Journal of Energy Chemistry 36 (2019) 7–14

Contents lists available at ScienceDirect

Journal of Energy Chemistry

journal homepage: www.elsevier.com/locate/jechem

http://www.journals.elsevier.com/
journal-of-energy-chemistry/

Regression model for stabilization energies associated with anion ordering in perovskite-type oxynitrides

Masanori Kaneko[a], Mikiya Fujii[a], Takashi Hisatomi[b], Koichi Yamashita[a],
Kazunari Domen[a,b,*]

[a] *Department of Chemical System Engineering, School of Engineering, The University of Tokyo, 7-3-1 Hongo Bunkyo-ku, Tokyo 113-8656, Japan*
[b] *Center for Energy & Environmental Science, Interdisciplinary Cluster for Cutting Edge Research, Shinshu University, 4-17-1 Wakasato, Nagano-shi, Nagano 380-8553, Japan*

Dedicated to the 70th anniversary of Dalian Institute of Chemical Physics, CAS, China.

ARTICLE INFO

Article history:
Received 26 November 2018
Revised 3 January 2019
Accepted 11 January 2019
Available online 16 January 2019

Keywords:
Machine learning
Total energy
Bandgap
Local anion ordering
Density functional theory

ABSTRACT

Certain perovskite-type oxynitrides have bandgaps suitable for renewable hydrogen production via photocatalytic and photoelectrochemical water splitting under visible light. Understanding the ordering of oxide and nitride anions in these materials is important because this ordering affects their semiconductor properties. However, the numerous possible orderings complicate systematic analyses based on density functional theory (DFT) calculations using defined elemental arrangements. This work shows that anion ordering in large-scale supercells within perovskite-type oxynitrides can be rapidly predicted based on machine learning, using $BaNbO_2N$ (capable of oxidizing water under irradiation up to 740 nm) as an example. Machine learning allows the calculation of the total energy of $BaNbO_2N$ directly from randomly selected initial atomic placements without costly structural optimization, thus reducing the computational cost by more than 99.99%. Combined with the Metropolis Monte Carlo method, machine learning permits exploration of the stable anion orderings of large supercells without costly DFT calculations. This work therefore demonstrates a means of predicting the properties of functional materials having complex compositions based on the most realistic elemental arrangements in conjunction with reasonable computational loads.

Kazunari Domen's Group for Sunlight-driven Renewable Hydrogen Production. Prof. Domen's primary research interest lies in overall water splitting using heterogeneous photocatalysts to generate clean and recyclable hydrogen. He focuses his effort on materials chemistry of particulate (oxy)nitride and (oxy)sulfide semiconductors suitable for water splitting under visible light, heterogeneous catalysis to improve the durability and reaction selectivity of photocatalysts, and design and development of scalable and economically viable photocatalytic systems, reactors and processes. He is leading the photocatalytic solar hydrogen production division of the Artificial Photosynthesis Project of the New Energy and Industrial Technology Development Organization (NEDO) to actualize sustainable artificial photosynthesis processes.

1. Introduction

Perovskite-type materials can have a variety of compositions and thus exhibit varying physical properties, including different bandgap energies, electronic states and formation energies [1]. In addition, some perovskite-type semiconductors have applications related to solar energy conversion. These include oxides [2,3], oxynitrides [3] and oxysulfides [4], which can be applied to photocatalytic and photoelectrochemical water splitting, and halides [5,6], which can be used for photovoltaics. Perovskite-type oxynitrides with the general formula $AB(O,N)_3$ have attracted particular interest because their bandgaps are narrower than those of the corresponding oxides [7–9]. The physical properties of these oxynitrides depend on the ordering of anions in their structures. As an example, the absorption edge wavelength of $SrTaO_2N$ has been predicted to be variable from approximately 600 to 720 nm [10] and the effective mass of charge carriers in $CaTaO_2N$ to be variable by a factor of three [11].

---

* Corresponding author at: Department of Chemical System Engineering, School of Engineering, The University of Tokyo, 7-3-1 Hongo Bunkyo-ku, Tokyo 113-8656, Japan.

E-mail address: domen@chemsys.t.u-tokyo.ac.jp (K. Domen).

https://doi.org/10.1016/j.jechem.2019.01.012
2095-4956/

Quantum chemistry calculations have been utilized to predict the physical properties of various functional materials [11–15]. In the case of perovskite-type oxynitrides, such calculations are required to identify thermodynamically stable anion ordering arrangements in various compositions of interest, because these stable arrangements are dependent on both chemical composition and temperature. However, it is impractical to employ conventional quantum chemistry calculations for this purpose because of the numerous possible anion orderings. In the case of a $3 \times 3 \times 3$ supercell within an ABO_2N system (consisting of 135 atoms), the quantity of possible anion orderings is on the order of 10^{18} (specifically, ${}_{81}C_{27}/3^3$ (translational symmetries)/24 (rotational symmetries of the regular octahedron)/2 (reflection symmetries)) at a rough estimate, and structural optimization calculations are needed for each structure before predicting its physical properties. As a result, the structural optimization step is the most computationally intensive (and hence the most expensive) aspect of the quantum chemical calculations. For this reason, calculations for only a few atomic arrangements of small supercells within perovskite-type materials have been reported to date. High-throughput screening of perovskite oxynitrides using first principles calculations has been carried out, but only limited arrangements of O and N anions were considered [16]. It is therefore essential to reduce the calculation costs associated with structural optimization so as to allow detailed theoretical predictions of the properties of materials such as these having complex compositions.

Recently, machine learning has emerged as a powerful means of screening various materials [17]. Machine learning based on linear regression, kernel ridge regression and artificial neural networks involves the processing of data to identify lurking patterns. Unlike conventional density functional theory (DFT) calculations, it is possible to efficiently predict realistic physical properties for a randomly-selected structure even if the most stable structures are not known, once proper models have been learned. In recent studies, machine learning models were developed to predict the thermodynamic phase stability of perovskite-type oxides as well as the bandgap energies of double perovskites, using datasets containing information regarding DFT calculations for more than 1900 perovskite oxides [18] and the Computational Materials Repository [19], respectively.

The present study demonstrates that the most stable anion ordering of perovskite-type oxynitrides having large supercells can be predicted using machine learning based on DFT calculations involving small supercells. In this work, the semiconducting properties of certain supercells were predicted by DFT calculations, after which a model for the prediction of the total energy was generated by machine learning. Subsequently, stable anion ordering was predicted using the Metropolis method [20]. Finally, the validity of the supercell structures obtained by the Metropolis method was confirmed by comparison with the results of DFT calculations. This work focused on anion ordering in $BaNbO_2N$ because this material is promising as a photocatalyst for solar hydrogen production [21–23]. Specifically, $BaNbO_2N$ can oxidize water under visible light irradiation up to 740 nm [21]. In addition, crystallization of the near-surface regions of $BaNbO_2N$ particles by annealing leads to a remarkable photoanodic current (attributable to water oxidation) of 5.2 mA cm^{-2} under simulated sunlight [23].

