Machine-learning-based prediction of first-principles XANES spectra for amorphous materials

Haruki Hirai,[1] Takumi Iizawa,[1] Tomoyuki Tamura,[1,*] Masayuki Karasuyama,[2] Ryo Kobayashi,[1] and Takakazu Hirose[3]

[1] *Department of Physical Science and Engineering, Nagoya Institute of Technology, Nagoya 466-8555, Japan*
[2] *Department of Computer Science, Nagoya Institute of Technology, Nagoya 466-8555, Japan*
[3] *Research & Development Department, Shin-Etsu Chemical Co., Ltd., Annaka, Gunma 379-0125, Japan*

(Received 27 June 2022; revised 17 August 2022; accepted 19 October 2022; published 17 November 2022)

In this paper, a machine-learning-based method is proposed for predicting the x-ray absorption near-edge structure (XANES) for local configurations specific to amorphous materials. A combination of molecular dynamics and first-principles XANES simulations was adopted. The XANES spectrum was assumed to be accurately represented by linear regression of the local atomic descriptors. A comprehensive prediction of Si *K*-edge XANES spectra was performed based on an atom-centered symmetry function, smooth overlap of atomic positions, local many-body tensor representation, and spectral neighbor analysis potential. Furthermore, prediction accuracy was improved by compression of XANES spectral data and efficient sampling of training data.

DOI: 10.1103/PhysRevMaterials.6.115601

I. INTRODUCTION

In recent years, x-ray absorption spectroscopy (XAS) has become important for the structural characterization of various materials. In particular, near-edge region spectra, such as x-ray absorption near-edge structure (XANES), can provide sensitive information regarding chemical bonding, valence states, and coordination around atoms of interest. Conventional interpretation of experimental XANES spectra is most commonly based on the fingerprinting technique, wherein the experimental spectrum of interest is compared with that of a reference crystalline material. For energy loss near edge structure (ELNES) analysis performed with a transmission electron microscope (TEM), identical to XANES, data-driven spectral analysis by in-database machine learning for crystalline materials via simulation was proposed [1]. However, unlike crystalline materials, the interpretation of experimental XANES spectra for glass systems remains difficult. There is no guarantee that spectra of crystalline material will be applicable to local configurations of amorphous or glass systems, and reference spectra for local configurations specific to those systems are lacking. Therefore, theoretical simulations are indispensable for the complete interpretation of experimental XANES spectra of glass systems.

Li-ion secondary batteries exhibiting high-energy densities have been developed with a focus on applications, such as small mobile devices and electric motorization. A possible high-energy density incorporation includes high-capacity negative electrodes fabricated from tin, silicon, and other materials [2–7]. In particular, SiO materials incorporated into a cell using a negative electrode mixed with carbon-active materials have attracted significant commercial attention [8,9]. It is believed that SiO is not a simple mixture of Si and SiO_2, but is composed of local atomic structures that can change during charge-discharge cycling. Structural changes during charge-discharge cycling were probed using XANES for Si, O, and Li *K*-edges. It has been proposed that the SiO material before charging contains Si atoms with valence states other than Si^0 in crystalline Si and Si^{4+} in amorphous SiO_2 [10,11]. However, Si atoms with these valence states are uncommon in glass materials, and the reference spectra of crystalline systems are insufficient for in-depth structural analysis. For example, there are 331 Si-O binary crystalline data sets in the Materials Project [12], but there are only three, three, and seven systems including Si^{1+}, Si^{2+}, and Si^{3+}, respectively. Therefore, it is reasonable to conclude that the experimental SiO XANES spectra are not yet fully understood.

Theoretical XANES spectra can be obtained using various first-principles methods. A large system is required when dealing with disordered systems that include defects, interfaces, and amorphous structures. To obtain XANES spectra of these systems, first-principles density functional theory (DFT) [13,14] simulations based on the projector augmented-wave (PAW) method [15–17] is effective. First-principles XANES simulations of Ti doped in SiO_2 glass [18] and oxygen-related defects in SiO glass have already been applied with great success [19]. Since various atomic configurations are possible in glass materials, it is necessary to perform several simulations to account for statistical effect. Generally, first-principles XANES calculations are computationally expensive because the XANES spectrum contains information regarding the excitation of core electrons to unoccupied states and several unoccupied states should be included. Furthermore, sufficiently large supercells are necessary to prevent strong physical interactions among the excited electrons in repeated cells.

XANES spectra provide local information around an atom of interest, thus it can be assumed that XANES spectra could be reproduced using only the easily calculated local

[*] tamura.tomoyuki@nitech.ac.jp

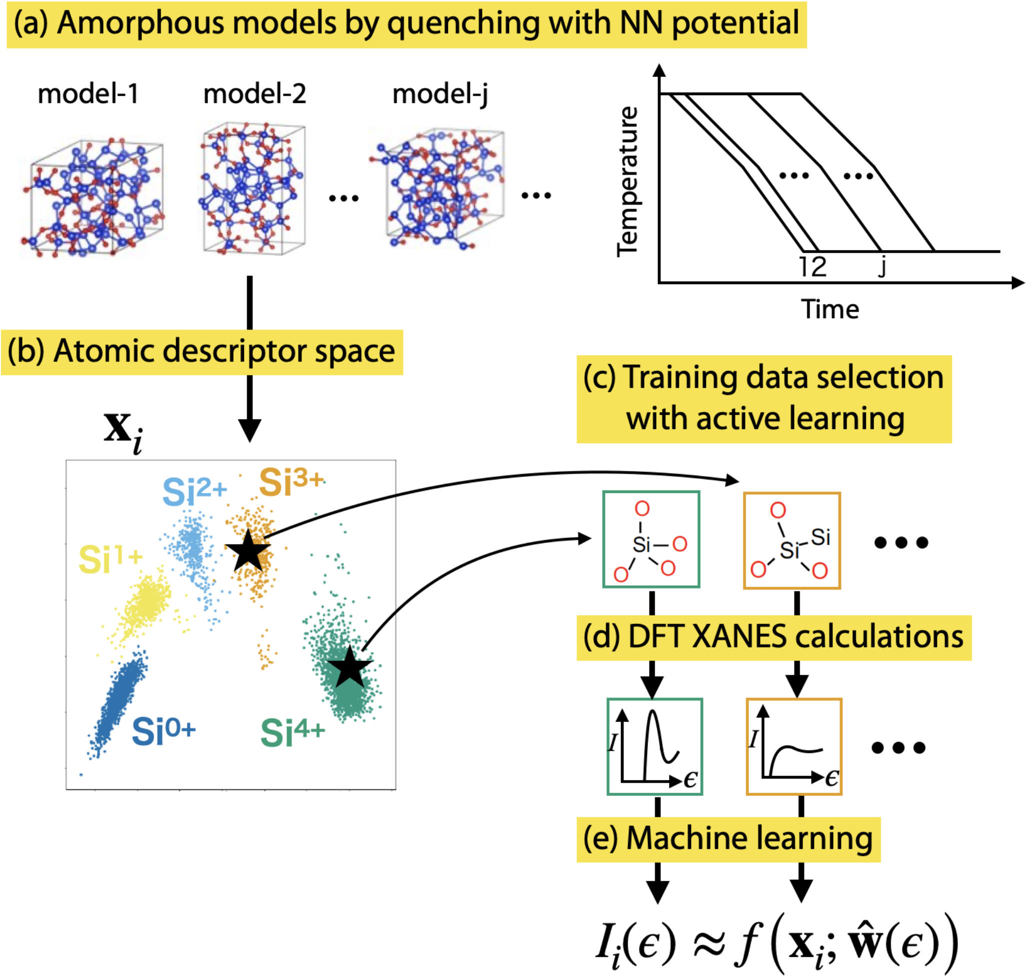
(a) Amorphous models by quenching with NN potential
model-1
model-2
model-j
Temperature
Time
1
2
j
(b) Atomic descriptor space
$\mathbf{x}_i$
Si1+
Si2+
Si3+
Si0+
Si4+
(c) Training data selection with active learning
(d) DFT XANES calculations
I
ϵ
(e) Machine learning
$I_i(\epsilon) \approx f(\mathbf{x}_i; \hat{\mathbf{w}}(\epsilon))$

