PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2026, **28**, 3682

Machine learning-based XANES analysis for predicting the local structure and valence in amorphous silicon suboxides

Yu Fujikata, *[ab] Hiroki Sugisawa [b] and Teruyasu Mizoguchi [a]

Silicon suboxide (SiO_x, $0 < x < 2$) has attracted considerable interest across various industrial fields due to its tunable physical properties that are afforded by its compositional ratio. However, the quantitative resolution of its atomistic structure–property correlations remains challenging using conventional approaches. In this study, nine compositionally controlled amorphous SiO_x networks were generated *via* molecular dynamics simulations, and a comprehensive dataset of Si K-edge X-ray absorption near-edge structure (XANES) spectra was constructed using first-principles calculations. Subsequently, a deep neural network was trained to develop a model capable of directly predicting both the local silicon atom valence state and the Si–O radial distribution function from single-site XANES spectra. Systematic sub-window analysis revealed that features near the absorption edge and the main peak provided information related to the valence state, whereas precise structural predictions required information from higher-energy regions. The model trained solely on the site-resolved spectra maintained a high predictive performance when applied to composition-averaged spectra, demonstrating robustness against the diverse atomic environments encountered in experimental measurements. By enabling the direct extraction of electronic valence and local structural descriptors from a single, ensemble-averaged XANES spectrum, this approach overcomes a key bottleneck in the atomistic analysis of amorphous materials. Consequently, it offers a transferable and experimentally viable framework for quantitatively characterizing the composition–structure–property relationships of complex, multivalent, amorphous systems. Moreover, this machine learning-based XANES approach provides a transferable framework for the quantitative characterization of such systems and may facilitate the accelerated development of SiO_x-based functional materials.

Received 4th September 2025,
Accepted 15th December 2025

DOI: 10.1039/d5cp03408f

rsc.li/pccp

Introduction

Silicon suboxide (SiO_x, $0 < x < 2$) is a non-stoichiometric system in which Si–O–Si bridges and Si–Si bonds coexist. Notably, this system is regarded as a critical material for various applications, such as barrier coatings, electronic and optical devices, and energy-storage electrodes.[1,2] By adjusting the oxygen ratio, it is possible to tune the band gap, electrical conductivity, and refractive index of the suboxide, which is difficult to achieve using either stoichiometric SiO_2 or elemental silicon alone.[3,4]

Owing to this unique property, SiO_x has been utilized in many industrial areas, such as in flexible organic light-emitting diode displays, and in food- and medical-packaging films. More specifically, in these applications, multilayer barrier coatings containing SiO_x serve as functional layers that combine a high gas-barrier performance with an optical transparency and mechanical flexibility.[5–8] Since the optical and electrical properties of SiO_x can be readily tailored, this system has also been employed as a transparent electrode for ultrathin photovoltaic cells, and as a resistive-switching layer in non-volatile memory applications.[4,9,10] Recent studies have demonstrated that SiO_x retains part of the high theoretical capacity attributed to elemental silicon ($\sim$4200 mA h g^{-1}), while acting as a mechanical buffer to mitigate volume expansion, thereby affording an improved cycling stability. Accordingly, intensive efforts are underway to develop SiO_x as a high-performance anode material for next-generation lithium-ion batteries.[11–15]

Despite such vast applications and the importance of SiO_x, the development of SiO_x systems exhibiting specific desired properties remains a significant challenge. For example, since the structural tunability of SiO_x originates from variations in the bonding motifs and defect states, the material properties

[a] *Institute of Industrial Science, The University of Tokyo, Tokyo, 153-8505, Japan. E-mail: fujikata@iis.u-tokyo.ac.jp*
[b] *Science & Innovation Center, Mitsubishi Chemical Corporation, Kanagawa, 227-8502, Japan*

can fluctuate markedly with only minor changes in these parameters, revealing an inherent vulnerability.[16,17] Consequently, a quantitative, atom-level understanding of the composition–structure–property relationships existing in SiO_x compounds is therefore expected to accelerate the development of innovative devices across the energy, semiconductor, and environmental application fields.

Understanding the composition–structure–property relationships of SiO_x compounds requires quantitative measurements of their microscopic chemical states (*e.g.*, degree of Si–O bridging, oxygen-vacancy concentration, and defect distribution), since these states are closely related to the resulting dielectric and electronic transport properties. However, such measurements remain challenging, even using state-of-the-art instrumentation. For instance, conventional X-ray diffractometry yields only averaged structural information due to the fact that amorphous SiO_x lacks long-range order, thereby hindering the capture of local structures.[18,19] Although X-ray photoelectron spectroscopy (XPS) and solid-state nuclear magnetic resonance (NMR) spectroscopy can selectively probe specific chemical bonds, they are not necessarily sufficient for comprehensively evaluating the full compositional range of the SiO_x structure, in which diverse bonding motifs are known to coexist.[20–23] Additionally, the use of transmission electron microscopy (TEM) also presents limitations, such as the projection of column-averaged structures, beam-induced damage, and thin specimens that do not represent the bulk state, thereby reducing its reliability.[19,24]

In this context, X-ray absorption spectroscopy (XAS) has been widely employed for the analysis of amorphous materials.[25] In particular, the X-ray absorption near-edge structure (XANES) technique is well suited to SiO_x because the edge chemical shift and intensity sensitively reflect changes in the oxidation state and coordination environment.[24,26–29] However, interpreting the complex spectral shapes, including multiple scatterings, relies heavily on researcher expertise, and conventional fingerprint matching or parametric fitting approaches have so far been insufficient for the quantification of mixed valence states and defect distributions.[30,31]

Recent advances in this field have focused on the integration of machine learning (ML) with large libraries of first-principles XANES calculations (ML-XAS).[32,33] Several studies have demonstrated that ML models trained on theoretical spectra can extract meaningful structural descriptors. For example, S. Kiyohara showed that the radial distribution function of amorphous SiO_2 can be predicted directly from Si K-edge XANES.[34] W. Jeong applied ML-XAS to carbon nitride materials and demonstrated that XANES contains sufficient information to recover local coordination environments in these systems.[35] H. Hirai developed an ML approach capable of predicting Si K-edge XANES spectra of amorphous SiO and explored the spectral sensitivity to local structural motifs.[36] These studies collectively indicate that ML models can learn subtle spectral variations and capture statistical features associated with changes in bonding environments.

Although these contributions provide important foundations for ML-XAS, the practical methodology for applying such models to technologically relevant multivalent systems remains underdeveloped. In particular, SiO_x exhibits a wide distribution of Si–O coordination numbers, continuous valence variations, and substantial amorphous structural disorder, making it significantly more complex than the systems investigated in previous ML-XAS studies. A systematic assessment is therefore lacking regarding the extent to which ML-XAS can generalize across the full compositional range of SiO_x and how these models can be utilized as a practical workflow for the structural analysis of SiO_x across different local environments.

The current study therefore aims to elucidate the composition-dependent structures of amorphous SiO_x by integrating ML into XANES analysis, thereby providing a systematical knowledge to facilitate material design. For this purpose, nine amorphous SiO_x networks with compositions ranging from $x = 0$–2.0 (0.25 increments) are generated *via* molecular dynamics (MD) simulations, and the corresponding high-precision Si K-edge XANES spectra are obtained for each Si site using first-principles calculations. Subsequently, deep NNs are constructed to regress the RDF (reflecting the local structure) and the Bader charge (representing the local valence environment which correlates with the dielectric and electronic transport properties). Finally, the model robustness is evaluated under experimental constraints by restricting the spectral window and by using system-averaged spectra that spatially average the site-resolved spectra. These numerical experiments aim to evaluate the feasibility of performing the quantitative atomistic and electronic structural analysis of amorphous SiO_x using ML-XAS.

