root mean squared error (RMSE):

$$\sqrt{\frac{\sum_{\mathbf{x}_i \in \mathcal{X}_{\text{Test}}} (y_i^{\text{DFT}} - \mathbf{x}_i^\top \mathbf{w})^2}{|\mathcal{X}_{\text{Test}}|}},$$

(13)

and the maximum absolute error (MAE):

$$\max_{\mathbf{x}_i \in \mathcal{X}_{\text{Test}}} |y_i^{\text{DFT}} - \mathbf{x}_i^\top \mathbf{w}|.$$

(14)
The RMSE was 0.055 eV/atom for Random 1 and 0.044 eV/atom for UR. The MAE was 0.212 eV/atom for Random 1 and 0.120 eV/atom for UR. First, both predictions, Fig. 4(a) and (b), were surprisingly accurate because the training data set had only 150 atoms, which is <0.02% (0.00014 ≈ 150/1037880) of the entire GB model. Further, UR outperformed Random 1 related to both RMSE and MAE. In particular, Random 1 does not have training instances for larger-energy regions (> 0.7 eV/atom). This lack of training data negatively affected the prediction accuracy of Random 1 for larger atoms.

The transitions of RMSE and MAE are shown in Fig. 5. As the dimension of the descriptor is \( d = 39 \), the errors widely fluctuate when the size of the training data is 40. It is clear that the UR steadily achieved the lowest errors among the three strategies with respect to RMSE and MAE.

Figure 6 shows the scatter plots of the training and test data in the two-dimensional space created by PCA. We observe that the test set \( \mathcal{X}_{\text{Test}} \) is diversified in the two-dimensional space, and our accuracy analysis covers a variety of structures. The training instances of UR are widely distributed compared with Random 1, which is concentrated around the center of the plot (Fig. 6 (b)). As UR attempts to reduce the uncertainty of the sum of all the atoms, it tends to select from a wide range of the input space, which makes the resulting estimation more stable.

We can plot the distribution of the predicted atomic energy values, as shown in Fig. 7. We can observe that the values of atomic energy at the GBs are much larger than those in the bulk region, and the atomic energy of the atoms surrounding the point defects in the bulk region is slightly larger than that of the bulk.

Various machine-learning-based potentials have been proposed for the high-precision prediction of defect structures. The transferability of potentials is evaluated by the prediction error, and the total energy of the system is basically used as

\[
\text{RMSE} = \frac{1}{M} \sqrt{\frac{1}{N_j} \sum_{j=1}^{M} \left( \frac{\Delta E_{\text{total}}^j}{N_j} \right)^2} = \frac{1}{M} \sqrt{\sum_{j=1}^{M} \frac{1}{N_j} \sum_{i=1}^{N_j} \left( \Delta E_{\text{atom}}^i \right)^2}.
\]

Conversely, we evaluate the prediction error using the atomic energy values
described in Eq. (13). Based on the error evaluation using the total energy, the average error is small if most atoms are close to a bulk-like environment. Therefore, the average error in the region within the cutoff radius $r_c$ from the central atom was evaluated as

$$\Delta \tilde{E}(r_c) = \left| \frac{1}{N_{r_i < r_c}} \sum_{r_i < r_c} \Delta E_{\text{atom}}^i \right|. \quad (16)$$

From the test data set with the coordination number $11 \sim 18$, the one with large error was selected as the central atom. Figure 8 shows the $r_c$ dependence of $\Delta \tilde{E}(r_c)$. For the bcc structure, the halfway value between the second and third coordination shells is $3.46$ Å for $a_{\text{bcc}}=2.87$, as described in Eq. (11). The averaged values of $\Delta \tilde{E}$ are $55.16$ meV/atom at $r_c = 0.0$ Å and $11.48$ meV/atom at $r_c = 3.50$ Å. As an atom with the coordination number 14 is chosen from the bulk-like region, almost all the surrounding atoms contain the same error because they have the same local environment and the average error does not decrease, even if the number of atoms increases. From these analyses, we can conclude that our method can be used to predict local energies near defect structures.

In this study, we aim to improve the DFT-level atomic energy of the GB model created using the EAM potential. As the correlation information between the atomic local environment and the atomic energy is general, there are at least two significant future directions for this study. One is the possibility of developing atomic-relaxation calculations using the correlation information between the local environment and energy/force field. We will be then able to predict the stable atomic configuration using DFT calculations based on using the empirical potential. The other is the possibility of developing the prediction of the atomic configuration and the energy distribution of various lattice defects, including amorphous structures.

4. Conclusions
We developed a method that can analyze huge random GB models with the accuracy of DFT calculations using active learning. It is assumed that the atomic energy is represented by linear regression of the atomic structural descriptor. Based on the DFT calculations, the atomic energy, called the replica DFT atomic energy, is obtained using a small cell extracted from a huge GB model. UR in
active learning is used to collect efficient training data concerning the atomic energy. In this method, atomic energy is not needed to search for candidate points; thus, there is no requirement for sequential DFT calculations. This method is suitable for massively parallel computers that can execute a large number of jobs simultaneously. We demonstrate the prediction of the atomic energy of a Fe GB model containing one million atoms using the UR approach. The rate of decrease of the prediction error is further compared with random sampling. We conclude that the UR approach with the replica DFT atomic energy is useful for modeling huge GBs and will be essential for modeling other structural defects.