FIG. 1. Schematic of nonsequential predictions of XANES spectra for amorphous materials. (a) Amorphous models are generated by quenching. (b) Atomic descriptors for all atoms are calculated. (c) Training data are selected with active learning. (d) XANES spectra is calculated within DFT. (e) Using the atomic descriptors and calculated XANES spectra, the machine learning model parameters are optimized.

atomic environment. Recently, a variety of atomic descriptors pertaining to the local environment have been proposed for machine-learned interatomic potentials, including symmetry functions [20], bispectrum components [21–24], and moment tensors [25]. In this paper, a procedure for predicting the XANES spectra of amorphous materials using the descriptors originally proposed for energies and forces was developed. Furthermore, the prediction accuracy was improved by efficient sampling of training data and compression of XANES data. Figure 1 shows a schematic of the proposed procedure, wherein a machine-learning model of the XANES spectra was developed based on atomic descriptor space.

II. METHODS

For *a*-SiO glass, periodic cells containing 50 Si and 50 O atoms were generated by quenching from the melt via classical and first-principles molecular dynamics (MD) simulations. Neural-network potentials expressing various Si valence states were constructed using the MD program package, nap [26] (see also Data-1 of the Supplemental Material [27] for more details), and 101 classical MD models were generated according to the following procedure with various cooling processes. The system was annealed at 5000 K in the NVT (constant-volume) ensemble for 200 ps and a structure was obtained every 2 ps, resulting in 101 initial structures. All structures were cooled linearly to 3000 K over 100 ps in the NVT ensemble, to 300 K over 100 ps in the NpT (constant-pressure) ensemble, and subsequently relaxed to stable configurations in the NpT ensemble. Additionally, a first-principles MD model was generated by rapid quenching from 3000 K with the cooling speed of 100 K/ps in the NVT ensemble. Figure 2 shows the distribution of the Si valence states in classical and first-principles MD models. The Si valence state and coordination number of oxygen were equivalent. The proportions of Si^{2+} and Si^{3+} in the first-principles MD model are more than those in classical MD models due to rapid cooling. The Si valence state was distributed between zero and four and no O–O bonds were observed.

Si *K*-edge XANES calculations were performed using the computational code QMAS (Quantum Materials Simulator) [28] to implement PAW calculations [15–17] with a generalized gradient approximation [29] for the exchange-correlation energy functional. The proper inclusion of a core hole in a supercell is essential for reproducing experimental XANES spectra [30,31]. The core-hole effect can be addressed through

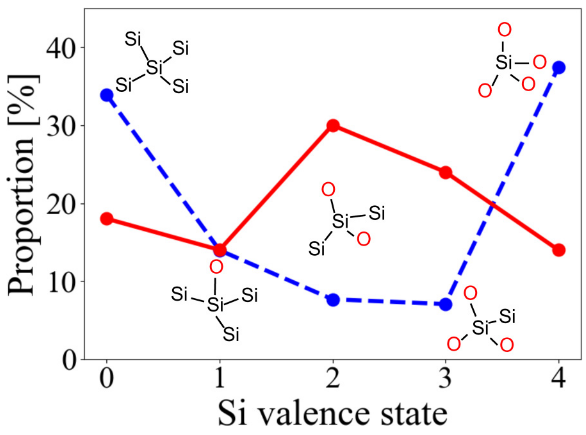

FIG. 2. Distribution of Si valence states in the models generated via classical MD (blue dotted line) and first-principles MD (red solid line).

the PAW pseudopotential of an excited atom with a core hole [18]. The **k**-point grids for the a-SiO models were set as the Γ point for self-consistent calculations and 4×4×4 for XANES calculations. The plane-wave energy cutoff was set to 476 eV (= 35 Ry). The theoretical spectra were broadened with Gaussian functions of $\sigma = 1.0$ eV. In our previous papers [18,32], it has already been confirmed that these parameters can be used to obtain sufficiently converged results (see also Data-3 of the Supplemental Material [27] for more details). The XANES spectra for diamond-type Si and α-quartz-type SiO_2 were also calculated as reference materials. Supercells were constructed with minimum distances between the excited atoms exceeding approximately 10 Å. The supercell contained 64 atoms for the diamond-type Si and 72 atoms for α-quartz-type SiO_2. The **k**-point sampling mesh is common to the a-SiO model.

The theoretical XANES spectra obtained by DFT-PAW calculations were predicted using only the local environment information. The proposed machine learning scheme was based on the assumption that a XANES spectrum can be represented by linear regression of the atomic structural descriptor. To obtain a highly accurate nonlinear regression, a large amount of training data must be prepared. Additionally, obtaining a single XANES spectrum is computationally expensive. Thus, linear ridge regression, which is a famous regression approach used in a wide range of statistical machine learning applications, was applied. Ridge regression minimizes the following objective function with input $\mathbf{x}_i$, output y_i, and regularization parameter λ:

$$L = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2, \tag{1}$$

where the ith row of $\mathbf{X}$ is $\mathbf{x}_i^\top$ and the ith element of $\mathbf{y}$ is y_i, for which the minimizer is written as

$$\hat{\mathbf{w}} = \mathbf{M}^{-1}\mathbf{X}^\top\mathbf{y}, \tag{2}$$

where $\mathbf{M} \equiv \mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ with the identity matrix $\mathbf{I} \in \mathbb{R}^{d\times d}$. It was assumed that the XANES spectral intensity at energy ϵ for ith atom can be represented as

$$I_i(\epsilon) = \mathbf{x}_i^\top\tilde{\mathbf{w}}(\epsilon), \tag{3}$$

where $I_i(\epsilon)$ denotes a 401-dimensional vector, wherein the region from 1835 to 1855 eV is divided by 0.05 eV increments, $\mathbf{x}_i \in \mathbb{R}^d$ denotes the d-dimensional structural descriptor vector, and $\tilde{\mathbf{w}}(\epsilon) \in \mathbb{R}^d$ denotes an unknown parameter vector for each energy level. Using the learned $\hat{\mathbf{w}}(\epsilon)$, a prediction of the XANES spectral intensity for the ith atom can be obtained as

$$I_i(\epsilon) \approx \mathbf{x}_i^\top\hat{\mathbf{w}}(\epsilon). \tag{4}$$

Parameter λ was determined by cross validation. We call this approach direct regression.