Methods

Construction of amorphous SiO_x structures

The initial SiO_x network structures were generated from a $4 \times 4 \times 4$ supercell of cubic SiO_2 (384 atoms, $\sim$15 Å per side) by randomly removing Si and/or O atoms to achieve the target compositions (*i.e.*, x = 0, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, and 2.00), yielding Si_{160}, $Si_{180}O_{45}$, $Si_{150}O_{75}$, $Si_{120}O_{90}$, $Si_{100}O_{100}$, $Si_{100}O_{125}$, $Si_{96}O_{144}$, $Si_{80}O_{140}$, and $Si_{72}O_{144}$, respectively. To minimize the dependence on the starting configuration, five distinct initial structures were prepared for each of the nine compositions. MD simulations were carried out with LAMMPS (large-scale atomic/molecular massively parallel simulator)[37] with the Tersoff potential.[38] Three-dimensional periodic boundary conditions were imposed; each system was heated to 5000 K in an atmospheric pressure NPT ensemble, held for 1 ns to promote complete melting, and then linearly cooled to 0 K over 3 ns to quench the structure below the glass transition temperature. Notably, the Tersoff potential has long been applied to SiO_x systems and is known to reproduce experimental atomic densities and structural features across the composition range of x = 0–2.[39]

First-principles calculation of Si K-Edge XANES spectra

For each Si site in the investigated structures, Si K-edge (1s → vacant orbital) XANES spectra were calculated using the

XSPECTRA[40,41] module implemented in Quantum Espresso (QE).[42,43] To account for core–hole effects, a reconstructed pseudopotential was automatically generated from a projector augmented wave (PAW) pseudopotential with one 1s electron removed. For this purpose, the GGA–PBE (generalized gradient approximation–Perdew–Burke–Ernzerhof) exchange–correlation functional[44] was employed. The plane-wave cut-off energy was set to 50 Ry. In addition, the self-consistent-field calculations employed a $1 \times 1 \times 1$ k-point mesh, whereas the transition-probability calculations employed a $2 \times 2 \times 2$ mesh, with the energy up to +50 eV above the absorption edge. The resulting transition intensities were convoluted with a Lorentzian broadening of γ = 0.8 eV to match the experimental resolution, and the transition energies were corrected using the total-energy difference between the excited and ground states.[45] All calculated spectra were shifted by the same constant to ensure that the main peak of α-quartz coincided with the experimental value.

To verify the computational accuracy of the developed approach, the calculated Si K-edge XANES spectra were compared with the experimental data (EXP), as shown in Fig. S1. Experimentally, the SiO_2 polymorph exhibits a sharp peak at ~1846 eV, whereas the metallic Si shows a weaker signal at ~1840 eV. Using this characteristic edge shift and peak shape as metrics, several codes were evaluated, namely QE-XSPEXTRA, CASTEP[46] (first-principles pseudopotential), and VASP[47–49] (first-principles PAW). Additionally, the CASTEP transition energies were evaluated using a procedure analogous to that employed in QE-XSPEXTRA, while VASP used the energies that were directly output by the code. QE-XSPEXTRA reproduced the overall spectral shape, including the main peak intensity and post-edge peak structure, together with an absorption-edge separation between metallic Si and SiO_2 that closely matched the experimental width. Although CASTEP captured the peak shapes well, the metallic/oxide energy shift was underestimated by >1 eV, thereby limiting its quantitative reliability for chemical-shift analysis. Furthermore, VASP yielded peak-intensity distributions that deviated markedly from the experimental values, and the onset energies for Si and SiO_2 almost overlapped, effectively eliminating the shift that is essential for distinguishing between the chemical states.

Based on these results, QE-XSPEXTRA was selected to perform the first-principles Si K-edge XANES calculations in the current study, since it can simultaneously reproduce the spectral shapes and chemical shifts with sufficient accuracy for the ML training data described below.

Dataset

A total of 45 configurations obtained from the MD simulations were analyzed. As described below, the validity of each structure was confirmed by its density and local bonding environment. Although the number of Si atoms varies among these configurations, two physicochemical quantities were defined as prediction targets for each Si atom, namely the Bader charge and the RDF. More specifically, in terms of the Bader charge, the ground-state electron densities were computed using QE, and the ionicity was defined according to modified Bader code (v1.05) described by Henkelman *et al.*[50] Subsequently, a charge was assigned to each Si atom. With regard to the RDF, for each absorbing Si atom, all Si and O atoms located within a distance of 5.0 Å were counted to construct the RDF. The radial distribution function $g(R)$ was defined as

$$g(R) = \frac{1}{4\pi R^2 \rho} \frac{\mathrm{d}N(R)}{\mathrm{d}R},$$

where $N(R)$ is the number of atoms found in a spherical shell of thickness $\mathrm{d}R$ at distance R, and ρ denotes the total atomic number density of the unit cell. These procedures yielded 5290 spectrum–target (Bader charge or RDF) pairs, which were randomly shuffled and split into a training/validation set (80%, 4232 pairs) and a test set (20%, 1058 pairs) for the final evaluation.

Neural-network architecture and training

Two NNs were constructed to regress distinct target properties from the computed spectra (1830–1880 eV, 256 energy points). One network predicts a scalar Bader charge, while the other predicts a 256-dimensional RDF. The charge model comprises three hidden layers of 256 nodes, each with leaky-ReLU activation, whereas the RDF model employs three hidden layers of 512 nodes with ReLU activation. Optimization was performed using the Adam algorithm,[51] employing an initial learning rate of 1×10^{-3} and a batch size of 128. Hyperparameters were tuned by five-fold cross-validation, and the configuration yielding the lowest root mean square error (RMSE) was adopted as the final model. Training was carried out on a single NVIDIA GeForce RT™ 4090 (24 GB). Consequently, the charge network converged in 271 epochs, while the RDF network converged in 583 epochs. In both cases, the loss function was represented by the mean squared error (MSE).

Results

Composition dependence of the amorphous network

Fig. 1 shows representative amorphous SiO_x models for each of the nine compositions (x = 0–2, 0.25 increments) generated by the MD simulations. No macroscopic defects (*e.g.*, crystalline clusters or large voids) were observed for any of the models, and in each case, a uniformly distributed network was formed within the periodic cell. The calculated bulk densities at x = 0 and 2 were determined to be 2.25 g cm^{-3} (amorphous Si) and 2.30 g cm^{-3} (amorphous SiO_2), respectively, which are in good agreement with reported experimental values of 2.29 and 2.20 g cm^{-3}, respectively.[52,53] These results indicate that the present models serve as reasonable representations of the glassy networks in the real systems.

To examine the local environment around Si in each composition, histograms were constructed for the different coordination types, classified according to the numbers of Si–Si and Si–O bonds (Fig. S2). In the case of SiO_2, the tetrahedral [SiO_4] unit containing four O neighbors dominates, yielding a highly uniform structure; an equally uniform configuration can be

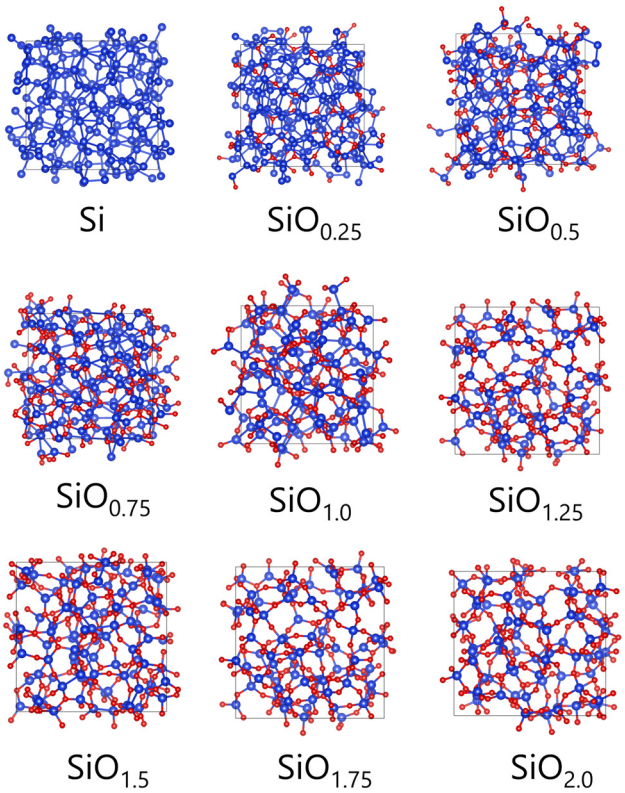

Fig. 1 Representative atomic configurations of amorphous SiO_x obtained from melt-quench molecular-dynamics simulations for nine compositions (x = 0–2 in 0.25 increments). Blue and red spheres denote Si and O atoms, respectively; gray lines indicate the periodic simulation cell.

