Quantification of the Properties of Organic Molecules Using Core-Loss Spectra as Neural Network Descriptors

Kakeru Kikumasa, Shin Kiyohara, Kiyou Shibata, and Teruyasu Mizoguchi*

Artificial neural networks are applied to quantify the properties of organic molecules by introducing a new descriptor, a core-loss spectrum, which is typically observed experimentally using electron or X-ray spectroscopy. Using the calculated C K-edge core-loss spectra of organic molecules as the descriptor, the neural network models quantitatively predict both intensive and extensive properties, such as the gap between highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) (HOMO–LUMO gap) and internal energy. The prediction accuracy estimated by the mean absolute errors for the HOMO–LUMO gap and internal energy is 0.205 and 97.3 eV, respectively, which are comparable with those of previously reported chemical descriptors. This study indicates that the neural network approach using the core-loss spectra as the descriptor has the potential to deconvolute the abundant information available in core-loss spectra for both prediction and experimental characterization of many physical properties. The study shows the practical potential of machine-learning-based material property measurements taking advantage of experimental core-loss spectra, which can be measured with high sensitivity, high spatial resolution, and high temporal resolution.

1. Introduction

The characterization of materials is key to their development as the knowledge obtained by various characterization techniques, such as atomic coordination and chemical bonding, is often correlated with material properties. In other words, material characterization is conducted to determine the relationships among the structure, properties, processing, and performance of materials.

Among the many available characterization techniques, core-loss spectroscopy using electrons and X-rays, specifically electron energy loss near-edge structure (ELNES) and X-ray near-edge structure (XANES), corresponds to the core electron adsorption spectrum reflecting the partial densities of state (PDOS) of unoccupied orbitals. These spectroscopic techniques offer experimental characterization with atomic-scale spatial resolution,[1–3] high temporal resolution,[4,5] and high sensitivity.[6] Thus, ELNES/XANES techniques have been used to characterize a wide variety of materials systems, including electronic devices, biological samples, and drug delivery systems.[7–9]

Although ELNES/XANES measurements can provide useful information for the development of materials, the relationships between their spectral features and material properties are ambiguous. Furthermore, the relationship between ELNES/XANES and extensive properties such as molecular weight and internal energy has not yet been explored.

If direct connections between the spectral features and those properties are established with machine learning (ML), measurements of properties, such as optical properties, electron conductivity, density, and stability, with high sensitivity, high spatial resolution, and high temporal resolution, can be achieved by utilizing the merit of ELNES/XANES. Furthermore, the prediction and design of material properties using the calculated ELNES/XANES spectra can be achieved through direct connection.

In recent years, ML approaches have been widely adopted in materials science to predict material properties with suitable descriptors for screening in a huge materials space and to overcome the limitations of the manual analysis of big data.[10–14] In the field of spectroscopy, ML approaches have been utilized from the viewpoint of characterization to analyze spectral data obtained from Raman spectroscopy, nuclear magnetic resonance (NMR) spectroscopy, and infrared spectroscopy (IR).[15–17]

From ELNES/XANES spectral data, ML can extract structural information, such as the bond length and angle,[18] oxidation state,[19] radial distribution function (RDF),[20] and 3D metallic structure.[21] In addition to extract structural information, interpretations and classifications of ELNES/XANES spectra have also been performed by combining supervised and unsupervised ML approaches.[22,23]

In a recent study by our group, we developed an ML approach to directly quantify the local structure and chemical bonding
from ELNES/XANES spectra as descriptors. This approach accurately determined the intensive properties at atomic sites, such as the Mulliken charge and bond overlap population; however, quantification of the properties of the material itself, which would be more useful for practical materials development, has not yet been achieved. Furthermore, the prediction of the extensive properties, such as molecular weight and internal energy, using the core-loss spectrum has not been achieved either.

In this study, we developed a new ML approach to directly and quantitatively unveil hidden information on material properties from ELNES/XANES spectral data. Using the molecular structures of organic molecules in computational databases we performed simulations on the carbon K-edge ELNES/XANES spectra of 22,155 organic molecules and used them as the input to the neural network model to predict the molecular properties. In other words, ELNES/XANES spectra were used as descriptors to quantify the properties of the organic molecules.