The direct regression of the intensity $I_i(\epsilon)$ is not always optimal since the intensity at each energy ϵ is regressed independently. XANES spectra are broadened with Gaussian functions. Therefore, we construct basis vectors by dimension reduction using the training set of XANES spectra. We adopt the principal component analysis (PCA) which is a typical method for dimension reduction. However, it is known that negative values appear in basis vectors with PCA. Thus, we also adopt the non-negative matrix factorization (NMF) [33] since the XANES spectrum should be a non-negative vector. The XANES spectrum can be approximated by a linear combination of m basis vectors, as

$$I_i(\epsilon) \approx \sum_{j=1}^{m} a_i^j\, G^j(\epsilon), \tag{5}$$

where $G^j(\epsilon)$ is a basis vector of the same energy mesh point as $I_i(\epsilon)$. a_i^j is a linear combination coefficient, and is predicted by ridge regression. Using the learned $\hat{\mathbf{w}}^j$, a_i^j can be predicted as

$$a_i^j \approx \mathbf{x}_i^\top\hat{\mathbf{w}}^j. \tag{6}$$

Finally, the intensity of XANES spectrum for the ith atom at energy ϵ can be predicted as

$$I_i(\epsilon) \approx \sum_{j=1}^{m} \mathbf{x}_i^\top\hat{\mathbf{w}}^j\, G^j(\epsilon). \tag{7}$$

For the atomic structural descriptor $\mathbf{x}_i$ of each atom, atom-centered symmetry function (ACSF) [20], smooth overlap of atomic position (SOAP) [21,22], local many-body tensor representation [34] (LMBTR), and spectral neighbor analysis potential (SNAP) [23,24] were applied. SOAP, ACSF, and LMBTR were obtained using the Python package DScribe [35]. Two types of SOAP were employed with spherical Gaussian-type orbitals (SOAP-gto) and polynomial basis (SOAP-poly) as radial basis functions (RBFs). The DScribe library was used to compute SOAP in user-friendly software packages that enable automatic analysis and mapping, and automatic selection and prediction tools for materials and molecules (ASAP) [36]. SNAP was obtained using LAMMPS [37] code. Two SNAP types were employed: a linear model in the original SNAP formulation [23] and a quadratic model (qSNAP) [24]. SNAP calculated by LAMMPS can correctly reproduce the physical properties of some metallic systems [38,39]. The details of the computational conditions for these descriptors are provided in Data-2 of the Supplemental Material [27]. The descriptors were standardized and subsequently dimensionally compressed using PCA, as described in the following section.

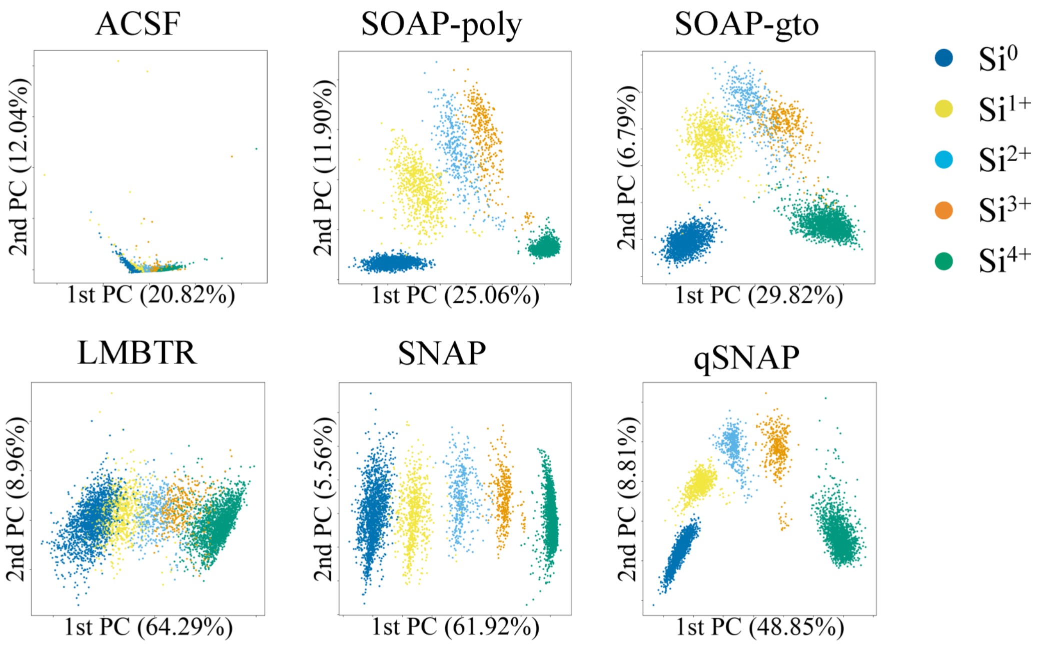

FIG. 3. PCA analysis of structural descriptors projected onto the plane of the first two principal components (PCs). The parentheses indicate the variances of the PCs.

III. RESULTS

The atomic structural descriptors **X** were calculated for all 5100 Si atoms contained in the generated models and standardized descriptors were applied to PCA maintaining 99.9% of the original variance. Figure 3 shows the projection of structural descriptors onto the first two principal components (PCs). Clusters formed based on the Si valence states and the first PC indicates the valence state. Local similarities were found in the amorphous structures. Only ACSF, originally proposed for neural-network potentials, showed different trends from the others, featuring outliers from the clusters.

Figure 4 shows the average Si *K*-edge XANES spectra calculated by randomly selecting 50 Si atoms from each of Si^0 to Si^{4+} along with the theoretical spectra of diamond-type Si and α-quartz-type SiO_2 and the experimental spectrum of *a*-SiO [10]. The calculated spectra were shifted to a lower energy by 10.2 eV to match the highest peaks in the theoretical spectrum of quartz SiO_2 and experimental spectrum of *a*-SiO. The highest peak intensity in the average spectrum of Si^{4+} was normalized to a value of one. All calculated spectra were shifted and normalized in the same manner. The average spectra of Si^0 and Si^{4+} in the amorphous models were broader than those in the bulk structures, but the peak positions were similar. The peak position shifted to the high-energy side with increasing valence state, consistent with that observed for other oxide materials.

The predictive performance of the structural descriptors was examined in detail. The descriptor dimensions after compression by PCA are listed in Table I. A total of 1000 training data points were randomly selected from the 5100 atoms. The test data consisting of 250 data points was generated by randomly selecting 50 data points from each Si valence state. The training and test data were common to all verifications. To

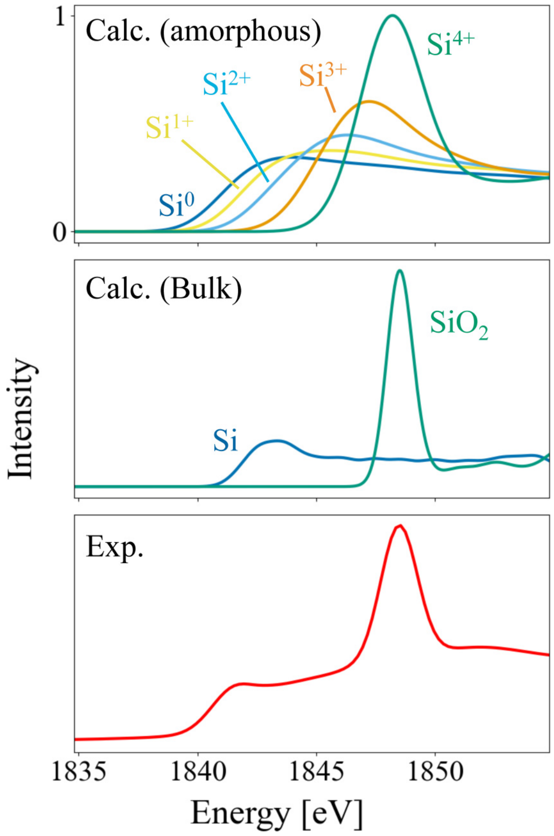

FIG. 4. Average Si *K*-edge XANES for Si valence states compared to theoretical calculations for bulk Si and quartz SiO_2 and experimental data for *a*-SiO [10]. The calculated absolute transition energy was adjusted to the experimental energy using a shift of −10.2 eV.