observed for elemental Si at x = 0. However, upon reducing the oxygen content below x = 2, diverse coordination environments emerged. For instance, in the case of the $SiO_{0.25}$ system, lower-order species such as $[SiSi_2O_2]$ become prominent, which contain two or three Si–Si bonds. Additionally, at intermediate compositions ($x \approx 1$), triangular and tetrahedral units containing mixed Si–O and Si–Si bonds connect randomly, producing the greatest degree of network heterogeneity. These coordination distributions capture the composition-dependent variation in the local environments, and provide essential context for the valence fluctuations and medium-range order discussed later with regard to the XANES and RDF analyses.

The composition dependence of the XANES spectra was also considered. More specifically, Fig. 2 shows the calculated Si K-edge XANES spectra for all 45 amorphous SiO_x species (x = 0–2, 0.25 increments), wherein each spectrum is color-coded according to the composition. The spectra were computed for all Si sites within each structure and subsequently averaged, as shown in Fig. S3. In the case of the SiO_2 system, a sharp main peak was observed at ~1846 eV, similar to the case of bulk α-quartz. Upon reducing the oxygen content, the peak energy remained unchanged down to x = 1.5, although its intensity markedly decreased. For structures were $x \leq 1.5$, the main peak shifted continuously toward a lower energy and its intensity further decreased, although this decrease occurred more gradually. Such behavior was attributed to a

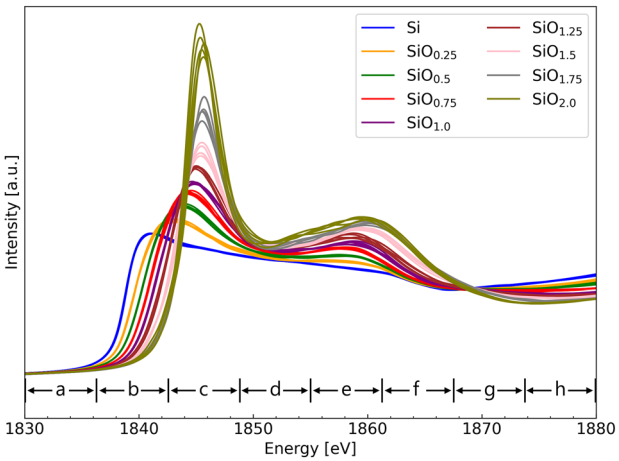

Fig. 2 Calculated Si K-edge XANES spectra for amorphous SiO_x (x = 0–2, step = 0.25). The line color denotes the composition, and five spectra are plotted for each composition, corresponding to the five MD-generated configurations. The energy windows a to h correspond to that in Fig. 5.

reduction in the average Si valence and an increased prevalence of coordinations containing Si–Si bonds, which together shifted the absorption edge to lower energy. The spectral evolution in Fig. 2 therefore reflects the composition-dependent changes in the local coordination and valence state, confirming the high sensitivity of the valence-estimation model introduced in the following section.

Prediction accuracy

Initially, the ability of the NN model to infer the silicon valence state and RDF from Si K-edge XANES spectra was evaluated. In principle, direct extraction of these descriptors should allow rational tuning of the dielectric constant, oxygen-vacancy content, and local structure, thereby supporting optimization of the conduction pathways, band gap, optical constants, and volumetric expansion in both electronic and photonic devices.

Considering a site-by-site analysis, the spectra simulated for individual Si atoms were used to predict their corresponding Bader charges and RDFs. Although XAFS lacks site-by-site information, and atomic-resolution scanning transmission electron microscopy–electron energy loss spectroscopy is difficult to apply to these glassy networks (despite its atomic resolution), this test quantitatively establishes the amount of information that a single spectrum carries regarding its corresponding electronic and structural descriptors. Thus, the results obtained from the composition-averaged spectra, which are more relevant to the experimental results, are presented later.

Fig. 3 compares the Si valence values predicted by the NN model (vertical axis) with reference values obtained from Bader-charge analysis (horizontal axis). Local spectra for each Si atom in the dataset were supplied to the model, and both the training data (blue) and test data (orange) are plotted. In both cases, the points cluster densely along the diagonal. In the case of the test data, the RMSE was calculated to be 0.110, while the coefficient of determination (R^2) was 0.991, thereby indicating that the model reproduces the valence within an average of

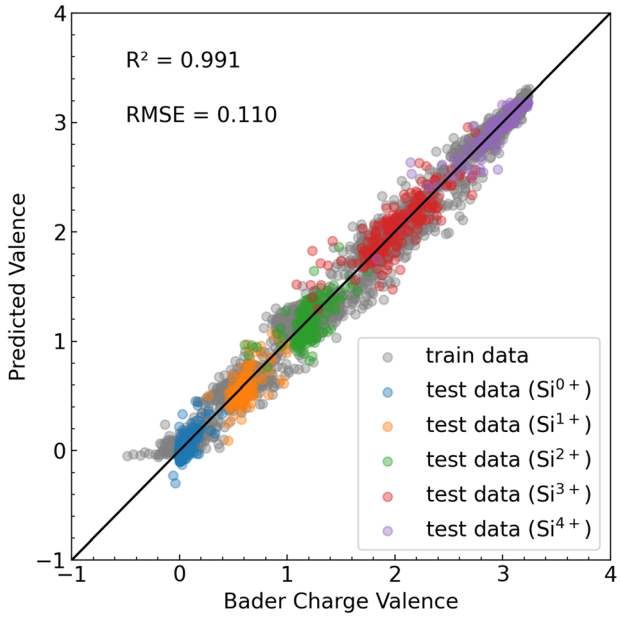

Fig. 3 Parity plot of the neural network-predicted Si valence against Bader-charge reference values. Gray dots represent training atoms. Test atoms are color-coded according to their formal valence states: blue, Si^{0+}; orange, Si^{1+}; green, Si^{2+}; red, Si^{3+}; and purple, Si^{4+}. Each valence category corresponds to the number of O neighbors in the first coordination shell (0–4).

$\sim$0.11 e^-. Due to the fact that XANES is highly sensitive to the edge shifts and white-line intensity changes that are associated with variations in the oxidation state, this model is expected to capture the subtle spectral differences that encode the valence information.[54] Color mapping revealed that the test data segregate into five well-defined clusters, each corresponding to a distinct Si–O coordination number. This partitioning confirms that the overall Si valence is dictated primarily by the number of first-shell oxygen ligands, and demonstrates that the NN reproduces this cluster structure with high fidelity. However, within each coordination class, the valence displays a finite spread, as captured by the model. Overall, these results demonstrate that the Si K-edge XANES spectra are sensitive not only to the nearest-neighbor geometry, but also to subtler electronic variations, thereby underscoring the ability of the NN to correlate spectral features with the electronic structure.

To benchmark the advantage of the present approach, we additionally performed a linear regression using three intuitive spectral descriptors, namely the white-line intensity, the integrated white-line intensity, and the edge position (Fig. S4). While this simple model captured the qualitative trend of the valence, the resulting RMSE was 0.311, considerably higher than that of the NN model, indicating that these heuristic descriptors alone are insufficient for achieving comparable accuracy.