Although accurate predictions of the molecular properties have been reported using databases such as SchNet, deep neural network (DNN), the graph convolutional (GC) multitask model, deep tensor neural network (DTNN), and message passing neural network (MPNN), these studies used dedicated chemical descriptors. For instance, MPNN used chemical information such as acceptor, donor, and hybridization. DTNN and SchNet used a large number of learning parameters, and DNN and multitasking required the coordination of constituent atoms. Thus, the atomic and electronic structures of molecules were indispensable for these predictions, but these descriptors are not always measurable in experiments.

The advantage of the descriptor proposed in this study, the core-loss spectrum, is that it is experimentally observable, and thus it has the potential to be measured with high sensitivity, high spatial resolution, and high temporal resolution by modern EELS and XAS instrumentation. This method is greatly beneficial for materials development because it has the potential to unveil where, when, and how the material properties arise.

2. Results and Discussion

In this study, we constructed an artificial feedforward neural network (FNN) to predict intensive and extensive material properties. This model can handle multidimensional input data, such as spectral data, and interpret complicated, nonlinear relationships between spectral data and molecular properties. A schematic of the constructed FNN architecture used to predict the properties based on the input spectra is shown in Figure 1.

To simulate the C K-edge, we selected 22,155 organic molecules with up to eight CONF atoms from the dataset. To obtain the spectra of 22,155 molecules, a total of 117,337 C K-edge simulations were performed owing to the presence of carbon atoms with different environments, and both the ground state and excited state were simulated to obtain the theoretical excitation energy.

Examples of the simulated C K-edge spectra of representative alkane, cycloalkane, alkyne, and alcohol molecules are shown in Figure 2a–d, respectively. The spectral features of molecules depend on their composition, structure, and bonding characteristics. The simulated spectra are compared with the experimental spectra in Figure 2e. The figure shows that the simulated spectra essentially reproduce the experimental spectrum features. However, it should be mentioned that the simulated spectra do not perfectly match the experimental spectra. This mismatch can be ascribed to the following three causes: 1) more accurate treatment is necessary for the core–hole effects, such as excitonic simulation based on the Bethe–Salpeter method. 2) The natural width of the respective transitions was not considered. 3) The Rydberg state, which is related to the decay process of the excited electron, was not considered in the present simulation.

As described in Section 4, all calculated spectra were shifted such that their thresholds were set at 0 eV. In other words, excitation energy information was intentionally eliminated and only the spectral shape was used as the input for the FNN because a quantitative measurement of the excitation energy in an experiment is often difficult to obtain owing to equipment stability. The calculated C K-edge spectra were discretized into 240 dimensions for use in the FNN as input data.

As outputs of our FNN model, 12 different molecular properties (listed in Table 1) were used. We trained one FNN for each of the 12 molecular properties. First, we constructed an FNN using only the C K-edge spectral features as input data (“Model A”). The hyperparameters were tuned using a grid search (Table S1, Supporting Information). Then, we constructed another model, “Model B,” by combining the C K-edge spectra and additional compositional inputs. In addition to these models dedicated to one of the 12 properties, we also examined the simultaneous prediction of all properties, namely, multiple outputs in the FNN, as described in Section 2.3.

2.1. Predictions Using Model A

The prediction accuracies of all the properties are shown in Table 1. The accuracy was evaluated using the coefficient of determination ($R^2$), mean absolute error (MAE), and root-mean-square error (RMSE) of the data. Selected results for the excitation energy, $sp^3$ carbon ratio, and highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital
(LUMO) (HOMO–LUMO gap) are shown in Figure 3a,b and Figure 4a, respectively, using the validation plots because of their high prediction accuracies ($R^2 > 0.93$). The validation plots for all properties are shown in Figure S1, Supporting Information, where the results are displayed in order of their accuracy; in other words, the accuracy decreases from (a) to (l).

As shown in Figure 3a, the excitation energy prediction was accurate, even though all the C K-edge spectra were aligned with their thresholds. The ELNES/XANES spectral features and excitation energies are often discussed separately. However, the current result indicates that information regarding the excitation energy is implicitly contained in the spectral features.

To understand the spectrum–property relationships, the spectral data were analyzed in detail, as shown in Figure 3c,e. First, all data were classified into ten groups according to the excitation energy, and the average spectrum for each group was obtained (Figure 3c). The average spectra of the groups with high (black) and low (yellow) excitation energies are shown in Figure 3c. (The same plots for all properties are shown in Figure S2, Supporting Information.) The spectral features around the threshold (0–2 eV) are mainly changed depending on the excitation energy, indicating that the peaks around the threshold are correlated with the excitation energy.