TABLE I. Descriptor dimensions before and after compression by PCA.

Descriptor	Before	After
ACSF	5202	85
SOAP-poly	2730	735
SOAP-gto	2730	531
LMBTR	1500	97
SNAP	1120	372
qSNAP	29 160	587

TABLE II. Total RMSE on direct and compressed regressions.

Descriptor	Direct	PCA	NMF
ACSF	3.214×10^{-2}	3.226×10^{-2}	3.210×10^{-2}
SOAP-poly	2.783	2.773	2.761
SOAP-gto	2.756	2.735	2.722
LMBTR	2.964	2.927	2.917
SNAP	2.450	2.446	2.424
qSNAP	2.321	2.278	2.259

evaluate prediction error, the root-mean-square error (RMSE) was determined for each energy mesh point as follows:

$$\mathrm{RMSE}(\epsilon) = \sqrt{\sum_{\mathbf{x}_i \in \mathcal{X}_{\mathrm{Test}}} \left\{ I_i^{\mathrm{DFT}}(\epsilon) - \mathbf{x}_i^{\top} \hat{\mathbf{w}}(\epsilon) \right\}^2 / |\mathcal{X}_{\mathrm{Test}}|}. \qquad (8)$$

The average RMSE was the smallest for qSNAP, as shown in Fig. 5(a). Figure 5(b) shows the RMSE for each valence-state of Si and qSNAP exhibited a smaller error for valence states between 1+ and 3+ than the other descriptors.

To improve prediction accuracy, the XANES spectra were predicted by regressing dimensionally reduced elements and then decoding them. PCA and NMF were applied to reduce the XANES dimensions. For PCA, the XANES spectrum was compressed until the cumulative contribution rate exceeded 99.9%, resulting in seven dimensions, whereas for NMF, the XANES spectrum was compressed to nine dimensions. The total RMSE values are listed in Table II. Compressed regression slightly improved the prediction accuracy, and NMF was more accurate than PCA. A comparison between the direct and compressed regressions for each energy level is shown in Fig. 6. The difference between the RMSEs of direct and compressed regressions is defined as

$$\Delta\mathrm{RMSE}(\epsilon) = \mathrm{RMSE}^{\mathrm{comp}}(\epsilon) - \mathrm{RMSE}^{\mathrm{direct}}(\epsilon), \qquad (9)$$

where the negative values indicate improvement. The compressed regression slightly improved prediction accuracy, but PCA regression was worse than direct regression at approximately 1840 eV. The reason behind this phenomenon is discussed in the following section.

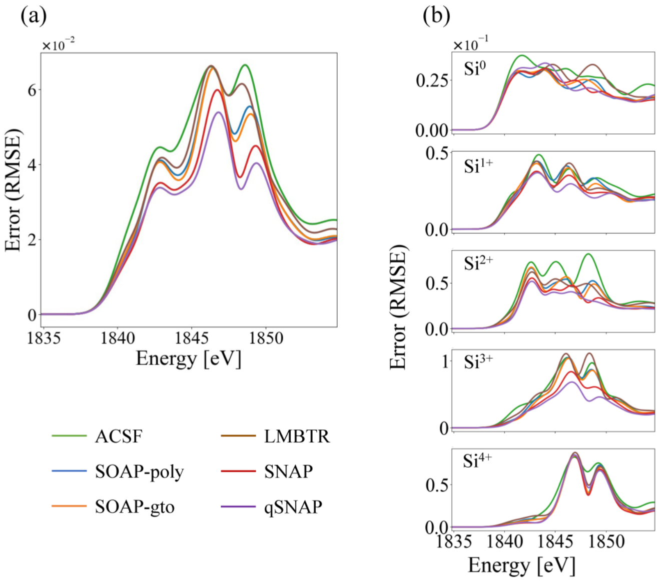

FIG. 5. (a) Average RMSE and (b) RMSE for each Si valence states obtained with direct regression.

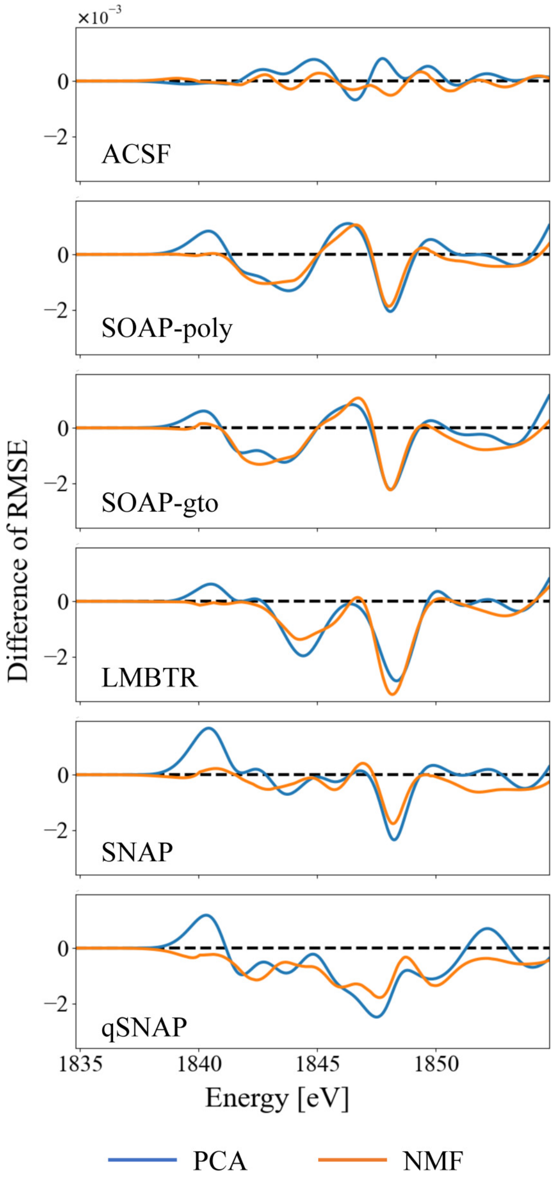

FIG. 6. Difference between the RMSEs of direct and compressed regressions. The difference of RMSE is defined in Eq. (9).

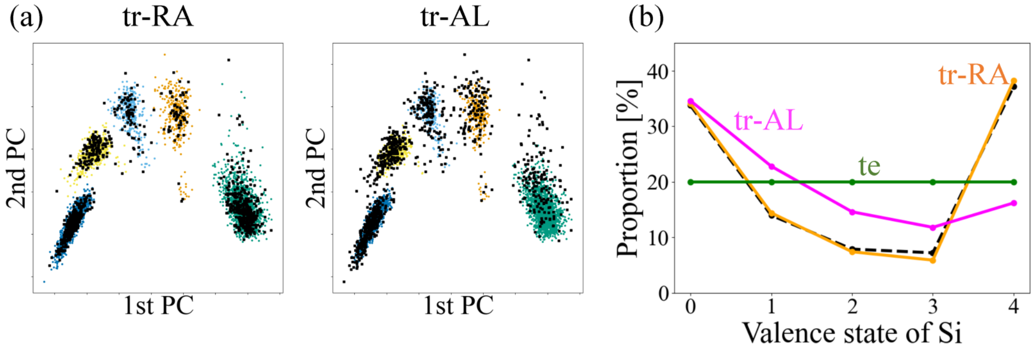

FIG. 7. (a) Training data obtained by active-learning (tr-AL) for qSNAP shown with black points in the reduced dimensional space created by PCA compared to training data obtained by random sampling (tr-RA). The distribution of all Si atoms is the same as in Fig. 3. (b) Distribution of test data (te) and training data by AL (tr-AL) and random sampling (tr-RA) for each valence state. The dotted black line represents the distribution of all Si atoms.