Subsequently, the RDFs were predicted using the NN based on the Si-K edge XANES results for the respective Si sites. Fig. 4(a) shows the RMSE for each RDF in the test set, ranking from smallest to largest. For the optimal case, the RMSE was $\sim$0.1, while the largest error was $\sim$0.5, thereby confirming the good structural predictive power of the developed model across diverse chemical environments. Fig. 4(b) compares the NN-predicted RDFs (dark red) with the reference RDFs (light red) for the six representative points (A–F) indicated in Fig. 4(a). It can be seen that the most accurate case, corresponding to point A (SiO_2), and the next most accurate case, corresponding to point B (metallic Si), both reproduce all peak positions and intensities almost perfectly. For the intermediate-valence cases,

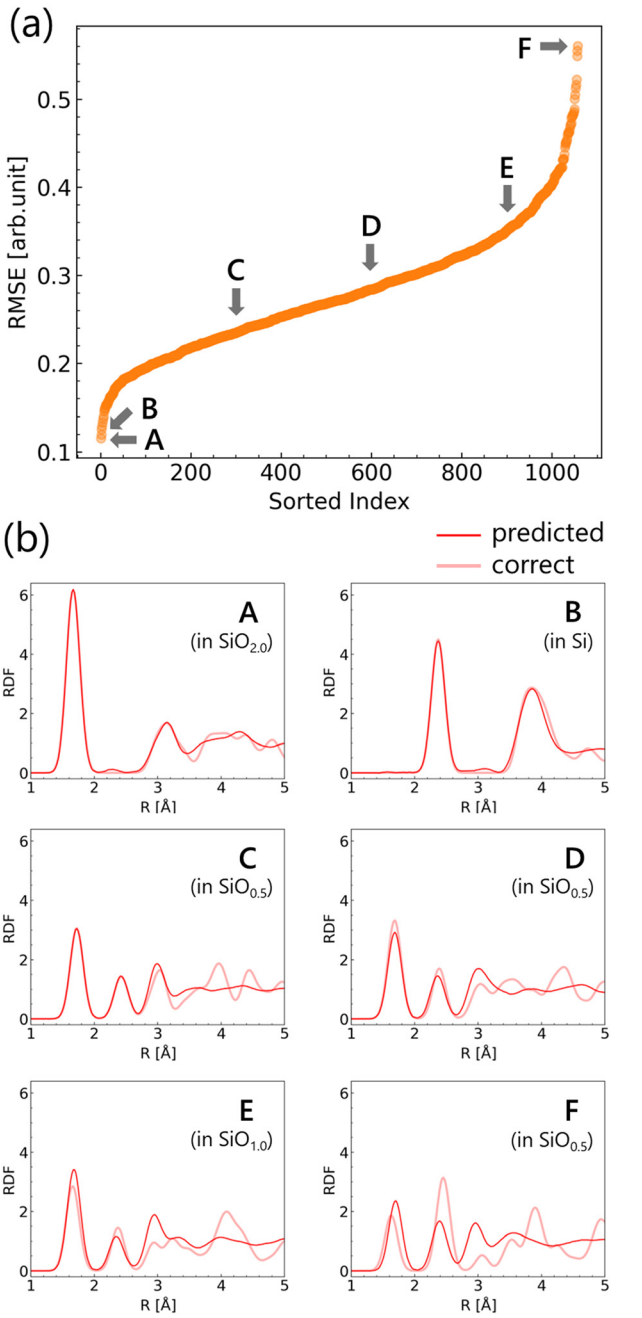

Fig. 4 (a) RMSEs of the neural network-predicted RDFs for all test structures, shown in ascending order. The arrows denote six representative cases (A)–(F). (b) Predicted (dark red) and reference (light red) RDFs for samples A–F, as identified in panel (a).

namely points C–F, the peaks within the first to third coordination shells ($R < 3$ Å) were well reproduced, whereas the intensities of the peaks corresponding to R = 3–5 Å were underestimated or exhibited slight shifts. These discrepancies likely arise because XANES is primarily sensitive to the local structure around the absorbing atom, and contributions beyond the third shell carry relatively little information.[55]

Overall, these results indicate that the NN accurately predicts the valence state along with short-to-medium-range ($\lesssim$3 Å) structural information over the full composition range from metallic Si to SiO_2 (*i.e.*, SiO_x, x = 0–2). To capture the details of longer-range coordination more precisely, a multimodal learning approach that integrates XANES with techniques possessing greater long-range sensitivity (*e.g.*, neutron diffraction or diffuse scattering) would be advantageous.

Discussion

Relationship between the spectra and the properties

The trained model was subsequently analyzed in detail to explore the means by which it can provide guidelines for optimizing XANES measurements and improve the model interpretability. For this purpose, the full Si K-edge XANES range (1830–1880 eV) was divided into eight sub-windows of 6.25 eV, and the prediction accuracy obtained from spectra within each energy window was examined.

Fig. 5 summarizes the valence-charge prediction accuracy for each sub-window (identified as a–h in Fig. 2); for comparison, the result obtained over the full energy range (*i.e.*, 1830–1880 eV) is also shown in the lower-right corner of the figure. It can be seen that the two windows spanning 1836.25–1842.50 eV