Our machine learning method is similar to the cluster expansion method [24–27] in that large systems can be handled. However, for the cluster expansion method, it is necessary to determine a crystal structure to expand the clusters, and a greater number of clusters are required when the symmetry of the crystal structure is lower. In addition, it is generally difficult to include long-range effects in heterovalent ionic systems as in oxynitrides, which could make significant contributions to electronic properties of bulk semiconductors, even though it has been attempted to include electrostatic energy and/or large number of pair clusters representing long-range interactions among clusters [28]. Comparatively, our machine learning approach can consider long-range effects in any crystal structure regardless of the symmetry by defining explanatory variables conveying chemically significant information explicitly.

2. Calculation and method details

2.1. Software and libraries

DFT calculations were performed using the VASP 5.4.4 software package [29–32], in conjunction with the Numpy [33], Scipy [34], Scikit-learn [35] and Atomic Simulation Environment (ASE) [36] libraries. Linear regression, ridge regression [37], lasso regression [38] and Random Forest [39] calculations were performed with the Scikit-learn software program. Both ridge and lasso regression are linear models that allow the suppression of over-fitting by applying a regularization term. Ridge regression considers factors with small contributions, while lasso regression does not. In contrast to the other models, Random Forest is a nonlinear method. All graphs were plotted with Gnuplot and all crystal structures were illustrated using VESTA [40].

2.2. DFT calculations

2.2.1. $BaNbO_2N$ models

Because $BaNbO_2N$ and $BaNbO_3$ both have similar cubic structures [41,42], the crystal structure of $Pm\bar{3}m$ $BaNbO_3$ was initially optimized. The results were then used to generate $BaNbO_2N$ supercells because the symmetry of the structure was well-suited to random elemental substitution. Due to the roughly equivalent unit cell structures of these two materials, the choice of the initial structure did not affect the results of the structural optimization process. In this optimization, one-third of the O atoms in four different $BaNbO_3$ supercells having different even-odd periodicities ($2 \times 2 \times 6$, $2 \times 3 \times 3$, $2 \times 3 \times 4$, and $3 \times 3 \times 3$) were randomly replaced with N atoms to generate $BaNbO_2N$ structures. The extent to which the stability of the supercells was affected by the even-odd periodicity and the span of the periodicity was subsequently assessed. In total, 140 structures were generated for each of the four types of supercells. It is considered that $BaNbO_2N$ structures in which each Nb atom is coordinated with two N atoms are more stable. Therefore, 20 structures in which 85%–100% of Nb atoms were coordinated with two N atoms and a further 20 structures in which 70%–85% of Nb atoms were coordinated with two N atoms were generated. Additionally, 100 structures with fully random O/N ordering were examined because the Metropolis method required information regarding unstable anion ordering. Overall, 560 structures were generated for the four supercells, and this quantity was sufficient to produce predictive models. In preparation for the machine learning process, these 560 structures were randomly divided into 420 structures (80%) for use as the training set and 140 structures (20%) for use as the test set.

2.2.2. Structural optimization and calculations of energy and bandgap values

Structural optimization and calculations of energy and bandgap values were carried out using the VASP software package. The projector augmented wave (PAW) [43,44] method with the regular GGA-PBE exchange-correlation functional [45,46] was used in conjunction with a cutoff energy of 520 eV. This process employed $3 \times 3 \times 1$, $3 \times 2 \times 2$, $3 \times 2 \times 2$ and $2 \times 2 \times 2$ Γ-centered k-point samplings of the Brillouin zone for the $2 \times 2 \times 6$, $2 \times 3 \times 3$, $2 \times 3 \times 4$, and $3 \times 3 \times 3$ supercells, respectively, and optimization was continued until the forces on all atoms were less than 0.05 eV/Å.

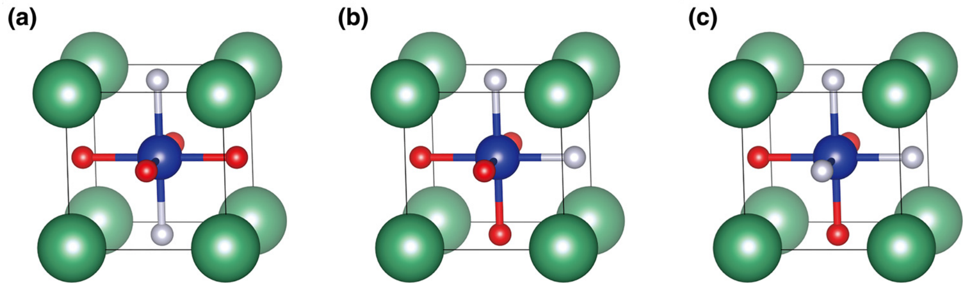


Fig. 1. $BaNbO_2N$ unit cell structures with different anion placements. Green, blue, red and gray balls represent Ba, Nb, O and N atoms, respectively.

The total energies and bandgaps of the optimized structures were calculated using $6\times6\times2$, $6\times4\times4$, $6\times4\times3$ and $4\times4\times4$ Γ-centered k-point samplings of the Brillouin zone for the $2\times2\times6$, $2\times3\times3$, $2\times3\times4$, and $3\times3\times3$ supercells, respectively. Each bandgap was calculated from the difference between the minimum and the maximum eigenvalues located above and below the Fermi level among all calculated k-points, respectively.

2.3. Machine learning

2.3.1. Explanatory and objective variables

Three explanatory variables were considered: D_{Coord}, D_{Order} and D_{exp2}. $D_{\mathrm{Coord}}(n)$ is the proportion of Nb atoms that are coordinated with n N $(0\le n\le6)$ atoms, and is associated with the electrical stability of the material. In a supercell in which each Nb atom is coordinated with two N atoms (Fig. 1(a) and (b)), $D_{\mathrm{Coord}}(2)$ is unity while the $D_{\mathrm{Coord}}(n)$ values ($n=0$, 1, 3, 4, 5 or 6) are zero.