The active-learning (AL) approach has been widely studied in the machine-learning community for selecting appropriate training data sets. It has been proposed that the *uncertainty reduction* (UR) approach in AL is effective for collecting training data for DFT atomic energy [40]. Our UR criterion evaluates the reduction of uncertainty (variance) for all the atoms after adding a candidate training point $\mathbf{x}_i$. Candidate points are iteratively selected to minimize the predicted total uncertainty. For this approach, the observed $\mathbf{y}$ is not required because the variance depends on $\mathbf{x}$. For each descriptor, a training data set was generated using the AL approach. For example, in Fig. 7(a), training data selected by AL for qSNAP are shown in the reduced dimensional space created by PCA alongside training data selected by random sampling. Compared to the random sampling selection, the training data was more widespread and representative when selected with AL. In Fig. 7(b), the selection points for each valence state are compared. Random sampling reproduced the distribution of all data, while the number of selection points for Si^{4+} was reduced, and those for Si^{+1}, Si^{2+} and Si^{3+} increased in AL for all descriptors. From these results, it is expected that the prediction accuracy for Si^{4+} will decrease, while that for Si^{+1}, Si^{2+}, and Si^{3+} will improve.

The average RMSEs for the data sets obtained by the AL and random sampling approaches are compared in Fig. 8. The difference between the RMSEs of direct regression on the data sets obtained by the AL and random sampling approaches is defined as

$$\Delta\mathrm{RMSE}(\epsilon) = \mathrm{RMSE}^{\mathrm{AL}}(\epsilon) - \mathrm{RMSE}^{\mathrm{RA-direct}}(\epsilon), \qquad (10)$$

where AL contains direct and compressed regressions, as described later. Direct regression with AL yielded a smaller RMSE than that obtained with random sampling for all descriptors. Therefore, the regression accuracy was shown to be improved by using AL. For qSNAP, which was the most accurate, the accuracy for each valence state is compared in Fig. 9. With AL, the prediction accuracy for Si^{1+}, Si^{2+}, and Si^{3+} improved compared with that of random sampling, but the accuracy for Si^{4+} decreased. This is likely due to the reduction in the selection points in the training data, as shown in Fig. 7(b).

The transition of the total RMSE with respect to the amount of training data generated by AL and random sampling is

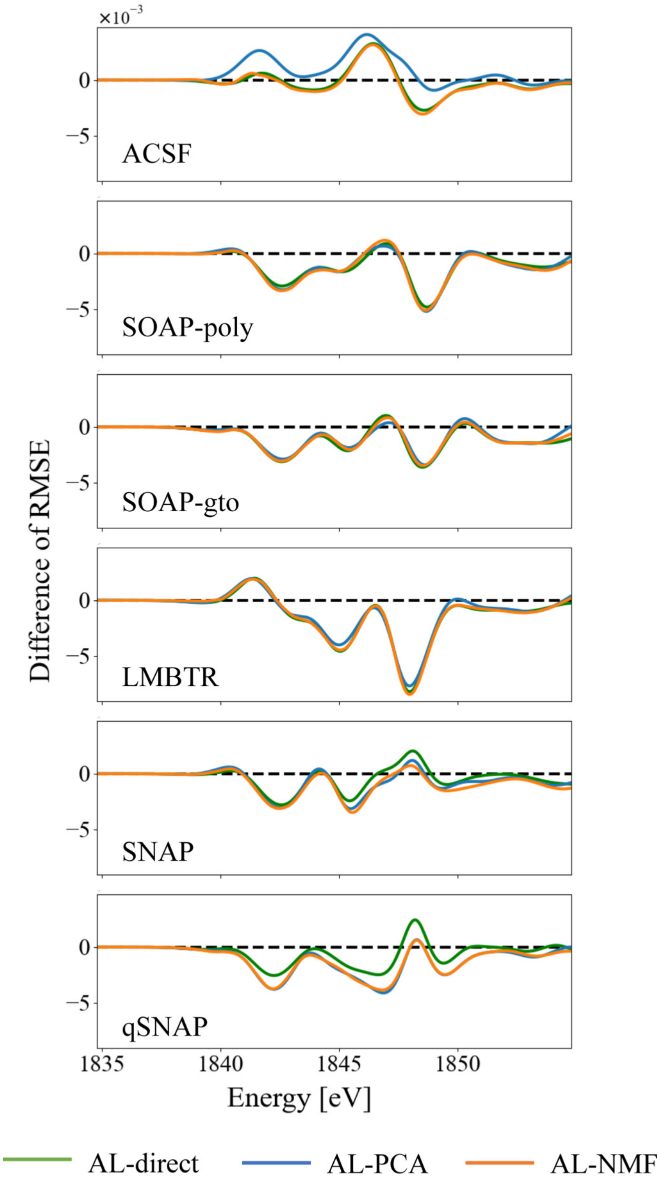

FIG. 8. Average differences in RMSEs for the data sets obtained by AL and random sampling. The difference of RMSE is defined in Eq. (10).

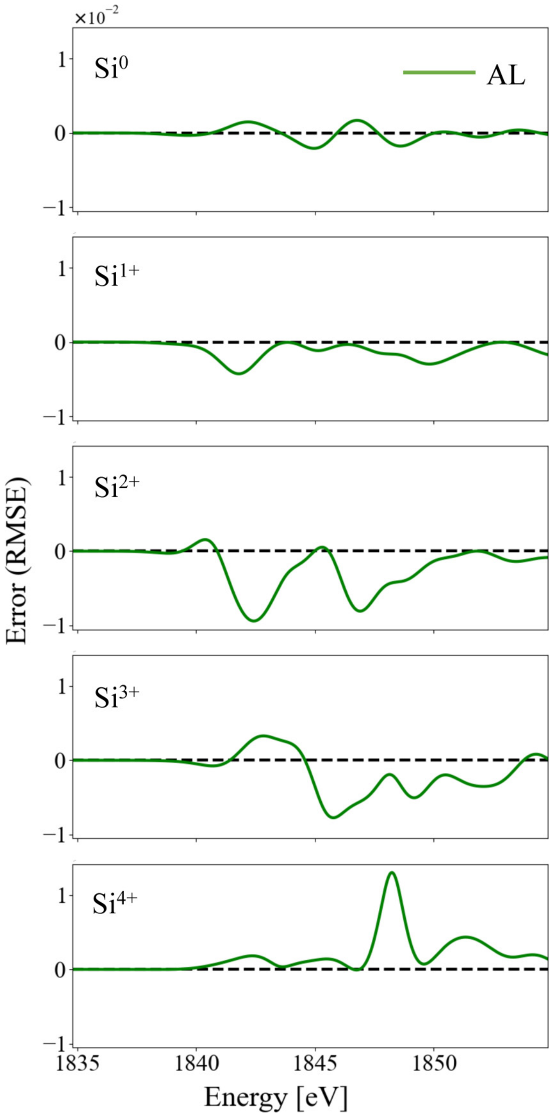

FIG. 9. Differences in RMSEs for Si valence states of the data sets obtained by AL and random sampling for qSNAP. The difference of RMSE is defined in Eq. (10).

shown in Fig. 10. Surprisingly, when the amount of training data was small, random sampling showed better prediction accuracy than AL. With increasing amounts of training data, AL improved the prediction accuracy. The accuracy of AL was poor for small training data because several specific data points exist in the training data and AL considers the variance, which gives priority to specific structures when few selection points are available.