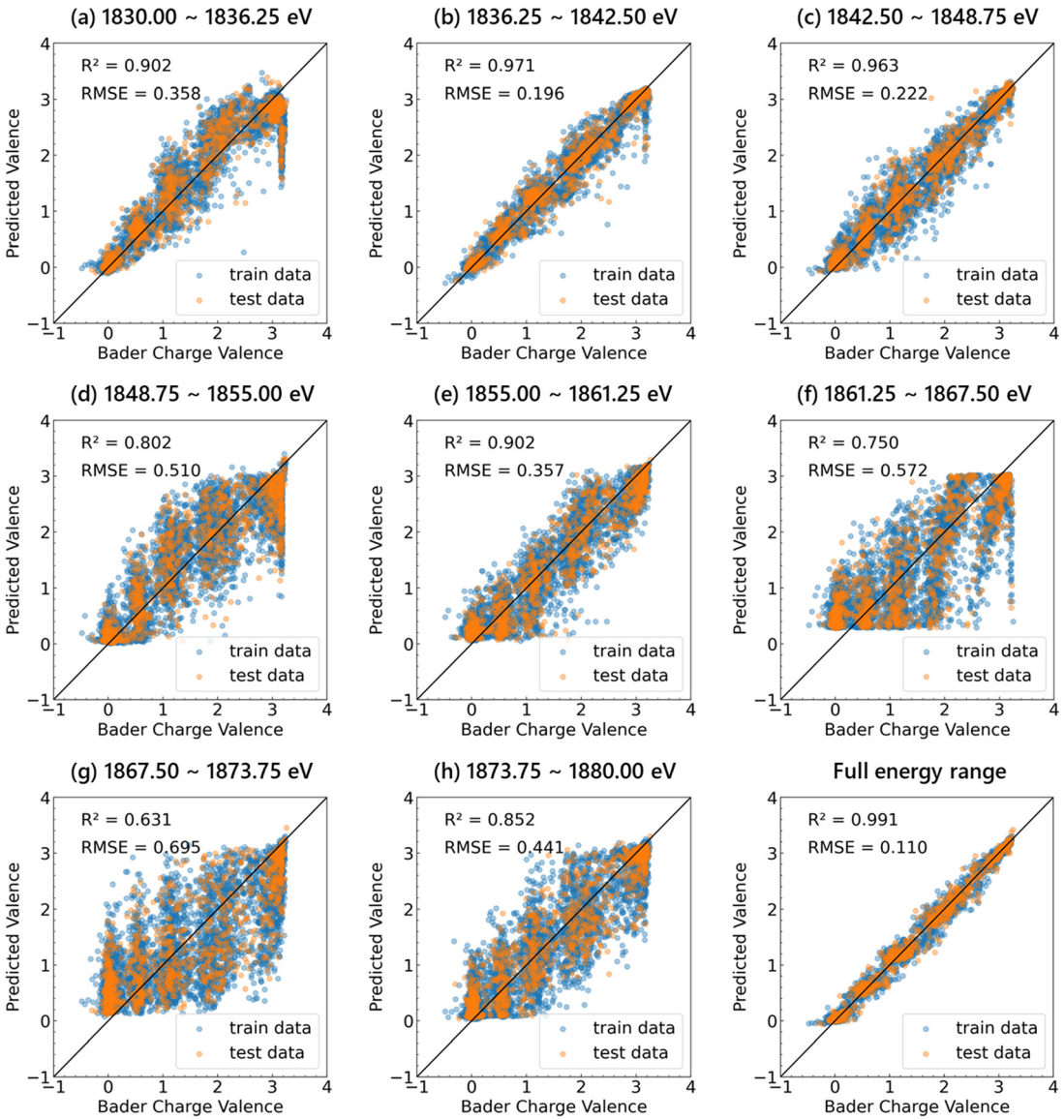

Fig. 5 Parity plots comparing the neural network-predicted Si valences with the Bader-charge references for eight models trained on individual 6.25 eV sub-windows of the Si K-edge XANES spectrum (panels a–h; the corresponding energy ranges are defined in Fig. 2). The lower-right plot corresponds to a model based on the full 1830–1880 eV range. Blue and orange markers denote the training and test data, respectively.

(Fig. 5(b)) and 1842.50–1848.75 eV (Fig. 5(c)), which cover the rising edge and apex of the main peak, yield the highest accuracies, with RMSE values of 0.196 and 0.222 and R^2 values of 0.971 and 0.963, respectively, for the test set. This superior performance is likely due to the pronounced edge shifts and peak intensity variations captured in these regions, features that are closely associated with changes in the oxidation state and which therefore carry the richest charge information. However, the prediction presented in Fig. 5(b) is less accurate for the highest-valence Si atoms (*i.e.*, SiO_2), whereas the accuracy for the lower-valence species is high; the same trend can be observed more clearly in Fig. 5(a). In these energy windows, the spectra for the low-valence Si species include the onset of the main peak, while spectra for the high-valence Si species are shifted to higher energies and lack this feature, leading to a reduced accuracy as shown in Fig. 2.

It was also found that the prediction accuracy deteriorated in the high-energy windows (>1861 eV). For example, the 1861.25–1867.50 eV window (Fig. 5(f)) gave an RMSE of 0.572 and an R^2 value of 0.750, presumably because the spectral features flatten with an increasing energy separation from the edge. When two adjacent sub-windows were merged into wider 25 eV bands (Fig. S5), the low-energy window (*i.e.*, 1830.00–1855.00 eV) achieved an RMSE of 0.157 and an R^2 value of 0.981, which were comparable to the results obtained for the full-range model. In contrast, the high-energy window (*i.e.*, 1855.00–1880.00 eV) exhibited only a modest improvement (RMSE = 0.268, R^2 = 0.945) but still contributed measurably to the charge prediction.

Fig. 6 assesses RDF prediction accuracy for each Si atom as a function of the sub-window. More specifically, from Fig. 6(a), it can be seen that as in the case of the valence estimation, the 1836.25–1842.50 and 1842.50–1848.75 eV windows yield superior accuracies, and each sub-window curve lies above the full-range curve (yellow line), thereby indicating that wide-range information is indispensable for reproducing the RDFs. Although the accuracy was found to decline markedly in the mid- to high-energy windows (>1861 eV), the RMSE values generally remained <1.0, indicating that broad, low-intensity features far from the edge carry complementary structural information. In the highest energy window (Fig. 5(h)), the RMSE decreased relative to that in the preceding windows (i.e., Fig. 5(f) and (g)). This improvement was attributed to the onset of the EXAFS region, where oscillations that strongly encode structural details become visible. When the windows were merged into 25 eV bands (Fig. S6), the low- and high-energy regions provided almost identical accuracies, which contrasts to the behavior observed for the valence estimations.

These analyses reveal that different portions of the Si K-edge XANES spectrum correlate more strongly with specific properties, and illustrate how ML models can provide guidance for designing more informed measurement strategies while offering physical insights into the investigated systems. In terms of the valence (charge), the edge shift and the main peak intensity constitute the principal information sources, and the spectra confined to this region delivered an accuracy comparable to

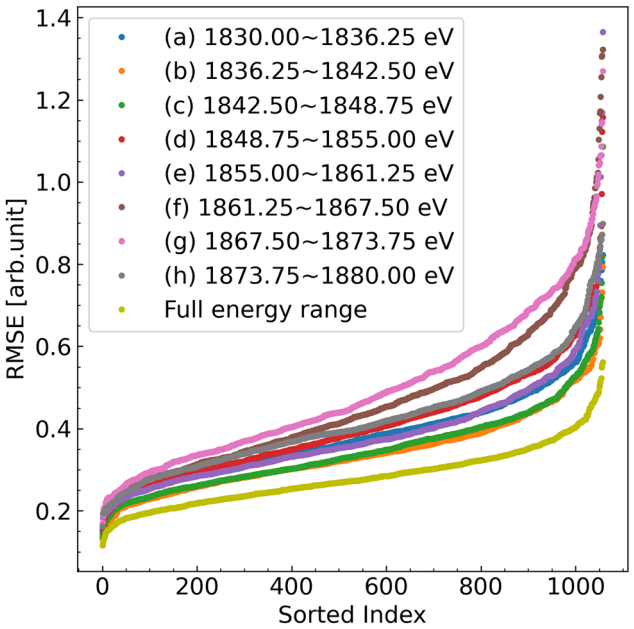

Fig. 6 Sorted RMSE distributions for RDF predictions generated by neural-network models trained on individual 6.25 eV sub-windows of the Si K-edge XANES spectrum (colored curves a–h; the corresponding energy ranges defined in Fig. 2) and by a model using the full 1830–1880 eV range (yellow curve).

that of the full spectrum. With regard to the RDF, although the same region remains important, higher-energy features were found to lower the degree of error, thereby suggesting that structural information is distributed across the entire spectrum.

Generalization to averaged spectra

In an actual XANES experiment, the signal detected corresponds to an averaged spectrum arising from numerous Si atoms. Consequently, before applying the NN model to experimental data, it is necessary to verify that a network trained exclusively on single-site spectra can also make reliable predictions from such averaged spectra. Thus, as presented in Fig. 7, using the NN trained on individual Si spectra (*i.e.*, the same model used in Fig. 3), the mean valences were predicted from the sample-averaged spectra (shown in Fig. 2) of the 45 systems spanning nine Si/O ratios from SiO_0 to SiO_2 (red dots). For comparison, valence predictions obtained from individual Si spectra (gray dots, identical to the Fig. 3 test set) are also displayed.