$D_{\mathrm{Order}}(n)$ is the proportion of Nb atoms having n trans NbN chain(s) $(0\le n\le3)$. This variable represents local anion ordering and is related to the overlap of Nb 4*d* and N 2*p* orbitals. The local anion ordering can be distinguished based on the D_{Coord} and D_{Order} values. Fig. 1 shows three $BaNbO_2N$ unit cells with different anion placements. The unit cells in Fig. 1(a) and (b) have the same D_{Coord} but different D_{Order} values ($D_{\mathrm{Order}}(1)$ is one and zero, respectively). In contrast, the unit cells in Fig. 1(b) and (c) have the same D_{Order} ($D_{\mathrm{Order}}(n)=1$ when $n=0$ and otherwise $D_{\mathrm{Order}}(n)=0$), but different D_{Coord} ($D_{\mathrm{Coord}}(n)=1$ when $n=2$ and otherwise $D_{\mathrm{Order}}(n)=0$ for the former and $D_{\mathrm{Coord}}(n)=1$ when $n=3$ and otherwise $D_{\mathrm{Order}}(n)=0$ for the latter).

$D_{\mathrm{exp2}}(I, J, n_x, n_y, n_z)$ is defined as follows.

$$D_{\mathrm{exp2}}(I,J,n_x,n_y,n_z)=\frac{1}{N}\sum_{i\in I,\ j\in J}\sum_{\boldsymbol{R}}\left|x_i-x_j+R_x\right|^{n_x}\left|y_i-y_j+R_y\right|^{n_y}\left|z_i-z_j+R_z\right|^{n_z}e^{-\left|\boldsymbol{r}_i-\boldsymbol{r}_j+\boldsymbol{R}\right|^2}$$

$n_x=0$, 1, 2 or 3 and $n_y=0$, 1, 2 or 3 and $n_z=0$, 1, 2 or 3
I, J: atomic species
N: number of atoms in a supercell
$\boldsymbol{r}_i$: position vector of atom I
x_i, y_i and z_i: x, y and z components of $\boldsymbol{r}_i$, respectively
$\boldsymbol{R}$: supercell lattice vector

The variables x_i, y_i, z_i, $\boldsymbol{r}_i$ and $\boldsymbol{R}$ are normalized so that 0.5 × lattice constant is 1 and the space is dimensionless. D_{exp2} is related to the overlap of atomic orbitals, being a function of two atomic species (I and J) and various powers (n_{i}) ($i=x$, y and z), and is similar to the overlap integral of Gaussian basis functions, which are often used in first-principles calculations [47]. Therefore, D_{exp2} can express the effects of chemical bonds and long-range interactions. All the explanatory variables were standardized using the training set. Employing the present method, the values of these explanatory variables could be calculated more than 10,000 times faster than possible when using the standard DFT approach. That is, in the same amount of CPU time, it was possible to consider 100 times the quantity of supercells by calculating the explanatory variables rather than running DFT calculations. During this process, the total energy per atom and the bandgap of the optimized structure were selected as objective variables. Our machine learning model predicts the properties of relaxed structures from explanatory variables of unoptimized cubic supercell structures based on an assumption that the atomic arrangement of unrelaxed structures and the atomic coordinates (including lattice vector) of relaxed structures have a one-to-one relationship. Therefore, the explanatory variables contain only information of the atomic arrangement. This assumption is considered to be correct taking the goodness of fit of the model (see Section 3.2).

2.3.2. Optimization of hyperparameters

Hyperparameters included in ridge regression, lasso regression and Random Forest calculations were determined by five-fold cross validation (CV). The original training set of 420 supercells was divided into a training set of 336 structures (80%) for fitting and a validation set of 84 structures (20%) for the calculation of the root mean square error (RMSE). This operation was repeated five times while exchanging the training and validation sets, and the average of the RMSE values was used as the estimated RMSE for the predictive model. In this approach, hyperparameters were selected so as to minimize the estimated RMSE.

2.4. Search for stable anion ordering

Based on a predictive model for total energy obtained from the machine learning process, stable anion orderings were searched for using the Metropolis method. The inverse temperature β was set to 50,000 eV^{-1} atom. Eight $3\times3\times3$ $BaNbO_2N$ supercell structures, in which O and N atoms were randomly ordered, were employed as initial structures. The simulation was stopped after 6000 steps because the total energy was sufficiently converged, as described in Section 3.3.

3. Results and discussion

3.1. Validity of explanatory variables

Fig. 2 shows the relationship between the D_{Coord} and total energy values for the 420 $BaNbO_2N$ supercells in the original training set as calculated by DFT in association with structural optimization. From Fig. 2(c), it is evident that the total energy decreased with increasing $D_{\mathrm{Coord}}(2)$, indicating that the $BaNbO_2N$ supercells were more stable when each Nb atom was coordinated with two N atoms, which in turn is associated with the spatial dispersion of N atoms in the supercell. Conversely, the total energy tended to increase along with increasing $D_{\mathrm{Coord}}(n)$ for $n\neq2$. Therefore, the number of N atoms coordinated with each Nb atom is expected to