To clearly visualize the spectrum–property relationships, we performed sensitivity analysis of the constructed FNN model (Figure 3e). This method is schematically illustrated in Figure S3, Supporting Information. First, artificial input spectra were generated by adding an artificial peak ($\Delta x$ in Figure S3, Supporting Information).
Table 1. The 12 properties investigated, listed according to the highest prediction accuracy achieved using either Model A or Model B. The prediction accuracy was based on the coefficient of determination ($R^2$), MAEs, and RMSEs of the test data. Each property was standardized so that the mean was equal to 0 and the variance was equal to 1. $a_0$ is the Bohr radius. MAEs obtained by previous studies using chemical descriptors, SchNet,[29] DNN,[30] Multitask,[31] DTNN,[31] and MPNN,[31] are also listed.

| Properties                      | Unit | Our models | MAE | RMSE | Previous studies using dedicated chemical descriptors | MAE |
|---------------------------------|------|------------|-----|------|-------------------------------------------------------|-----|
| Excitation energy               | eV   | 0.988      | 0.0959 | 0.0575 | 0.0898                                               | 0.0829 | SchNet         | 0.0959 |
| sp$^3$ carbon ratio             | –    | 0.985      | 0.0991 | 0.0116 | 0.017                                               | 0.0291 | DNN            | 0.0991 |
| LUMO energy                     | eV   | 0.949      | 0.964  | 0.17  | 0.298                                               | 0.25  | 0.034          | 0.076  |
| HOMO–LUMO gap                   | eV   | 0.935      | 0.952  | 0.235 | 0.332                                               | 0.286 | 0.063          | 0.102  |
| Zero point vibrational energy   | eV   | 0.904      | 0.937  | 0.214 | 0.276                                               | 0.224 | 0.0017         | 0.0015 |
| HOMO energy                     | eV   | 0.765      | 0.833  | 0.219 | 0.302                                               | 0.254 | 0.041          | 0.088  |
| Isotropic polarizability        | $a_0$ | 0.753      | 0.880  | 2.86  | 4.02                                               | 2.8   | 0.235          | 0.044  |
| Internal energy                 | eV   | 0.752      | 0.933  | 363   | 549                                               | 285   | 0.014          | 92.8   |
| Heat capacity                   | cal/(mol K) | 0.701     | 0.810  | 1.68  | 2.21                                               | 1.76  | 0.033          | 0.65   |
| Dipole moment                   | D    | 0.610      | 0.610  | 0.628 | 0.901                                               | 0.901 | 0.033          | 0.583  |
| Electronic spatial extent       | $a_0$ | 0.600      | 0.726  | 106   | 148                                               | 122   | 0.073          | 35.9   |
| Molecular weight                | –    | 0.336      | 0.828  | 3.24  | 5.38                                               | 3.27  | 0.073          | 35.9   |

Information, made by a Gaussian function) overlaid onto each training spectrum at different energies; these spectra were then applied to the constructed FNN. Both positive and negative additional peaks were generated to understand the influence of either a peak increase or decrease on the predicted excitation energy. Then, the average influence of the peak deviations on the predicted property, $\Delta y$, was plotted at the corresponding energy and $\Delta x$.

Figure 3e shows the resultant deviation in the predicted excitation energy (red and blue) induced by the increase/decrease in the additional peak (vertical axis) at various energies (horizontal axis). The sensitivity analysis of all properties is shown in Figure S4, Supporting Information. Based on this figure, we can understand the increase/decrease in peak intensity and the change in property. Figure 3e shows that an intensity perturbation around the threshold (0–2 eV) results in the prediction of higher (red) and lower (blue) excitation energies. These results indicate that the spectral features around the threshold are closely related to the excitation energy and that the constructed FNN correctly learned the relationships through training.

As with the excitation energy, the prediction of the sp$^3$ carbon ratio in the investigated molecules was highly accurate ($R^2 = 0.985$ in Figure 3b). Using an approach similar to the excitation energy analysis described earlier, the average spectra of molecules with high (black) and low (yellow) sp$^3$ carbon ratios were investigated, and the results are shown in Figure 3d. In this case, the spectral data were classified into 11 groups based on the sp$^3$ carbon ratio, and their average spectra are shown. Groups of 11 were used to analyze the sp$^3$ carbon ratio because their data distribution was discrete (Figure 3b). The sensitivity analysis of the sp$^3$ carbon ratio is shown in Figure 3f. The figure clearly shows that the predicted sp$^3$ carbon ratio decreases (blue) approximately 0–2 eV and increases (red) approximately 2–5 eV with increasing sp$^3$ carbon ratio (positive side on the vertical axis).