Although the NN had been trained only on single-atom spectra, the predictions based on the ensemble spectra retained high degrees of accuracy (RMSE = 0.105, R^2 = 0.989), indicating that the model had successfully learned statistically robust valence-dependent features. Notably, a slight underestimation was observed for the $SiO_{0.75}$–$SiO_{1.25}$ compositions, where low-, intermediate-, and high-valence Si atoms coexist. In this range, the overlap of the respective spectral signatures broadened the averaged spectrum and shifted it toward features characteristic of a lower valence. Nevertheless, the deviation remained

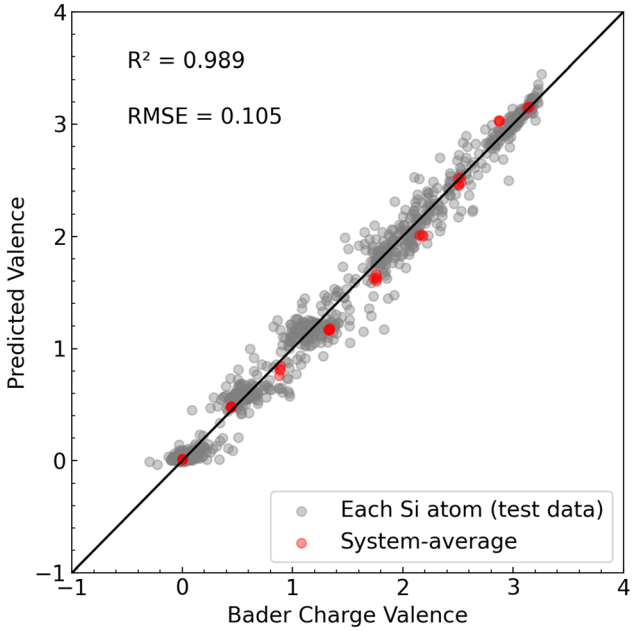

Fig. 7 Correlation between the Bader-charge valence and the valence predicted by Si K-edge XANES using the neural network. Gray dots represent individual Si sites in the test set, while red dots represent values obtained from the spectra averaged over each of the 45 amorphous SiO_x systems (nine compositions, five configurations each).

<0.2 e^-, thereby confirming that the mean valence can be retrieved with practical accuracy.

Fig. 8 assesses the accuracy with which the same ensemble spectra reproduce the mean RDF. According to the RMSE distribution presented in Fig. 8(a), the best agreement can be observed for metallic Si, followed by $SiO_{0.75}$; however, even the largest RMSE (*i.e.*, for $SiO_{0.25}$) does not exceed $\sim$0.23. Notably, this is comparable to the top 10% of single-atom RDF predictions. In the case where the largest RMSE was observed (Fig. 8(b)), the NN reproduces both the positions and the intensities of peaks within the first to third coordination shells ($R < 3$ Å) and matches the medium-range peaks at 3–5 Å reasonably well. These results imply that, during training on individual spectra, the NN captured not only the local structural motifs, but also their statistical fluctuations, thereby acquiring the capacity to infer average structures from ensemble-averaged spectra.

While the present study demonstrates that our model can reliably recover ensemble-averaged valence and RDF from simulated sample-averaged Si K-edge XANES spectra, extending this capability to experimental spectra remains an important direction for future work. In particular, the overall performance will ultimately be limited by the accuracy of first-principles XANES calculations for amorphous SiO_x, and further theoretical advances in excited-state electronic structure and core–hole treatments will be required before truly quantitative experiment–theory consistency can be assumed.

A promising route toward experimental applicability is to combine large-scale training on simulated spectra with transfer-learning or domain-adaptation schemes that fine-tune the model on a comparatively small set of experimental XANES data with

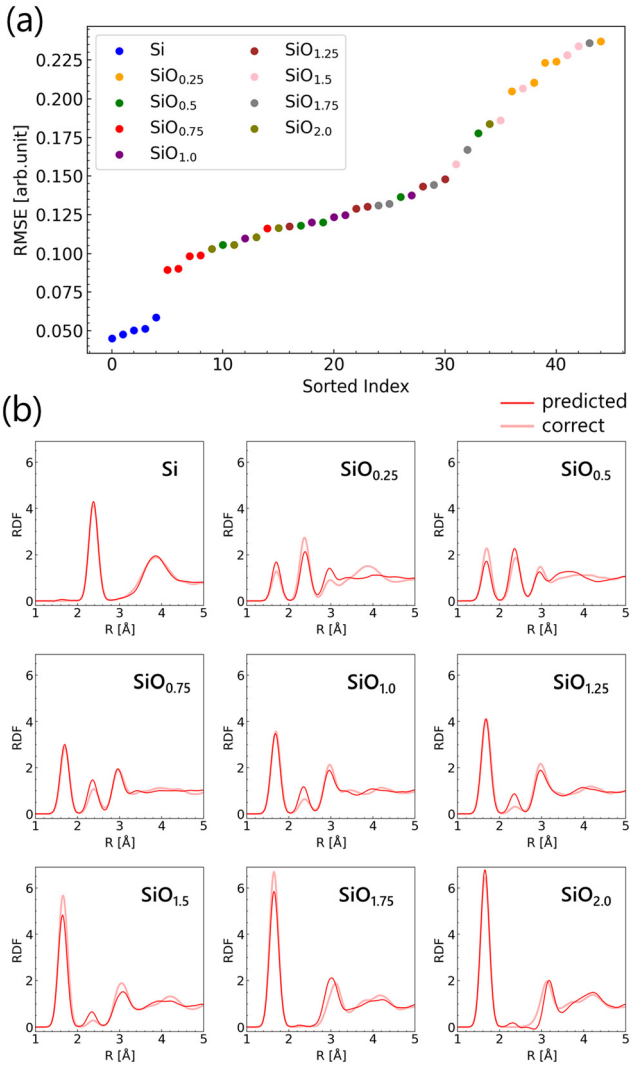

Fig. 8 (a) Composition-coded RMSE distributions for the mean RDFs predicted by the neural network from the system-averaged Si K-edge XANES results of 45 amorphous SiO_x models (nine compositions, five configurations each). (b) Predicted (dark red) *versus* reference (light red) mean RDFs for the least-accurate sample in each composition, illustrating agreement across the Si–SiO_2 stoichiometric range.

independently characterized structures and oxidation states. In some XAS studies, models trained on theory-based spectra have already been applied to experimental XANES/EXAFS data,[34,56] and related spectroscopies such as Raman have reported successful examples of transfer learning,[57] suggesting that similar sim-to-real strategies are feasible in this context. In this sense, our results should be viewed as a proof of concept on simulated data that outlines a concrete pathway toward an experiment-ready framework once improved theoretical spectra and carefully curated experimental benchmarks become available.

Conclusions

In this study, the composition-dependent structure of amorphous SiO_x ($0 \leq x \leq 2$) was systematically clarified at the atomic scale by integrating machine learning (ML) into Si

K-edge X-ray absorption near-edge structure (XANES) analysis. Nine compositions, spaced by x intervals of 0.25, were generated using molecular dynamics simulations, yielding a total of 45 network models. Using the XSPECTRA module of Quantum Espresso (QE), high-precision spectra were calculated for each Si site, establishing a database that demonstrated the ability of QE-XSPECTRA to accurately reproduce both the SiO_2/Si edge shift and the overall spectral shape. Additionally, using deep neural networks, two target properties were regressed from the XANES inputs corresponding to 256 energy points, namely the Bader valence (root mean squared error, RMSE = 0.117; coefficient of determination, R^2 = 0.990) and the radial distribution function (RDF) to a distance of 5 Å (RMSE $\leq$ 0.5). The valence model also distinguished discrete Si–O coordination numbers (0–4), indicating its suitability for application with mixed-valence systems. Furthermore, sub-spectrum analysis revealed that the edge-onset to the main peak region (1836–1849 eV) contained the greatest amount of valence information, whereas an accurate RDF prediction required a wider energy range, including high-energy features. Notably, the network trained solely on single-site spectra reproduced the mean valence and mean RDF from ensemble-averaged spectra with practical accuracy, providing a proof of concept for applying this approach to experimental XANES data in future studies. These results demonstrate that ML combined with X-ray absorption spectroscopy can simultaneously quantify the valence, local structure, and medium-range order across the continuous composition space of SiO_x, while also providing guidance for high-throughput analysis and for optimizing the measurement energy ranges. Looking ahead, an important direction will be to develop transfer-learning strategies that adapt models trained on large-scale simulated XANES datasets to experimental spectra using a comparatively small number of well-characterized reference samples. Such sim-to-real transfer has the potential to make the present framework applicable to quantitative interpretation and high-throughput screening of experimental XANES data for amorphous functional materials, including SiO_x.