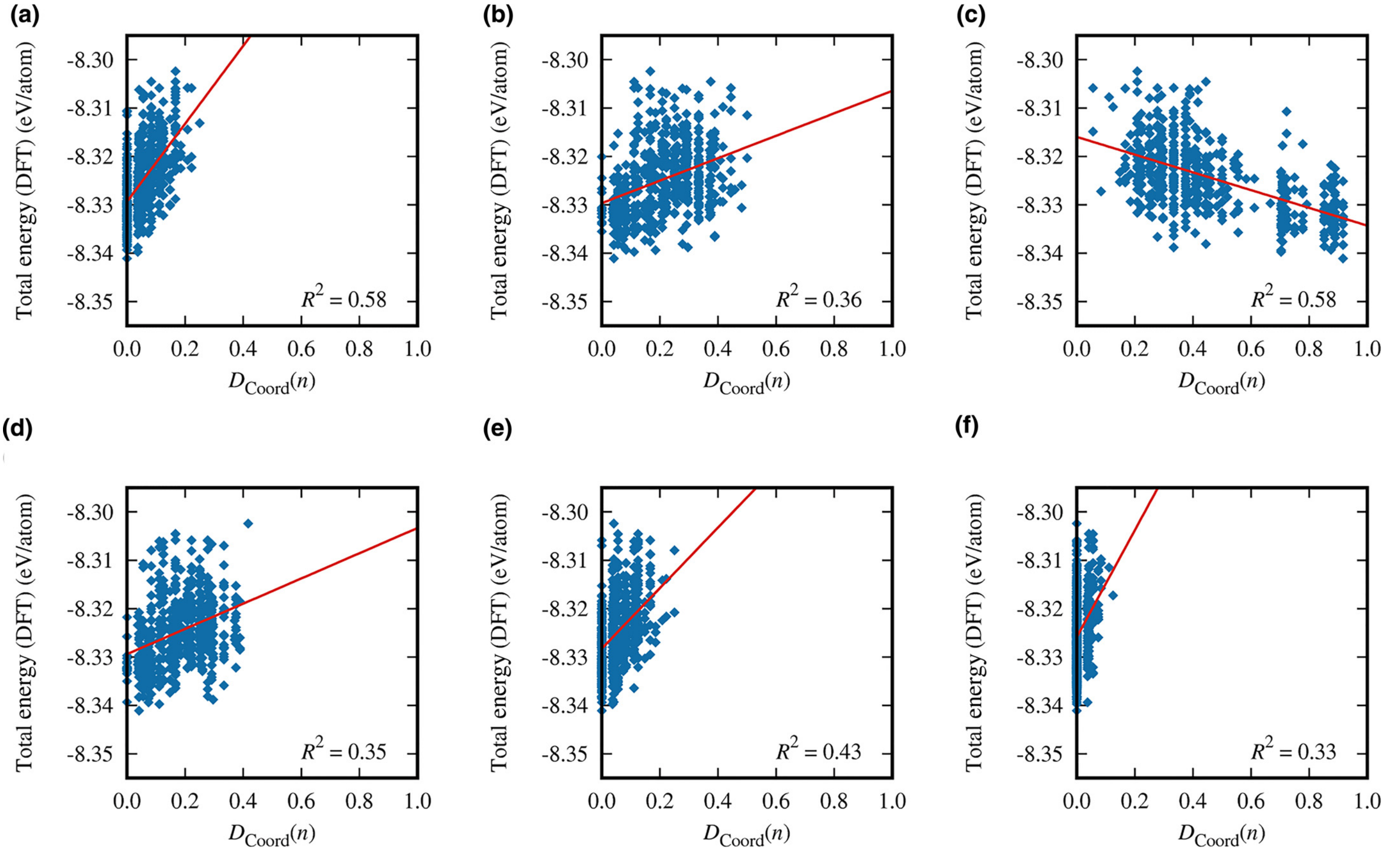


Fig. 2. Relationship between the $D_{\mathrm{Coord}}(n)$ and total energy values for $BaNbO_2N$ supercells as calculated using DFT. The values of n are (a) 0, (b) 1, (c) 2, (d) 3, (e) 4, and (f) 5.

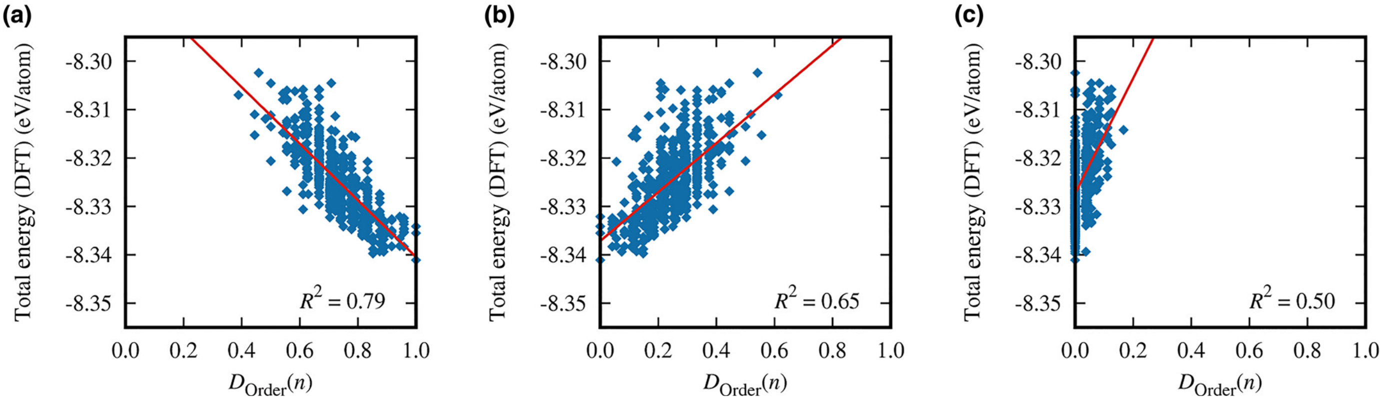


Fig. 3. Relationship between the $D_{\mathrm{Order}}(n)$ and total energy values for $BaNbO_2N$ supercells as calculated using DFT. The values of n are (a) 0, (b) 1, and (c) 2.

converge to a value of two during the search for stable anion orderings via the Metropolis method, regardless of the initial anion orderings. Because $D_{\mathrm{Coord}}(n)$ was correlated with the total energy, it is suitable for use as an explanatory variable.

Fig. 3 presents the relationship between the D_{Order} and total energy values for the same 420 $BaNbO_2N$ supercells as calculated by DFT. Fig. 3(a) demonstrates that the total energy decreases as the proportion of Nb atoms without trans-NbN chains (that is, the D_{Order} (0) value) becomes greater. In contrast, the total energy increases with increasing D_{Order} (1) and D_{Order} (2). These results indicate that cis-NbN chains are more stable than trans-NbN chains in the supercells. The structures of perovskite-type oxynitrides have been studied experimentally and theoretically [48–53]. It has been shown that $BaTaO_2N$ with a d^0-type electronic configuration had TaN chains in the cis-configuration [48,51–53]. It is expected that $BaNbO_2N$ and $BaTaO_2N$ exhibit similar properties because both Nb^{5+} and Ta^{5+} are group V elements and have the same valency. In this regard, our calculation result is considered to be consistent with the findings in these earlier studies. Because the $D_{\mathrm{Order}}(n)$ results obtained in the present work are in agreement with the earlier DFT calculations and also correlate with the total energy, and this parameter is also an appropriate explanatory variable.

3.2. Prediction of the total energy

Table 1 summarizes the RMSE values obtained by predicting the total energy using D_{Order} and D_{Coord}. Linear combinations of these

Table 1. Total energy RMSE values based on predictions using D_{Order} and D_{Coord}[a].

Data	Model			
	Linear regression	Lasso regression	Ridge regression	Random forest
Training (five-fold CV)	3.6	3.6	3.6	3.7
Training (non CV)	3.5	3.5	3.5	2.5
Test	3.8	3.8	3.8	3.7

[a] Reported in units of meV/atom.

explanatory variables were evidently sufficient because the RMSEs associated with the CV and test calculations were almost equal in the case of the linear, ridge and lasso regressions. As in the group contribution method, the total energy can be represented by the sum of the local structures. Among these methods, ridge regression was the most accurate and thus was employed in subsequent trials.