AL was further investigated to determine whether it improves the accuracy of compression regression. The RMSEs for the data sets generated by AL and random sampling are shown in Fig. 8. Similar to direct regression, AL exhibited a smaller RMSE, showing that AL is effective for compression regression. Figure 11 shows the spectra with the smallest and largest RMSE in each valence state predicted for qSNAP with AL and NMF. For the smallest RMSE, the predicted spectra are in good agreement with the theoretical spectra. However, for the largest RMSE, the peak positions could not be predicted accurately.

We demonstrate the prediction of XANES spectrum for a large-scale model. An *a*-SiO model containing 4913 Si and 4913 O atoms was generated via classical MD simulation in the same way as for small 100-atoms models, as shown in Fig. 12(a). The proportions of Si^0, Si^{1+}, Si^{2+}, Si^{3+}, and Si^{4+} are 34.7, 12.6, 9.5, 3.5, and 39.6%, respectively. The proportion of Si^{3+} is much less than that in small 100-atom models shown in Fig. 2. Figure 12(b) shows PCA analysis of qSNAP for 5,100 Si atoms contained in 100-atoms models and 4913 Si atoms in the large-scale model. The distribution of the large-scale model is almost within the range of that of small-scale models, except for Si^0. This means that the parameter vector learned with small-scale models can be applied to large-scale models. Figure 12(c) shows the predicted average Si *K*-edge XANES for 4913 Si atoms in the large-scale model in comparison with the simple sum of bulk Si and SiO$_2$ shown in Fig. 4. In the energy range from 1843 to 1846 eV, the *a*-SiO model shows the intensity that is not seen in the simple sum of bulk Si and SiO$_2$, similar to experimental spectra. These results provide strong evidence for the presence of various Si valence other than Si^0 and Si^{4+} in SiO materials.

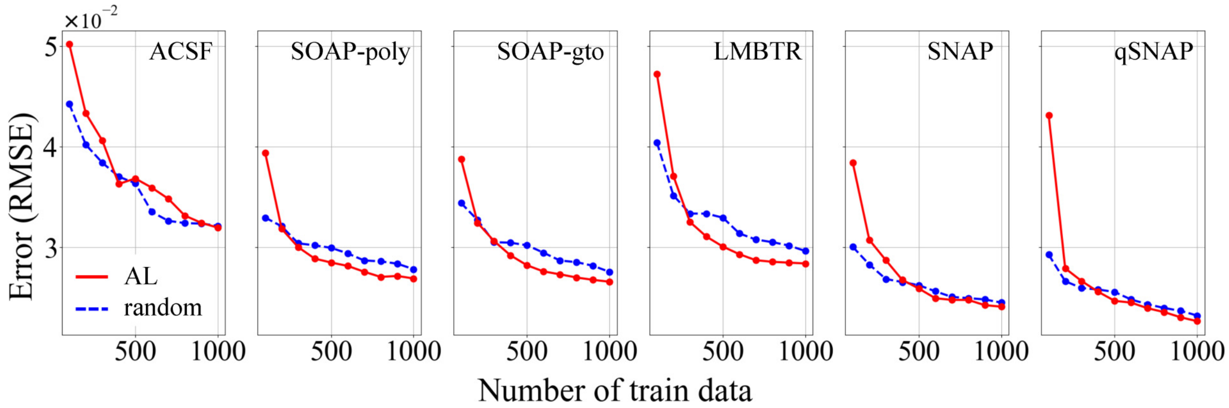

FIG. 10. The transition of total RMSE with respect to the number of training data points generated by AL (red solid line) and random sampling (blue dotted line).

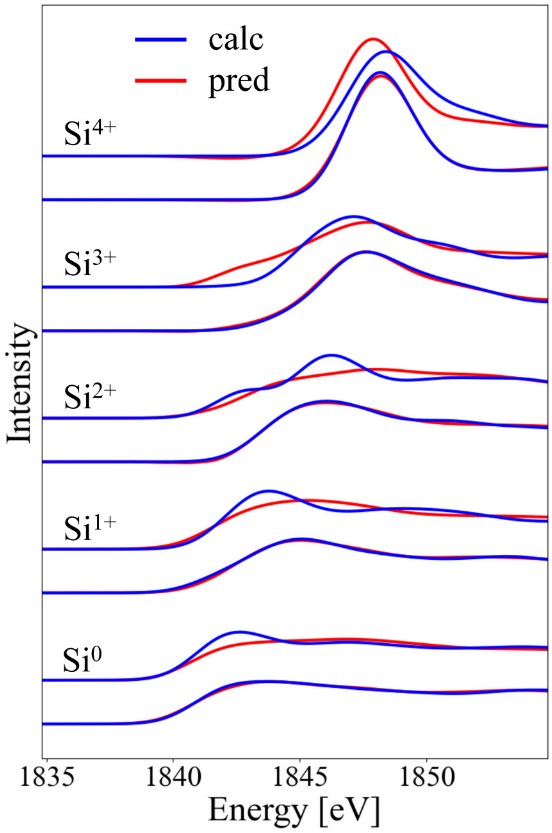

FIG. 11. The spectra with the smallest and largest RMSE in each valence state predicted for qSNAP with AL and NMF.

IV. DISCUSSIONS

The Si *K*-edge XANES contains information regarding the excitation of 1*s* electrons to unoccupied *p* states and reflects the projected density of states (DOSs) of *p* orbitals for the Si atom of interest. DOSs are the result of hybridization with surrounding atom electrons. Thus, the XANES spectra can be reproduced using the local atomic environment. Various atomic descriptors have been proposed to describe the local environment for machine-learned interatomic potentials. ACSF and LMBTR incorporate information regarding distances and angles. SOAP and SNAP are atomic descriptors of the correlation between atomic density functions, containing information relating to three- and four-body correlations, respectively. Previous studies have systematically examined the regression performance of the atomic descriptors. In the Gaussian process for the regression of energies and forces of the Si crystal, the prediction accuracy follows the order SOAP-poly, qSNAP, and SNAP [41]. For the kernel ridge regression of small organic molecules with ionic charges, the prediction accuracy follows the order SOAP-gto, SOAP-poly, and ACSF [35]. Pozdnyakov *et al.* showed that atomic descriptors calculated using three-body correlation were incomplete because of degeneracy [42]. Among the descriptors adopted in this paper, the atomic descriptors for three-body correlations were ACSF, SOAP, and LMBTR. SNAP contains information related to four-body correlations and was considered to be the most accurate in this paper. However, it has also been highlighted that SNAP is not complete because it does not distinguish chirality (mirror images) as the tetrahedra are not chiral [42]. Parsaeifard and Goedecker proposed an overlap matrix (OM) containing four-body correlation information [43]. Using the OM, the helix angle, which is specific to a four-body correlation, can be accurately regressed. XANES reflects additional information regarding the surrounding atoms, and it is highly possible that XANES is a vector quantity containing four-body correlation information. To further improve prediction accuracy, it is necessary to develop a structural descriptor that includes information relating to the four-body and higher-body correlations. For example, more recently, the atomic cluster expansion (ACE) [44,45] was developed to provide a complete and efficient representation of atomic properties as a function of local atomic environment in terms of many-body correlation functions, and SOAP and SNAP are equivalent to the three- and four-body terms in ACE, respectively. Although simple linear regression was adopted in this paper, results show that nonlinear regression, including the kernel regression, slightly improves prediction accuracy. The descriptors can be intrinsically connected to the target function via a simple nonlinear function. Our future work will focus on the verification of prediction performance by a variety of descriptors and models.