Author contributions

Y. F. and H. S. conceived and designed the study. Y. F. performed the calculations and wrote the original draft. T. M. supervised the project. All authors reviewed and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

The data supporting this article have been included as part of the supplementary information (SI). Supplementary information is available. See DOI: https://doi.org/10.1039/d5cp03408f.

Acknowledgements

This study was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) (Grant Number 24H00042), the Japan Science and Technology Agency (JST), and the New Energy and Industrial Technology Development Organization (NEDO). All calculations were performed on the Mitsubishi Chemical Corporation (MCC) high-performance computer (HPC) system ''NAYUTA'', where ''NAYUTA'' is a nickname for the MCC HPC, and is not a product or service name of MCC.

Notes and references

1 N. Tomozeiu, *Optoelectronics – Materials and Techniques*, IntechOpen, 2011.
2 G. Lucovsky, S. Y. Lin, P. D. Richard, S. S. Chao, Y. Takagi, P. Pai, J. E. Keem and J. E. Tyler, Local atomic structure of silicon suboxides (SiO_x, $x < 2$), *J. Non-Cryst. Solids*, 1985, **75**, 429–434.
3 H.-P. Ma, J.-H. Yang, J.-G. Yang, L.-Y. Zhu, W. Huang, G.-J. Yuan, J.-J. Feng, T.-C. Jen and H.-L. Lu, Systematic Study of the SiO_x Film with Different Stoichiometry by Plasma-Enhanced Atomic Layer Deposition and Its Application in SiO_x/SiO_2 Super-Lattice, *Nanomaterials*, 2019, **9**, 55.
4 P. Cuony, D. T. L. Alexander, I. Perez-Wurfl, M. Despeisse, G. Bugnon, M. Boccard, T. Söderström, A. Hessler-Wyser, C. Hébert and C. Ballif, Silicon Filaments in Silicon Oxide for Next-Generation Photovoltaics, *Adv. Mater.*, 2012, **24**, 1182–1186.
5 J. Lange and Y. Wyser, Recent innovations in barrier technologies for plastic packaging—a review, *Packag. Technol. Sci.*, 2003, **16**, 149–158.
6 C.-Y. Wu, R.-M. Liao, L.-W. Lai, M.-S. Jeng and D.-S. Liu, Organosilicon/silicon oxide gas barrier structure encapsulated flexible plastic substrate by using plasma-enhanced chemical vapor deposition, *Surf. Coat. Technol.*, 2012, **206**, 4685–4691.
7 S. Zhang, W. Xue and Z. Yu, Moisture barrier evaluation of SiO_x/SiN_x stacks on polyimide substrates using electrical calcium test, *Thin Solid Films*, 2015, **580**, 101–105.
8 S. Qin, S. Xiang, B. Eberle, K. Xie and J. C. Grunlan, High Moisture Barrier with Synergistic Combination of SiO_x and Polyelectrolyte Nanolayers, *Adv. Mater. Interfaces*, 2019, **6**, 1900740.
9 A. Mehonic, A. L. Shluger, D. Gao, I. Valov, E. Miranda, D. Ielmini, A. Bricalli, E. Ambrosi, C. Li, J. J. Yang, Q. Xia and A. J. Kenyon, Silicon Oxide (SiO_x): A Promising Material for Resistance Switching?, *Adv. Mater.*, 2018, **30**, 1801187.
10 D. Qiu, A. Lambertz, W. Duan, L. Mazzarella, P. Wagner, A. B. Morales-Vilches, G. Yang, P. Procel, O. Isabella, B. Stannowski and K. Ding, A Review: Application of Doped Hydrogenated Nanocrystalline Silicon Oxide in High Efficiency Solar Cell Devices, *Adv. Sci.*, 2024, **11**, 2403728.
11 M. Zhang, N. Liang, D. Hao, Z. Chen, F. Zhang, J. Yin, Y. Yang and L. Yang, Recent advances of SiO_x-based anodes

for sustainable lithium-ion batteries, *Nano Res. Energy*, 2023, **2**, e9120077.
12 X. Zhou, Z. Qi, Q. Liu, J. Tian, M. Liu, K. Dong and Z. Lei, Research Progress of Silicon Suboxide-Based Anodes for Lithium-Ion Batteries, *Front. Mater.*, 2021, **7**, 628233.
13 T. Chen, J. Wu, Q. Zhang and X. Su, Recent advancement of SiO_x based anodes for lithium-ion batteries, *J. Power Sources*, 2017, **363**, 126–144.
14 H. Li, H. Li, Z. Yang, L. Yang, J. Gong, Y. Liu, G. Wang, Z. Zheng, B. Zhong, Y. Song, Y. Zhong, Z. Wu and X. Guo, SiO_x Anode: From Fundamental Mechanism toward Industrial Application, *Small*, 2021, **17**, 2102641.
15 Y. Yu, H. Gong, X. He, L. Ming, X. Wang and X. Ou, Alleviating the volume expansion of silicon anodes by constructing a high-strength ordered multidimensional encapsulation structure, *Chem. Sci.*, 2024, **15**, 15891–15899.
16 P. Ragonese, B. Kalinic, L. Franco, L. Girardi, B. M. Fernández Peréz, D. Carbonera, G. Mattei, G.-A. Rizzi and C. Maurizio, Effect of Interfacial SiO_x Defects on the Functional Properties of Si-Transition Metal Oxide Photoanodes for Water Splitting, *ACS Appl. Mater. Interfaces*, 2023, **15**, 46933–46940.
17 Y. Yang, Y. Liu, X. Jiang, L. Zhao, P. Wang and Y. Zhang, Rational design of SiO_x based anode materials for next generation lithium-ion batteries, *Mater. Adv.*, 2024, **5**, 896–919.
18 O. M. Feroughi, C. Sternemann, C. J. Sahle, M. A. Schroer, H. Sternemann, H. Conrad, A. Hohl, G. T. Seidler, J. Bradley, T. T. Fister, M. Balasubramanian, A. Sakko, K. Pirkkalainen, K. Hämäläinen and M. Tolan, Phase separation and Si nanocrystal formation in bulk SiO studied by X-ray scattering, *Appl. Phys. Lett.*, 2010, **96**, 081912.
19 A. Hirata, S. Kohara, T. Asada, M. Arao, C. Yogi, H. Imai, Y. Tan, T. Fujita and M. Chen, Atomic-scale disproportionation in amorphous silicon monoxide, *Nat. Commun.*, 2016, **7**, 11591.
20 R. Alfonsetti, L. Lozzi, M. Passacantando, P. Picozzi and S. Santucci, XPS studies on SiO_x thin films, *Appl. Surf. Sci.*, 1993, **70–71**, 222–225.
21 C.-M. Park, W. Choi, Y. Hwa, J.-H. Kim, G. Jeong and H.-J. Sohn, Characterizations and electrochemical behaviors of disproportionated SiO and its composite for rechargeable Li-ion batteries, *J. Mater. Chem.*, 2010, **20**, 4854–4860.
22 Z. Ma, X. Liao, J. He, W. Cheng, G. Yue, Y. Wang and G. Kong, Annealing behaviors of photoluminescence from SiO_x:H, *J. Appl. Phys.*, 1998, **83**, 7934–7939.
23 K. Kitada, O. Pecher, P. C. M. M. Magusin, M. F. Groh, R. S. Weatherup and C. P. Grey, Unraveling the Reaction Mechanisms of SiO Anodes for Li-Ion Batteries by Combining *in Situ* ^{7}Li and *ex Situ* ^{7}Li/^{29}Si Solid-State NMR Spectroscopy, *J. Am. Chem. Soc.*, 2019, **141**, 7014–7027.
24 A. Hohl, T. Wieder, P. A. van Aken, T. E. 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*, 2003, **320**, 255–280.
25 E. Brack, M. Plodinec, M.-G. Willinger and C. Copéret, Implications of Ga promotion and metal–oxide interface from tailored PtGa propane dehydrogenation catalysts supported on carbon, *Chem. Sci.*, 2023, **14**, 12739–12746.
26 A. Barranco, F. Yubero, J. P. Espinós, P. Groening and A. R. González-Elipe, Electronic state characterization of SiO_x thin films prepared by evaporation, *J. Appl. Phys.*, 2005, **97**, 113714.
27 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.*, 2021, **196**, 110555.
28 S. Yu Turishchev, V. A. Terekhov, D. A. Koyuda, A. V. Ershov, A. I. Mashin, E. V. Parinova, D. N. Nesterov, D. A. Grachev, I. A. Karabanova and E. P. Domashevskaya, Formation of silicon nanocrystals in multilayer nanoperiodic a-SiO_x/insulator structures from the results of synchrotron investigations, *Semiconductors*, 2017, **51**, 349–352.
29 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*, 2017, **304**, 1–6.
30 A. A. Guda, S. A. Guda, A. Martini, A. N. Kravtsova, A. Algasov, A. Bugaev, S. P. Kubrin, L. V. Guda, P. Šot, J. A. van Bokhoven, C. Copéret and A. V. Soldatov, Understanding X-ray absorption spectra by means of descriptors and machine learning algorithms, *npj Comput. Mater.*, 2021, **7**, 203.
31 S. Xiang, M. R. Knecht and A. Frenkel, Machine learning-assisted X-ray absorption analysis of bimetallic catalysts, *Mater. Horiz.*, 2025, **12**, 4487–4495.
32 J. Timoshenko, D. Lu, Y. Lin and A. I. Frenkel, Supervised Machine-Learning-Based Determination of Three-Dimensional Structure of Metallic Nanoparticles, *J. Phys. Chem. Lett.*, 2017, **8**, 5091–5098.
33 K. Kikumasa, S. Kiyohara, K. Shibata and T. Mizoguchi, Quantification of the Properties of Organic Molecules Using Core-Loss Spectra as Neural Network Descriptors, *Adv. Intell. Syst.*, 2022, **4**, 2100103.
34 S. Kiyohara and T. Mizoguchi, Radial Distribution Function from X-ray Absorption near Edge Structure with an Artificial Neural Network, *J. Phys. Soc. Jpn.*, 2020, **89**, 103001.
35 W. Jeong, W. Sun, M. F. Calegari Andrade, L. F. Wan, T. M. Willey, M. H. Nielsen and T. A. Pham, Integrating Machine Learning Potential and X-ray Absorption Spectroscopy for Predicting the Chemical Speciation of Disordered Carbon Nitrides, *Chem. Mater.*, 2024, **36**, 4144–4156.
36 H. Hirai, T. Iizawa, T. Tamura, M. Karasuyama, R. Kobayashi and T. Hirose, Machine-learning-based prediction of first-principles XANES spectra for amorphous materials, *Phys. Rev. Mater.*, 2022, **6**, 115601.
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.*, 2022, **271**, 108171.