Fig. 4(a) summarizes the relationships between the total energy values for $BaNbO_2N$ supercells predicted using the ridge regression method in conjunction with D_{Order} and D_{Coord} and those calculated by DFT. The ridge regression approach was found to produce a consistent trend for the $BaNbO_2N$ supercells in the test set (that were not included in the training set) based on the DFT calculation outputs, without outliers. The total energy values were predicted accurately (with a high R^2 value of 0.84 for the test set) when using D_{Order} and D_{Coord} alone, indicating that the energy was highly dependent on the local anion ordering structure.

D_{exp2} was incorporated into the predictive model because it was thought that the total energy could be predicted more accurately by incorporating long-range interactions. Table 2 provides the RMSE values obtained by predicting the total energy using D_{Order}, D_{Coord} and D_{exp2}. In contrast to the results obtained when considering only the local anion ordering (Table 1), the RMSE values resulting from CV and test calculations were found to depend on the method. Specifically, the Random Forest model produced a RMSE for the training set that was an order of magnitude smaller than that for the test set, indicating overlearning. It is believed that the Random Forest approach overestimated D_{exp2} terms with small contributions. However, the accuracy of the total energy predictions obtained from the other regression methods was improved by including D_{exp2} in the model. Considering the RMSE values associated with the five-fold CV calculations, ridge regression was the most accurate regression method. This approach was presumably more accurate than lasso regression because many D_{exp2} terms made slight contributions to the total energy. However, it is difficult to identify which specific terms made significant contributions in terms of improving the prediction accuracy because of issues related to multiple collinearity.

Fig. 4(b) shows the relationships between the total energy values for $BaNbO_2N$ supercells predicted by ridge regression using D_{Order}, D_{Coord} and D_{exp2} and those obtained by DFT. Compared to the predictions based on D_{Order} and D_{Coord} (Fig. 4(a)), the accuracy was improved, suggesting that not only local steric structures but also long-range interactions affect the stability of the supercells. A one-to-one relationship between the atomic arrangement and atomic coordinates (including lattice vectors) of optimized structures is correct in this system because the R^2 value was as high as 0.91.

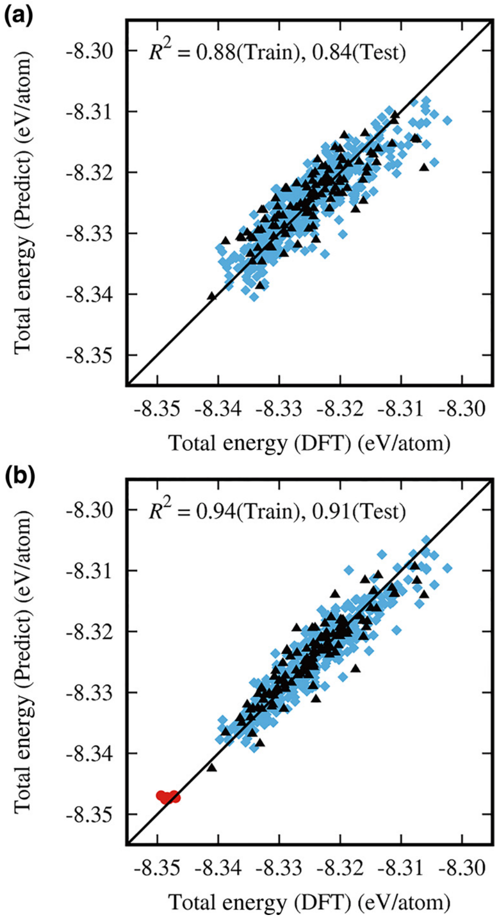


Fig. 4. Relationships between the total energy values for $BaNbO_2N$ supercells in the training set (rhombuses) and the test set (triangles) predicted by ridge regression and those calculated by DFT. The explanatory variables employed were (a) D_{Order} and D_{Coord}, and (b) D_{Order}, D_{Coord} and D_{exp2}. The total energy values for the eight $BaNbO_2N$ supercells obtained using the Metropolis method are shown as red circles in (b).

3.3. Searching for stable anion ordering

The ridge regression model predicted the total energy with the greatest accuracy using D_{Order}, D_{Coord} and D_{exp2} as the explanatory variables. Therefore, stable anion orders were predicted based on this model in conjunction with the Metropolis method. Fig. 5 provides the total energy for each step of the Metropolis method

Table 2. Total energy RMSE values based on predictions using D_{Order}, D_{Coord} and D_{exp2}[a].

Data	Method			
	Linear regression	Lasso regression	Ridge regression	Random forest
Training (five-fold CV)	3.2	3.5	2.7	3.6
Training (non CV)	2.0	3.3	2.5	0.38
Test	2.7	3.6	2.8	3.9

[a] Reported in units of meV/atom.

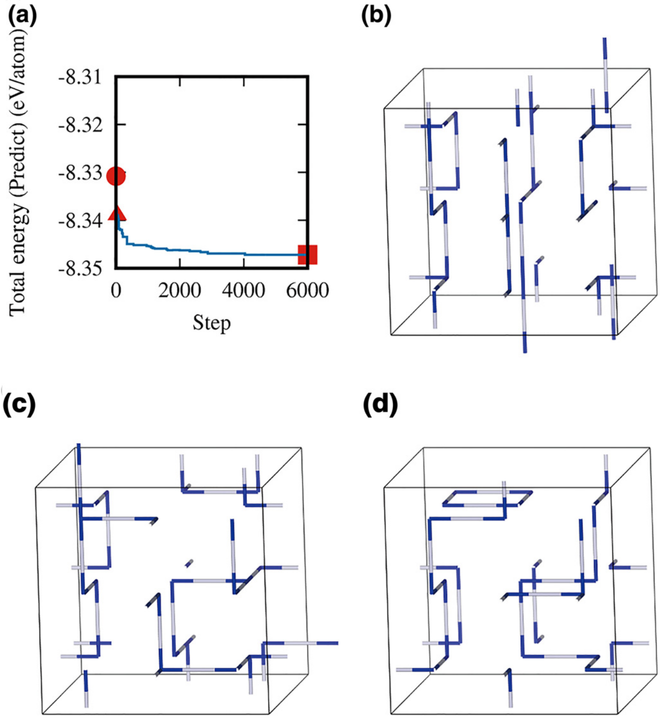


Fig. 5. (a) Total energy values for a $3 \times 3 \times 3$ $BaNbO_2N$ supercell at each step in the Metropolis method, and drawings of NbN chains in the supercell at the (b) initial, (c) 40th and (d) final steps of the Metropolis method.