Acknowledgments
This work was supported by ‘Materials research by Information Integration’ Initiative (MI²I) project of the Support Program for Starting Up Innovation Hub from Japan Science and Technology Agency (JST), MEXT as a social and scientific priority issue (Creation of new functional devices and high-performance materials to support next-generation industries; CDMSI) to be tackled by using post-K computer, MEXT KAKENHI awarded to T.T. (18K04700) and M.K. (17H04694), and PRESTO awarded to M.K. (JPMJPR15N2). We would like to thank M Kohyama for helpful discussions on the ab-initio atomic-energy analysis, Y Shibuta for providing data of large-scale MD simulations and R Kobayashi for valuable information on the calculation of the SOAP.

References
[1] A. P. Sutton and R. W. Balluffi. Interfaces in crystalline materials. Clarendon Press, Oxford, 1995.
[2] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal. The SIESTA method for ab initio order-n materials simulation. J. Phys.: Condens. Matter, 14:2745–2779, 2002.
[3] http://www.openmx-square.org.
[4] http://www.order-n.org.
[5] Y. Shibuta, S. Sakane, T. Takaki, and M. Ohno. Submicrometer-scale molecular dynamics simulation of nucleation and solidification from undercooled melt: Linkage between empirical interpretation and atomistic nature. Acta Materialia, 105:328 – 337, 2016.
[6] Y. Shibuta, S. Sakane, E. Miyoshi, S. Okita, T. Takaki, and M. Ohno. Heterogeneity in homogeneous nucleation from billion-atom molecular dynamics simulation of solidification of pure metal. Nature Communications, 8:10–1–9, 2017.

[7] Y. Shiihara, M. Kohyama, and S. Ishibashi. Phys. Rev. B, 81:075441, 2010.

[8] T. Tamura, M. Karasuyama, R. Kobayashi, R. Arakawa, Y. Shiihara, and I. Takeuchi. Fast and scalable prediction of local energy at grain boundaries: machine-learning based modeling of first-principles calculations. Modell. Simul. Mater. Sci., 25:075003–1–21, 2017.

[9] Burr Settles. Active Learning. Morgan & Claypool, 2012.

[10] J. Song and W. A. Curtin. Atomic mechanism and prediction of hydrogen embrittlement in iron. Nature Materials, 12:145–151, 2013.

[11] A. Stukowski. Visualization and analysis of atomistic simulation data with ovito the open visualization tool. Modell. Simul. Mater. Sci., 18:015012–1–7, 2010.

[12] A. Stukowski. Modell. Simul. Mater. Sci., 20:045021–1–5, 2012.

[13] R. F. Bader. Atoms in Molecules: A Quantum Theory. Oxford University Press, 1990.

[14] http://www.qmas.jp.

[15] H. Wang, M. Kohyama, S. Tanaka, and Y. Shiihara. J. Phys.: Condens. Matter, 25:305006, 2013.

[16] H. Wang, M. Kohyama, S. Tanaka, and Y. Shiihara. J. Mater. Sci., 50:6864, 2015.

[17] H. Wang, M. Kohyama, S. Tanaka, and Y. Shiihara. Model. Simul. Mater. Sci. Eng., 25:015005, 2017.

[18] S.K. Bhattacharyya, S. Tanaka, Y. Shiihara, and M. Kohyama. J. Mater. Sci., 49:3980, 2014.

[19] S.K. Bhattacharyya, M. Kohyama, S. Tanaka, and Y. Shiihara. J. Phys.: Condens. Matter., 26:355005, 2014.

[20] J. P. Perdew, K. Burke, and M. Ernzerhof. Generalized gradient approximation made simple. Phys. Rev. Lett., 77:3865, 1996.

[21] Albert P Bartók and Gábor Csányi. Gaussian approximation potentials: A brief tutorial introduction. Int. J. Quant. Chem., 115(16):1051–1057, April 2015.

[22] Sandip De, Albert P Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. Phys. Chem. Chem. Phys., 18(20):13754–13769, May 2016.
Figure 1: Schematic of the non-sequential prediction of the \textit{ab initio} atomic energy of a random GB model. (a) Atomic descriptors for all atoms are calculated. (b) Training data are selected based on the uncertain reduction criterion. (c) Replica DFT atomic energy calculations are performed non-sequentially for selected atoms. (d) Using the atomic descriptors and the calculated atomic energy, the machine-learning model parameters are optimized. Then, the DFT-based atomic energy of a random GB model can be predicted quickly without time-consuming computations.
Figure 2: Atomic configurations on a plane. Blue, yellow, red, green, and gray represent atoms with BCC, ICO, HCP, FCC, and OTH configurations, respectively, as defined by the adaptive CNA.
Figure 3: PCA analysis of the SOAP descriptor, projected onto the plane of the first two PCs. Data points are the color-coded labels of the CNA structure.
Figure 4: Comparison between the DFT and predicted values of the atomic energy (eV/atom).
Figure 5: Transition of prediction error. (Left) RMSE (eV/atom) and (Right) MAE (eV/atom).
Figure 6: Training and test data in the reduced dimensional space created by PCA. The data points with the CNA structure type are the same as those in Fig. 3.
Figure 7: Distribution of the predicted atomic energy values. Energy differences from the bulk value are indicated by colors.
Figure 8: Averaged error in the region within the cutoff radius $r_c$ from the central atom.