In core-loss spectroscopy, it is well known that the first sharp peak in the C K-edge originates from $\pi^*$ orbitals, and higher energy peaks are attributed to the $\sigma^*$ orbitals. In other words, the peak at the threshold is caused by the presence of $\pi^*$ orbitals and corresponds to a decrease in the sp$^3$ carbon ratio, whereas peaks beyond the threshold are attributed to $\sigma^*$ orbitals and correspond to an increase in the sp$^3$ carbon ratio. Our FNN model successfully extracted this knowledge, which usually requires manual interpretation from the spectral dataset by experts, and successfully quantified the sp$^3$ carbon ratio in various molecules.

Our FNN model also quantitatively predicted the HOMO–LUMO gap, as shown in Figure 4a. The HOMO–LUMO gap is a molecular property correlated with both occupied and unoccupied orbitals and is directly linked to material performance, such as optical and electrical properties; thus, our model can contribute significantly to materials development. The prediction accuracy of the HOMO–LUMO gap using Model A achieved MAE = 0.235 eV (Table 1). The prediction accuracy of the dedicated chemical descriptors is compared later in this article.

Typically, the C K-edge spectral features reflect the p-type PDOS of the unoccupied orbitals of carbon at the excited state. However, as shown in Figure 3 and 4, and Figure S1, Supporting Information, we have demonstrated that the core-loss is a molecular property correlated with both occupied and unoccupied orbitals, and is directly linked to material performance, such as optical and electrical properties; thus, our model can contribute significantly to materials development. The prediction accuracy of the HOMO–LUMO gap using Model A achieved MAE = 0.235 eV (Table 1). The prediction accuracy of the dedicated chemical descriptors is compared later in this article.
Our previous study revealed that ML can successfully predict the core-loss spectral feature from the PDOS at the ground state, indicating that the core-loss spectrum has information on the ground-state electronic structure and ML has the potential to extract the information.\[41\]

The spectrum–property relationships were investigated using the same method as described earlier. As shown in Figure 4b, the intensity of the average spectra at approximately 2–5 eV appears to increase as the HOMO–LUMO gap increases (yellow → red → black); however, the peak near the threshold (0–2 eV) shows complex behavior. Although the peak intensity increases when the HOMO–LUMO gap increases from 4 to 8 eV (yellow → red), it decreases for molecules with a larger HOMO–LUMO gap (black lines). The sensitivity analysis in Figure 4c shows a non-linear response. Dark red and blue regions appear at approximately 0–3 eV, indicating that the predicted HOMO–LUMO gap is sensitive to the increase/decrease of the peak near the spectral threshold. Deviations in the predicted values exhibit complex behavior with increasing or decreasing peak intensities. Although these results are difficult to interpret, the FNN model can successfully capture the complicated spectrum–property relationships from the dataset and achieve high-accuracy predictions.

Contrary to the excitation energy, the sp\(^3\) carbon ratio, and the HOMO–LUMO gap, the R\(^2\) values for the HOMO energy, isotropic polarizability, internal energy, and heat capacity were
0.77–0.70 (Figure S1f–i, Supporting Information, respectively, and Table 1) and The worst $R^2$ values were observed for the molecular weight, which was 0.536, as shown in Figure 5a, as well as in Table 1 and Figure S1i, Supporting Information.

To understand why some properties could be predicted from the ELNES/XANES spectral features whereas others could not, the correlation coefficients among the 12 properties were obtained by dividing the covariance by the respective standard deviations of the pairs of properties, as shown in Figure S5, Supporting Information. Positive and negative values near unity correspond to strong correlations. Based on this correlation map, the properties that were predicted with high accuracy, namely, the excitation energy, sp$^3$ carbon ratio, LUMO energy, HOMO–LUMO gap, and zero-point vibration energy, showed relatively large positive correlations (0.6–0.8, dark green color). Furthermore, the properties that were predicted with high accuracy and those that were predicted with low accuracy showed relatively low correlations (thin green or thin red, respectively). The properties that were predicted with low accuracy are mainly extensive properties, such as internal energy and molecular weight, which are correlated with molecular size. Because ELNES/XANES spectra reflect the electronic structure, the use of these spectra as descriptors for predicting such extensive properties is quite challenging.