applied to a randomly-generated $3 \times 3 \times 3$ $BaNbO_2N$ supercell. The total energy decreased rapidly in the initial stage of this procedure and converged sufficiently following 2000 steps. The predicted total energy values were lower than those for the initial structure by at least 0.01 eV/atom, which exceeds the RMSE of the predictive model. Therefore, the supercell structures obtained following 2000 steps can be regarded as stable. Fig. 5(b)–(d) provides images of NbN chains in a $3 \times 3 \times 3$ $BaNbO_2N$ supercell during the search for stable anion ordering by the Metropolis method. Following stabilization of the supercell, all the NbN chains had a cis conformation and N atoms were not localized but rather were dispersed three-dimensionally. The other seven randomly-generated $3 \times 3 \times 3$ supercells and those with different even/odd periodicities also exhibited the same features, characterized by delocalization of N atoms and the dominance of NbN chains in the cis conformation (Fig. S1 in the Supporting Information). Note that the supercells including short periodicities tended to be less stable, because the ability of the NbN chains to adopt a cis conformation was restricted (Fig. S2). Even so, these results are in agreement with earlier studies of ABO_2N-type material $CaTaO_2N$ based on DFT calculations, in that three-dimensional anion ordering with d^0-cation—N chains in the cis conformation was found to be the most thermodynamically stable structure [11]. These data demonstrate the feasibility of readily reproducing the results of DFT calculations by machine learning.

The total energy values for the eight $3 \times 3 \times 3$ $BaNbO_2N$ supercells obtained by the Metropolis method were also calculated using DFT and the results are superimposed in Fig. 4(b). The predicted total energies closely match the total energies calculated by DFT even though supercells with random anion orders were adopted as the initial structures. It should be noted that the supercell structures identified by the Metropolis method were not included in either the training or test sets because thermodynamically stable supercell structures were actively explored based on the predictive model. Moreover, the total energies for the eight $BaNbO_2N$ supercells were lower than those for the 560 $BaNbO_2N$ supercells initially generated as the training and test data sets. This result indicates the ability of the predictive model to explore stable anion ordering with accuracy comparable to that of DFT calculations. Successful extrapolation of the predictions suggests that overlearning was avoided and that the explanatory variables used were not only statistically but also chemically meaningful.

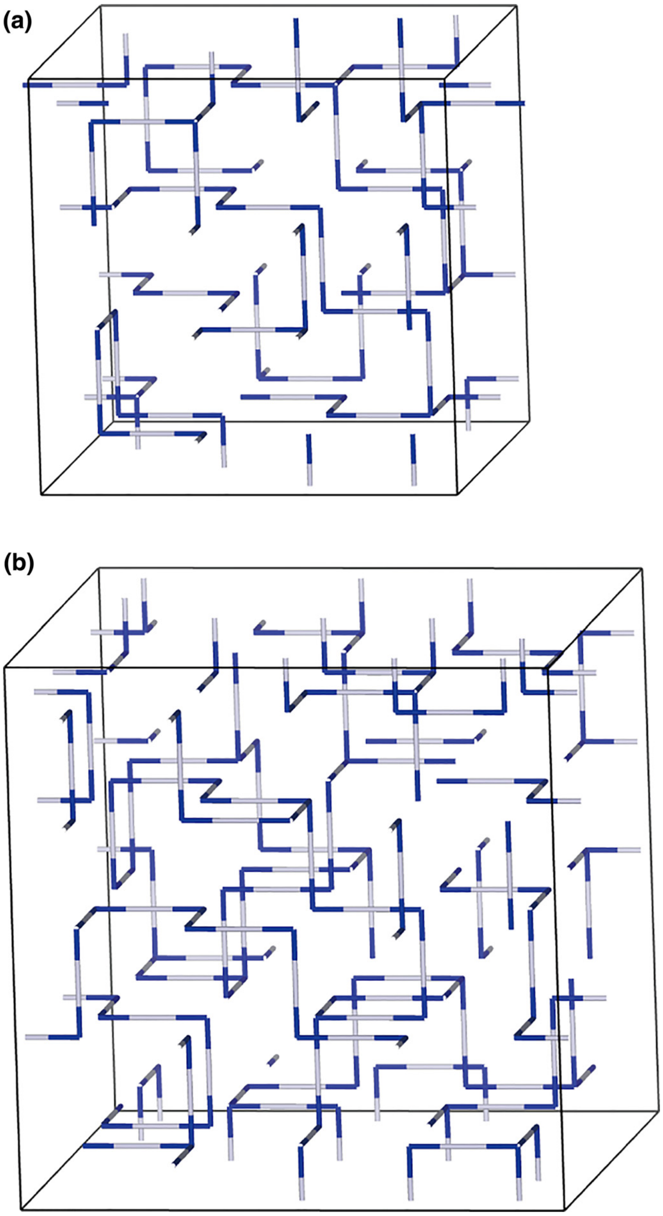
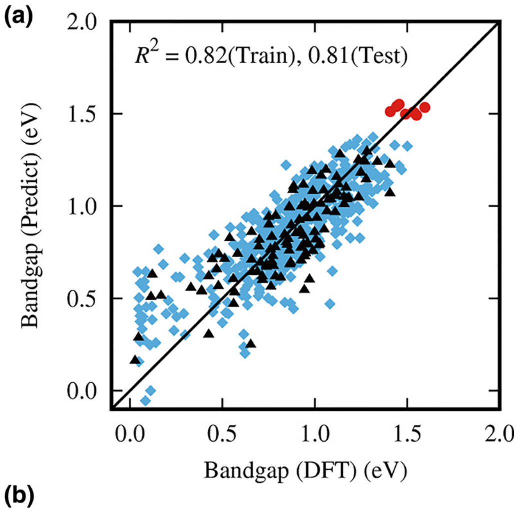


Fig. 6. Drawings of NbN chains in (a) $4 \times 4 \times 4$ and (b) $5 \times 5 \times 5$ supercells as predicted by the Metropolis method. The inverse temperature β and the number of steps were set to 50,000 eV^{-1} atom and 3000, respectively.

Fig. 6 shows images of NbN chains in $4 \times 4 \times 4$ and $5 \times 5 \times 5$ $BaNbO_2N$ supercells as generated by the Metropolis method based on the predictive model established using the $3 \times 3 \times 3$ supercells. The $4 \times 4 \times 4$ and $5 \times 5 \times 5$ supercells had NbN chains in the cis conformation exclusively, with N atoms evenly distributed throughout. These results are consistent with the data obtained for the $3 \times 3 \times 3$ supercells, and indicate the applicability of the predictive model to larger supercells. DFT calculations involving large supercells are typically computationally expensive, although it is possible to estimate the total energy by applying a predictive model based on DFT calculations with relatively small supercells.