The structural descriptors mentioned above were developed for interatomic potentials that represent energies and forces. Recently, vector data prediction using these structural descriptors has been reported for atom-projected DOS prediction using SOAP [46] and graph neural networks [47]. The

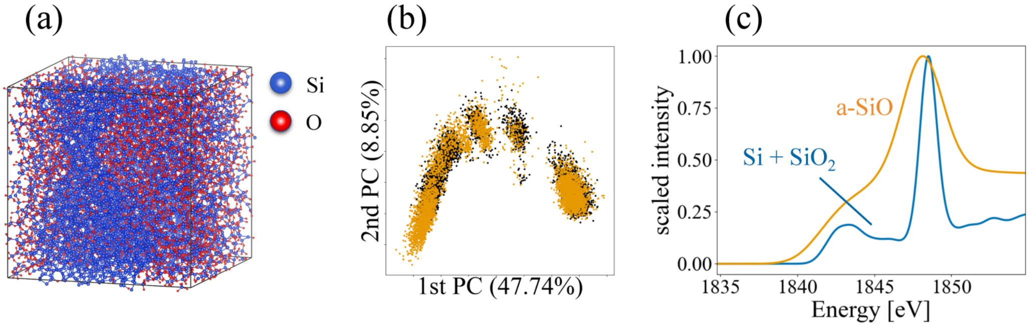

FIG. 12. (a) Atomic configuration for a large-scale *a*-SiO glass model containing 4913 Si atoms and 4913 O atoms. (b) PCA analysis of qSNAP for 5100 Si atoms contained in 100-atom models (black) and 4913 Si atoms in the 9826-atom model (orange). (c) The predicted average Si *K*-edge XANES for all Si atoms in the large-scale model in comparison with the simple sum of bulk Si and SiO_2 shown in Fig. 4.

simplest regression of vector data is a direct regression at each point. However, since there is a correlation with adjacent points in XANES spectral data, compressed regression is more useful. PCA is widely used to reduce dimensionality by compression. It was previously demonstrated that compressed regression by PCA is more useful than direct regression, and the cumulative distribution function (CDF) is more useful than PCA for predicting DOS based on Gaussian process regression [46]. In this paper, as shown in Table II, PCA improved the accuracy as a whole. However, as illustrated in Fig. 6, poor accuracy was obtained for some energy points. In principle, eigenvectors with negative values can appear in the PCA and the analysis confirmed a negative eigenvector. By contrast, eigenvectors with negative values did not appear in NMF. Shiga *et al.* proposed a spectral imaging technique for electron energy loss and energy-dispersive x-ray spectral data sets observed with scanning TEM combined with NMF [48]. As shown in Fig. 6, NMF compression improved the accuracy at many energy points, displaying higher regression accuracy than direct regression and PCA. In this paper, based on linear ridge regression, CDF did not show any advantages over direct regression.

Since the efficient sampling of training data directly affects regression accuracy, AL has been extensively studied in the field of machine learning. The uncertainty reduction (UR) technique was adopted in this paper, similar to a previous paper [40]. An important property of UR is that it does not require the observed $\mathbf{y}$ because the variance does not depend on $\mathbf{y}$. Thus, sequential DFT-XANES calculations are not necessary and it is possible to execute several calculations simultaneously using a massive parallel computer. However, as illustrated in Fig. 10, when the training data is limited, random sampling is more accurate than AL. This is because the AL tends to select from specific points to reduce the overall variance. When the number of training data points was $\geqslant 400$, AL was more accurate than random sampling for all descriptors. Therefore, it can be concluded that the UR technique is an efficient sampling method for training data when DFT calculations can be performed in parallel.

V. CONCLUSIONS

In this paper, a machine learning-based method was proposed for predicting XANES spectra for local configurations specific to amorphous materials using a combination of MD simulations and first-principles XANES simulations. It was assumed that the XANES spectrum can be represented by linear regression of the local atomic descriptors. Comprehensive predictions of the Si K-edge XANES spectra were made based on ACSF, SOAP, LMBTR, and SNAP methodologies. Furthermore, the prediction accuracy was improved by compression of the XANES spectral data and efficient sampling of training data.

ACKNOWLEDGMENTS

This work was supported by a MEXT KAKENHI grant awarded to T.T. (Grant No. 22H01808) and M.K. (Grant No. 21H03498). The authors would like to thank editage [55] for the English language review.

H.H.: Data curation, formal analysis, visualization, writing—original draft. T.I.: Data curation and formal analysis. T.T.: Conceptualization, methodology, writing—review and editing. M.K.: Conceptualization, methodology, writing—review and editing. R.K.: Data curation, methodology. T.H.: Investigation and validation.

The authors declare no competing interests.

[1] T. Mizoguchi and S. Kiyohara, Machine learning approaches for ELNES/XANES, Microscopy **69**, 92 (2020).

[2] Y. Idota, T. Kubota, A. Matsufuji, Y. Maekawa, and T. Miyasaka, Tin-based amorphous oxide: A high-capacity lithium-ion-storage material, Science **276**, 1395 (1997).

[3] A. D. W. Todd, R. E. Mar, and J. R. Dahn, Tin–transition metal–carbon systems for lithium-ion battery negative electrodes, J. Electrochem. Soc. **154**, A597 (2007).

[4] A. M. Wilson and J. R. Dahn, Lithium insertion in carbons containing nanodispersed silicon, J. Electrochem. Soc. **142**, 326 (1995).

[5] J. Yin, M. Wada, K. Yamamoto, Y. Kitano, S. Tanase, and T. Sakai, Micrometer-scale amorphous Si thin-film electrodes fabricated by electron-beam deposition for Li-ion batteries, J. Electrochem. Soc. **153**, A472 (2006).

[6] G. Wang, L. Sun, D. Bradhurst, S. Zhong, S. Dou, and H. Liu, Innovative nanosize lithium storage alloys with silica as active centre, J. Power Sources **88**, 278 (2000).

[7] C. Chan, H. Peng, G. Liu, K. McIlwrath, X. Zhang, R. Huggins, and Y. Cui, High-performance lithium battery anodes using silicon nanowires, Nat. Nanotechnol. **3**, 31 (2008).

[8] A. Hohl, T. Wieder, P. van Aken, T. Weirich, G. Denninger, M. Vidal, S. Oswald, C. Deneke, J. Mayer, and H. Fuess, An interface clusters mixture model for the structure of amorphous silicon monoxide (SiO), J. Non-Cryst. Solids **320**, 255 (2003).

[9] Y. Nagao, H. Sakaguchi, H. Honda, T. Fukunaga, and T. Esaka, Structural analysis of pure and electrochemically lithiated SiO using neutron elastic scattering, J. Electrochem. Soc. **151**, A1572 (2004).

[10] T. Hirose, M. Morishita, H. Yoshitake, and T. Sakai, Investigation of carbon-coated SiO phase changes during charge/discharge by x-ray absorption fine structure, Solid State Ionics **304**, 1 (2017).