Figure 4. Analysis of the HOMO–LUMO gap prediction accuracy using Model A. a) Parity plot of the HOMO–LUMO gap showing the distribution of training and test data. b) Averaged spectra of the HOMO–LUMO gap over various value ranges. c) A type of sensitivity analysis for the HOMO–LUMO gap. Color plot showing deviations in the predicted HOMO–LUMO gap value induced by the perturbation of an artificial Gaussian peak.

Figure 5. Comparison of Model A and Model B for molecular weight predictions. Parity plots of the molecular weights predicted by a) Model A and b) Model B.
2.2. Predictions Using Model B

To improve the prediction accuracy, we focused on the relative compositions, in particular, the ratios of nitrogen, oxygen, and fluorine to carbon, because the compositional information can be easily accessed experimentally using X-ray spectroscopy or other spectroscopic methods. Furthermore, the possible values for the output can be limited by including this additional information in the model. 3D data—the ratio of nitrogen, oxygen, and fluorine atoms to carbon atoms in a molecule—were used as input data in addition to the 240D spectral input data. In total, 243D input data were used. This prediction model is referred to as "Model B." The hyperparameters in Model B were the same as those in Model A (Table S1, Supporting Information).

The molecular weight prediction obtained using Model B is shown in Figure 5b. The prediction accuracy ($R^2 = 0.536$) achieved using Model A was significantly improved ($R^2 = 0.828$) using Model B. The results for all properties obtained by Model B are shown in Table 1 (validation plots are shown in Figure S6, Supporting Information). The prediction accuracies for the properties that were predicted with higher accuracies using Model A, such as excitation energy, $sp^3$ carbon ratio, and HOMO–LUMO gap, slightly improved using Model B. Furthermore, it is noteworthy that the prediction accuracy of the HOMO–LUMO gap by Model B ($MAE = 0.205$ eV) is comparable to that of the molecular chemical descriptors, such as multitask, DTNN, and MPNN, where $MAE = 0.180–0.234$ eV $^{[31–33]}$ Although the accuracy achieved by SchNet and DNN is much higher ($0.063–0.091$ eV), $^{[29,30]}$ detailed information on chemical bonding, hybridization, and coordination of constituent atoms is necessary for these descriptors. The advantage of the descriptor proposed in this study is that it is experimentally observable. Our study proposes that an ML approach can derive electronic and structural properties from core-loss spectra, which can be measured with high sensitivity, high spatial resolution, and high temporal resolution using modern EELS and XAS instrumentation.

Furthermore, Model B significantly improved the prediction performance for several properties where Model A showed poor prediction performance. Specifically, the MAEs for the internal energy and molecular weight improved by more than 50%. A possible explanation for the improvement in the prediction accuracy of the extensive properties is that the additional molecular composition data effectively served as constraints. To confirm this, the distribution of the molecular weights of molecules used in the present study was investigated. Figure S7a,b, Supporting Information, shows the molecular weight distributions for all molecules and molecules with specific molecular compositions, respectively. Figure S7b, Supporting Information, indicates that the molecules with specific molecular compositions (red and blue histograms) have distributions that are narrower than the total distribution (Figure S7a, Supporting Information). This result indicates that the additional compositional information (i.e., the nitrogen, oxygen, and fluorine ratios to carbon) effectively works as a constraint in the data space.

We also found that the properties of molecules with neither nitrogen nor oxygen atoms could not be predicted accurately. They appear as outliers in Figure 5b because of the low accuracy of the predictions. Based on Figure S7b, Supporting Information, we found that the molecules that were predicted with low accuracy had broad molecular weight distributions (green histogram). These observations suggest that the additional molecular composition data effectively served as a constraint to improve the prediction accuracy.

2.3. Limitations and Future Perspective

Based on the prediction ability of the present method, the limitations and future perspectives are discussed in this section. First, we investigated the importance of the ELNES/XANES spectral features for prediction. Various sets of descriptors were created by varying the combination of the C K-edge spectral data, molecular composition, and molecular weight, and the prediction accuracies were confirmed. Figure S8a,b, Supporting Information, shows the prediction accuracies in terms of the MAE and RMSE, respectively. Although some properties, such as internal energy, can be predicted well without spectral data, prediction performance is considerably improved when C K-edge spectral data are used as the input. These results clearly indicate that the C K-edge spectral data contain sufficient molecular property information and that our FNN can accurately extract these properties.