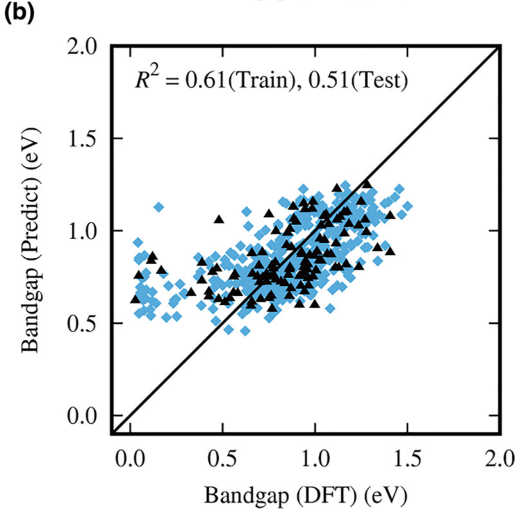


Fig. 7. Relationships between the bandgap energy values for $BaNbO_2N$ supercells in the training set (rhombuses) and the test set (triangles) as predicted using (a) D_{Order}, D_{Coord} and D_{exp2}, and (b) solely D_{Order} and D_{Coord} (employing ridge regression) and the values calculated by DFT. The bandgap energies for the eight $BaNbO_2N$ supercells obtained by the Metropolis method are shown in (a) as red circles.

It is anticipated that machine learning could therefore allow calculations involving low concentrations of dopants and defect generation, which require large supercells.

3.4. Prediction of the bandgap energy

Because stable anion ordering was successfully predicted, the bandgap energy of $BaNbO_2N$ supercells was also analyzed, since this is related to the electronic state of the material. Fig. S3 plots the relationship between the total energy and the bandgap obtained by DFT for the 420 $BaNbO_2N$ supercells in the training set. The bandgap became wider as the supercell became more stable, because the energy of the N 2*p* and O 2*p* orbitals constituting the valence band maximum was reduced. The data points also converged to a greater extent as the total energy of the supercell became lower. Therefore, the bandgap can be predicted from the total energy once stable supercell structures have been identified.

Fig. 7(a) shows the relationship between the bandgaps of $BaNbO_2N$ supercells as predicted by ridge regression using D_{Order}, D_{Coord} and D_{exp2} and the values calculated by DFT. The relationship based on predictions using solely D_{Order} and D_{Coord} is also presented for comparison in Fig. 7(b). The bandgap energy was evidently not accurately predicted using the local anion arrangements alone, as the R^2 values for the training and test sets were 0.58 and 0.49, respectively. This lack of correlation is attributed to continuous spreading of the wave function representing the electronic state in the crystal, such that local information regarding anion ordering alone cannot reproduce the bandgap correctly. In particular, supercell structures having extremely small bandgap energies based on DFT were not well reproduced, although it should be noted that such structures were unstable and thus unrealistic.

The bandgap prediction accuracy was remarkably improved when long-range interactions were considered by incorporating D_{exp2} as an explanatory variable. The average of the estimated bandgap energy for the supercells was 1.44 eV and therefore closer to the experimental value of 1.68 eV [21] than those of the initially generated 560 supercells. Notably, DFT calculations for some of the supercell structures identified by the Metropolis method produced even closer bandgap energies. Improvements in the explanatory variables should enable more accurate prediction of the bandgap energy based on machine learning in future.

4. Conclusions

Stable anion ordering in perovskite-type $BaNbO_2N$ supercells was predicted using a regression model established by machine learning based on DFT calculation outputs. In this process, the total energies for small $BaNbO_2N$ supercells (up to 27 unit cells) with random anion ordering were calculated by DFT. The explanatory variables D_{Order} and D_{Coord} (reflecting the local anion ordering) and D_{exp2} (associated with chemical bonds and long-range interactions) were found to be applicable to predictions of the total energy, and including D_{exp2} was shown to improve the prediction accuracy. The most accurate model, based on ridge regression, reproduced the local anion ordering generated by DFT calculations, in which the most stable supercells had each Nb atom coordinated with two N atoms, along with NbN chains in a cis conformation.

By combining this predictive model with the Metropolis Monte Carlo method, stable anion orders were rapidly obtained without the need to input the most stable supercell having optimized anion ordering or the requirement for costly structural optimization based on DFT calculations. Supercells predicted by the Metropolis method following the predictive model were more stable than any of the supercells used for machine learning and had a total energy close to that calculated by DFT. These results indicate the ability of the predictive model developed herein to explore stable anion ordering with a level of accuracy comparable to that obtained from costly DFT calculations. It was also possible to predict stable anion ordering in larger (e.g., $4 \times 4 \times 4$ and $5 \times 5 \times 5$) supercells by applying the same predictive model. The stable anion ordering generated in this manner can be used to accurately predict electronic properties, such as bandgap energy. This work suggests a means of predicting the properties of functional materials with complex compositions at reasonable computational costs via the appropriate choice of the explanatory variables and the use of machine learning. We expect that this method is applicable to other semiconductors through modifications of explanatory variables considering material systems and crystal structures.

Acknowledgments

This work was financially supported by Grants-in-Aid for Scientific Research (A) (no. 16H02417) and for Young Scientists (A) (no. 15H05494) from the Japan Society for the Promotion of Science (JSPS). The DFT calculations in this study were performed using the facilities at the Supercomputer Center in the Institute for Solid State Physics (ISSP) at the University of Tokyo and at the Research Center for Computational Science in the Institute for Molecular Science (IMS), Okazaki, Japan. This research was partly supported by MEXT as "Priority Issue on Post-K computer" (Development of new fundamental technologies for high-efficiency energy creation, conversion/storage and use).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jechem.2019.01.012.