[11] T. Hirose, M. Morishita, H. Yoshitake, and T. Sakai, Investigation of carbon-coated silicon oxide phase changes during charge/discharge by oxygen and lithium K-edge x-ray absorption fine structure spectroscopy, Solid State Commun. **269**, 39 (2018).

[12] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, The Materials Project: A materials genome approach to accelerating materials innovation, APL Mater. **1**, 011002 (2013).

[13] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, Phys. Rev. **136**, B864 (1964).

[14] W. Kohn and L. J. Sham, Self-consistent equations including exchange and correlation effects, Phys. Rev. **140**, A1133 (1965).

[15] P. E. Blöchl, Projector augmented-wave method, Phys. Rev. B **50**, 17953 (1994).

[16] N. A. W. Holzwarth, G. E. Matthews, R. B. Dunning, A. R. Tackett, and Y. Zeng, Comparison of the projector augmented-wave, pseudopotential, and linearized augmented-plane-wave formalisms for density-functional calculations of solids, Phys. Rev. B **55**, 2005 (1997).

[17] G. Kresse and D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B **59**, 1758 (1999).

[18] T. Tamura, S. Tanaka, and M. Kohyama, Full-paw calculations of XANES/ELNES spectra of Ti-bearing oxide crystals and TiO-SiO glasses: Relation between pre-edge peaks and Ti coordination, Phys. Rev. B **85**, 205210 (2012).

[19] W. Katayama, T. Tamura, Y. Nishino, and T. Hirose, First-principles XANES simulation for oxygen-related defects in Si-O amorphous materials, Comput. Mater. Sci. **196**, 110555 (2021).

[20] J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, J. Chem. Phys. **134**, 074106 (2011).

[21] A. P. Bartók and G. Csányi, Gaussian approximation potentials: A brief tutorial introduction, Int. J. Quantum Chem. **115**, 1051 (2015).

[22] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, Comparing molecules and solids across structural and alchemical space, Phys. Chem. Chem. Phys. **18**, 13754 (2016).

[23] A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials, J. Comput. Phys. **285**, 316 (2015).

[24] M. A. Wood and A. P. Thompson, Extending the accuracy of the snap interatomic potential form, J. Chem. Phys. **148**, 241721 (2018).

[25] A. V. Shapeev, Moment tensor potentials, Multiscale Model. Simul. **14**, 1153 (2016).

[26] R. Kobayashi, nap: A molecular dynamics package with parameter-optimization programs for classical and machine-learning potentials, J. Open Source Softw. **6**, 2768 (2021).

[27] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevMaterials.6.115601 for more details, which includes Refs. [49–54].

[28] S. Ishibashi, T. Tamura, S. Tanaka, M. Kohyama, and K. Terakura, Ab initio calculations of electric-field-induced stress profiles for diamond/$c-$BN (110) superlattices, Phys. Rev. B **76**, 153310 (2007).

[29] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized Gradient Approximation Made Simple, Phys. Rev. Lett. **77**, 3865 (1996).

[30] S.-D. Mo and W. Y. Ching, Ab initio calculation of the core-hole effect in the electron energy-loss near-edge structure, Phys. Rev. B **62**, 7901 (2000).

[31] S.-D. Mo and W. Y. Ching, X-ray absorption near-edge structure in alpha-quartz and stishovite: Ab initio calculation with core–hole interaction, Appl. Phys. Lett. **78**, 3809 (2001).

[32] T. Tamura, S. Ishibashi, S. Tanaka, M. Kohyama, and M.-H. Lee, First-principles analysis of the optical properties of structural disorder in SiO_2 glass, Phys. Rev. B **77**, 085207 (2008).

[33] D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature (London) **401**, 788 (1999).

[34] H. Huo and M. Rupp, Unified representation of molecules and crystals for machine learning, arXiv:1704.06439.

[35] L. Himanen, M. O. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, Dscribe: Library of descriptors for machine learning in materials science, Comput. Phys. Commun. **247**, 106949 (2020).

[36] B. Cheng, R.-R. Griffiths, S. Wengert, C. Kunkel, T. Stenczel, B. Zhu, V. L. Deringer, N. Bernstein, J. T. Margraf, K. Reuter, *et al.*, Mapping materials and molecules, Acc. Chem. Res. **53**, 1981 (2020).

[37] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton, LAMMPS: A flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales, Comput. Phys. Commun. **271**, 108171 (2022).

[38] X.-G. Li, C. Hu, C. Chen, Z. Deng, J. Luo, and S. P. Ong, Quantum-accurate spectral neighbor analysis potential models for Ni-Mo binary alloys and FCC metals, Phys. Rev. B **98**, 094104 (2018).

[39] M. A. Wood and A. P. Thompson, Quantum-accurate molecular dynamics potential for tungsten, arXiv:1702.07042.

[40] T. Tamura and M. Karasuyama, Prediction of formation energies of large-scale disordered systems via active-learning-based executions of ab initio local-energy calculations: A case study on a Fe random grain boundary model with millions of atoms, Phys. Rev. Mater. **4**, 113602 (2020).

[41] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood *et al.*, Performance and Cost Assessment of Machine Learning Interatomic Potentials, J. Phys. Chem. A **124**, 731 (2020).

[42] S. N. Pozdnyakov, M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, Incompleteness of Atomic Structure Representations, Phys. Rev. Lett. **125**, 166001 (2020).

[43] B. Parsaeifard and S. Goedecker, Manifolds of quasi-constant SOAP and ACSF fingerprints and the resulting failure to machine learn four-body interactions, J. Chem. Phys. **156**, 034302 (2022).

[44] R. Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, Phys. Rev. B **99**, 014104 (2019).

[45] G. Dusson, M. Bachmayr, G. Csányi, R. Drautz, S. Etter, C. van der Oord, and C. Ortner, Atomic cluster expansion: Completeness, efficiency and stability, J. Comput. Phys. **454**, 110946 (2022).

[46] C. Ben Mahmoud, A. Anelli, G. Csányi, and M. Ceriotti, Learning the electronic density of states in condensed matter, Phys. Rev. B **102**, 235130 (2020).

[47] V. Fung, P. Ganesh, and B. G. Sumpter, Physically informed machine learning prediction of electronic density of states, Chem. Mater. **34**, 4848 (2022).

[48] M. Shiga, K. Tatsumi, S. Muto, K. Tsuda, Y. Yamamoto, T. Mori, and T. Tanji, Sparse modeling of EELS and EDX spectral imaging data by nonnegative matrix factorization, Ultramicroscopy **170**, 43 (2016).

[49] G. Kresse and J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B **54**, 11169 (1996).

[50] J. Behler and M. Parrinello, Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces, Phys. Rev. Lett. **98**, 146401 (2007).

[51] Y. Huang, J. Kang, W. A. Goddard, and L.-W. Wang, Density functional theory based neural network force fields from energy decompositions, Phys. Rev. B **99**, 064103 (2019).

[52] H. Hay, G. Ferlat, M. Casula, A. P. Seitsonen, and F. Mauri, Dispersion effects in SiO_2 polymorphs: An ab initio study, Phys. Rev. B **92**, 144111 (2015).

[53] A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, Machine Learning a General-Purpose Interatomic Potential for Silicon, Phys. Rev. X **8**, 041048 (2018).

[54] L. C. Erhard, J. Rohrer, K. Albe, and V. L. Deringer, A machine-learned interatomic potential for silica and its relation to empirical models, npj Comput. Mater. **8**, 90 (2022).

[55] www.editage.jp.