38 S. Munetoh, T. Motooka, K. Moriguchi and A. Shintani, Interatomic potential for Si–O systems using Tersoff parameterization, *Comput. Mater. Sci.*, 2007, **39**, 334–339.

39 C. Ugwumadu, K. N. Subedi, R. Thapa, P. Apsangi, S. Swain, M. N. Kozicki and D. A. Drabold, Structure, vibrations and electronic transport in silicon suboxides: Application to physical unclonable functions, *J. Non-Cryst. Solids: X*, 2023, **18**, 100179.

40 M. Taillefumier, D. Cabaret, A.-M. Flank and F. Mauri, X-ray absorption near-edge structure calculations with the pseudopotentials: Application to the K edge in diamond and α-quartz, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2002, **66**, 195107.

41 C. Gougoussis, M. Calandra, A. P. Seitsonen and F. Mauri, First-principles calculations of X-ray absorption in a scheme based on ultrasoft pseudopotentials: From α-quartz to high-T_c compounds, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2009, **80**, 075102.

42 P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari and R. M. Wentzcovitch, QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials, *J. Phys.: Condens. Matter*, 2009, **21**, 395502.

43 P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. Buongiorno Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, N. Colonna, I. Carnimeo, A. Dal Corso, S. de Gironcoli, P. Delugas, R. A. DiStasio, A. Ferretti, A. Floris, G. Fratesi, G. Fugallo, R. Gebauer, U. Gerstmann, F. Giustino, T. Gorni, J. Jia, M. Kawamura, H.-Y. Ko, A. Kokalj, E. Küçükbenli, M. Lazzeri, M. Marsili, N. Marzari, F. Mauri, N. L. Nguyen, H.-V. Nguyen, A. Otero-de-la-Roza, L. Paulatto, S. Poncé, D. Rocca, R. Sabatini, B. Santra, M. Schlipf, A. P. Seitsonen, A. Smogunov, I. Timrov, T. Thonhauser, P. Umari, N. Vast, X. Wu and S. Baroni, Advanced capabilities for materials modelling with Quantum ESPRESSO, *J. Phys.: Condens. Matter*, 2017, **29**, 465901.

44 J. P. Perdew, K. Burke and M. Ernzerhof, Generalized Gradient Approximation Made Simple, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

45 T. Mizoguchi, I. Tanaka, S.-P. Gao and C. J. Pickard, First-principles calculation of spectral features, chemical shift and absolute threshold of ELNES and XANES using a plane wave pseudopotential method, *J. Phys.: Condens. Matter*, 2009, **21**, 104204.

46 S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. J. Probert, K. Refson and M. C. Payne, First principles methods using CASTEP, *Z. Kristallogr. – Cryst. Mater.*, 2005, **220**, 567–570.

47 G. Kresse and J. Hafner, *Ab initio* molecular dynamics for liquid metals, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1993, **47**, 558–561.

48 G. Kresse and J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, *Comput. Mater. Sci.*, 1996, **6**, 15–50.

49 G. Kresse and J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 11169–11186.

50 G. Henkelman, A. Arnaldsson and H. Jónsson, A fast and robust algorithm for Bader decomposition of charge density, *Comput. Mater. Sci.*, 2006, **36**, 354–360.

51 D. P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, *arXiv*, 2017, preprint, arXiv:1412.6980, DOI: 10.48550/arXiv.1412.6980.

52 J. S. Custer, M. O. Thompson, D. C. Jacobson, J. M. Poate, S. Roorda, W. C. Sinke and F. Spaepen, Density of amorphous Si, *Appl. Phys. Lett.*, 1994, **64**, 437–439.

53 *CRC Handbook of Chemistry and Physics*, ed. W. M. Haynes, CRC Press, Boca Raton, 92nd edn, 2011.

54 A. A. Guda, S. A. Guda, A. Martini, A. N. Kravtsova, A. Algasov, A. Bugaev, S. P. Kubrin, L. V. Guda, P. Šot, J. A. van Bokhoven, C. Copéret and A. V. Soldatov, Understanding X-ray absorption spectra by means of descriptors and machine learning algorithms, *npj Comput. Mater.*, 2021, **7**, 1–13.

55 C. S. Spanjers, P. Guillo, T. D. Tilley, M. J. Janik and R. M. Rioux, Identification of Second Shell Coordination in Transition Metal Species Using Theoretical XANES: Example of Ti–O–(C, Si, Ge) Complexes, *J. Phys. Chem. A*, 2017, **121**, 162–167.

56 J. Timoshenko, A. Anspoks, A. Cintins, A. Kuzmin, J. Purans and A. I. Frenkel, Neural Network Approach for Characterizing Structural Transformations by X-Ray Absorption Fine Structure Spectroscopy, *Phys. Rev. Lett.*, 2018, **120**, 225502.

57 J. Kamran, J. Hniopek and T. Bocklitz, Transfer-Learning Deep Raman Models Using Semiempirical Quantum Chemistry, *J. Chem. Inf. Model.*, 2025, **65**, 6632–6643.