References

[1] J. Zhu, H. Li, L. Zhong, P. Xiao, X. Xu, X. Yang, Z. Zhao, J. Li, ACS Catal. 4 (2014) 2917–2940.
[2] H. Kato, A. Kudo, J. Phys. Chem. B 105 (2001) 4285–4292.
[3] K. Maeda, J. Photochem. Photobiol. C 12 (2011) 237–268.
[4] A. Ishikawa, T. Takata, T. Matsumura, J.N. Kondo, M. Hara, H. Kobayashi, K. Domen, J. Phys. Chem. B 108 (2004) 2637–2642.
[5] A. Kojima, K. Teshima, Y. Shirai, T. Miyasaka, J. Am. Chem. Soc. 131 (2009) 6050–6051.
[6] J. Burschka, N. Pellet, S.-J. Moon, R. Humphry-Baker, P. Gao, M.K. Nazeeruddin, M. Grätzel, Nature 499 (2013) 316–319.
[7] Y.-I. Kim, P.M. Woodward, K.Z. Baba-Kishi, C.W. Tai, Chem. Mater. 16 (2004) 1267–1276.
[8] K. Maeda, K. Domen, J. Phys. Chem. C 111 (2007) 7851–7861.
[9] B. Siritanaratkul, K. Maeda, T. Hisatomi, K. Domen, ChemSusChem 4 (2001) 74–78.
[10] A. Ziani, C. Le Paven, L. Le Gendre, F. Marlec, R. Benzerga, F. Tessier, F. Cheviré, M.N. Hedhili, A.T. Garcia-Esparza, S. Melissen, P. Sautet, T. Le Bahers, K. Takanabe, Chem. Mater. 29 (2017) 3989–3998.
[11] A. Kubo, G. Giorgi, K. Yamashita, Chem. Mater. 29 (2017) 539–545.
[12] H. Wolff, R. Dronskowski, J. Comput. Chem. 29 (2008) 2260–2267.
[13] I.E. Castelli, D.D. Landis, K.S. Thygesen, S. Dahl, I. Chorkendorff, T.F. Jaramillo, K.W. Jacobsen, Energy Environ. Sci. 5 (2012) 9034–9043.
[14] A. Grimaud, K.J. May, C.E. Carlton, Y.-L. Lee, M. Risch, W.T. Hong, J. Zhou, Y. Shao-Horn, Nat. Commun. 4 (2013) 2439.
[15] B. Modak, S.K. Ghosh, J. Phys. Chem. C 119 (2015) 23503–23514.
[16] Y. Wu, P. Lazic, G. Hautier, K. Persson, G. Ceder, Energy Environ. Sci. 6 (2013) 157.
[17] Y. Liu, T. Zhao, W. Ju, S. Shi, J. Materiomics 3 (2017) 159–177.
[18] W. Li, R. Jacobs, D. Morgan, Comput. Mater. Sci. 150 (2018) 454–463.
[19] G. Pilania, A. Mannodi-Kanakkithodi, B.P. Uberuaga, R. Ramprasad, J.E. Gubernatis, T. Lookman, Sci. Rep. 6 (2016) 19375.
[20] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, J. Chem. Phys. 21 (1953) 1087–1092.
[21] T. Hisatomi, C. Katayama, Y. Moriya, T. Minegishi, M. Katayama, H. Nishiyama, T. Yamada, K. Domen, Energy Environ. Sci. 6 (2013) 3595–3599.
[22] T. Yamada, Y. Murata, S. Suzuki, H. Wagata, S. Oishi, K. Teshima, J. Phys. Chem. C 122 (2018) 8037–8044.
[23] J. Seo, T. Hisatomi, M. Nakabayashi, N. Shibata, T. Minegishi, M. Katayama, K. Domen, Adv. Energy Mater. 8 (2018) 1800094.
[24] J.M. Sanchez, F. Ducastell, D. Gratias, Physica 128 A (1984) 334.
[25] J.M. Sanchez, Phys. Rev. B 48 (1993) 14013.
[26] J.M. Sanchez, Phys. Rev. B 81 (2010) 224202.
[27] A.R. Natarajan, A. Van der Ven, npj Comput. Mater. 4 (2018) 56.
[28] A. Seko, I. Tanaka, J. Phys.: Condens. Matter 26 (2014) 115403.
[29] G. Kresse, J. Hafner, Phys. Rev. B 47 (1993) 558.
[30] G. Kresse, J. Hafner, Phys. Rev. B 49 (1994) 14251–14269.
[31] G. Kresse, J. Furthmüller, Comput. Mater. Sci. 6 (1996) 15–50.
[32] G. Kresse, J. Furthmüller, Phys. Rev. B 54 (1996) 11169–11186.
[33] T.E. Oliphant, A Guide to NumPy, Trelgol Publishing, USA, 2006.
[34] E. Jones, T. Oliphant, P. Peterson, et al., SciPy: Open Source Scientific Tools for Python. , 2001 http://www.scipy.org/ (accessed 19 November 2018).
[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[36] A.H. Larsen, J.J. Mortensen, J. Blomqvist, I.E. Castelli, R. Christensen, M. Dułak, J. Friis, M.N. Groves, B. Hammer, C. Hargus, E.D. Hermes, P.C. Jennings, P.B. Jensen, J. Kermode, J.R. Kitchin, E.L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J.B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K.S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, K.W. Jacobsen, J. Phys.: Condens. Matter 29 (2017) 273002.
[37] A.E. Hoerl, R.W. Kennard, Technometrics 12 (1970) 55–67.
[38] T. Robert, J. R. Stat. Soc. Ser. B. 58 (1996) 267–288.
[39] L. Breiman, Mach. Learn. 45 (2001) 5–32.
[40] K. Momma, F. Izumi, J. Appl. Crystallogr. 44 (2011) 1272–1276.
[41] M.T. Casais, J.A. Alonso, I. Rasines, M.A. Hidalgo, Mater. Res. Bull. 30 (1995) 201–208.
[42] J. Seo, Y. Moriya, M. Kodera, T. Hisatomi, T. Minegishi, M. Katayama, M. Katayama, K. Domen, Chem. Mater. 28 (2016) 6869–6876.
[43] P.E. Blochl, Phys. Rev. B 50 (1994) 17953–17979.
[44] G. Kresse, D. Joubert, Phys. Rev. B 59 (1999) 1758–1775.
[45] J.P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. 77 (1996) 3865–3868.
[46] J.P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. 78 (1997) 1396.
[47] M. Dupuis, J. Rys, H.F. King, J. Chem. Phys. 65 (1976) 111–116.
[48] H. Wolff, R. Dronskowski, J. Comput. Chem. 29 (2008) 2260–2267.
[49] M. Yang, J. Oró-Solé, J.A. Rodgers, A.B. Jorge, A. Fuertes, J.P. Attfield, Nat. Chem. 3 (2011) 47–52.
[50] A. Fuertes, J. Mater. Chem. 22 (2012) 3293–3299.
[51] K. Page, M.W. Stoltzfus, Y.-I. Kim, T. Proffen, P.M. Woodward, A.K. Cheetham, R. Seshadri, Chem. Mater. 19 (2007) 4037–4042.
[52] C.M. Fang, G.A. de Wijs, E. Orhan, G. de With, R.A. de Groot, H.T. Hintzen, R. Marchand, J. Phys. Chem. Solids 64 (2003) 281.
[53] Y. Hinuma, H. Moriwake, Y. Zhang, T. Motohashi, S. Kikkawa, I. Tanaka, J. Phys. Chem. Solids 24 (2012) 4343.