In contrast to the molecular properties that were predicted accurately, the dipole moment could not be predicted well using either Model A or Model B (Figure S1j and S6j, Supporting Information, respectively, and Table 1). This poor accuracy may be due to the anisotropic character of the dipole moment, in that the value of the dipole moment depends on the direction of the axis being measured. In this research, we used isotropic spectral data, that is, the total spectra calculated along different directions. As a result, it is difficult to extract information on the dipole moment hidden in spectral features, even with the use of a neural network.

Furthermore, the present method is applicable only to relatively small molecules (fewer than nine carbon, nitrogen, oxygen, and fluorine atoms). Although the spectral features of small molecules can be correlated with the overall molecular structure, the spectrum at each carbon site collects only local structural information within a limited spatial range. In addition, average spectra were used in our predictions, and the features of the core-loss spectra reflect the electronic structure but do not represent the number of atomic sites in the molecule. Because the spectrum of a given site is derived from relatively local information around the simulated carbon site, the influence of the molecule far from the site on the spectrum is small. Similar average spectra were observed for molecules with different numbers of carbon sites but similar local environments, such as heptane and octane (Figure 2a). Therefore, it is difficult to obtain sufficient information to distinguish these large molecules from the C K-edge spectral data alone. Thus, our approach may not work well for large molecules, especially polymers or proteins, wherein the physical parameters require nonlocal and entire-molecule structural information. These tendencies, however, indicate that our model is suitable for predicting the average intensive properties, irrespective of the molecular size.

Finally, we emphasize the potential of the proposed method in terms of its versatility. The experimental spectra shown in
predicted all properties. This outcome indicates that C K-edge spectral data can act as immensely powerful descriptors for molecular property predictions. Furthermore, these results suggest that the output of the second to the last layer of the FNN possesses enough information to represent all 12 physical properties in just one fully connected layer; thus, the output is a good general descriptor for all molecular properties. The output can possibly be diverted to effectively predict other physical properties via transfer learning.

3. Conclusion

In summary, we quantified the intensive and extensive properties using core-loss spectra via an FNN model. Our neural network model using only C K-edge spectra as input data accurately predicted some intensive properties, such as excitation energy, sp³ carbon ratio, and HOMO–LUMO gap. The prediction accuracy of the HOMO–LUMO gap using the core-loss spectra as a descriptor was confirmed to be comparable with that reported in previous studies. However, we found that this prediction model did not adequately predict extensive properties, such as internal energy and molecular weight. To improve the prediction accuracy, three additional descriptors, namely, the ratios of N, O, and F to carbon, were used. Consequently, high-accuracy predictions of extensive properties, including internal energy and molecular weight, were achieved using the updated prediction model. Furthermore, it was demonstrated that the property predictions of unknown (hypothetical) molecules and predictions using experimental spectra are also available. In addition to single-output predictions, we also achieved the simultaneous prediction of all 12 properties investigated using the C K-edge spectra as descriptors.

Finally, we discuss the limitations and capabilities of the proposed method. Because the information observed using ELNES/ XANES is localized to the atomic site, sufficient information to distinguish these molecules using only C K-edge data alone is difficult. Through our successful prediction of various organic molecule properties from C K-edge spectra, our study highlights the potential of neural networks to access the abundant information hidden in core-loss spectra for the prediction of versatile physical properties. We expect that our ML approach using the core-loss spectrum as a descriptor paves the way for direct measurement of material properties with high sensitivity, high spatial resolution, and high temporal resolution and for high-throughput screening of materials through prediction of various properties, which will accelerate materials development.

4. Experimental Section

ML Models for the Prediction of Molecular Properties: The Adam optimizer\[43\] was utilized to determine the trained parameters by minimizing the mean squared error loss function. All nodes except those in the last hidden layer used the rectified linear unit (ReLU) as the activation function and a 50% dropout to prevent overfitting to the training data. Fivefold cross-validation was applied to tune the hyperparameters, such as the number of hidden layers and nodes. The hyperparameters were tuned by using a grid search. Details regarding the setup are shown in Table S1, Supporting Information.

In practical situations, molecules can undergo structural deformation according to the surrounding environment, and these structural changes induce differences in physical properties, such as chemical shifts. We also examined whether our models could be used to predict the properties of molecules with structural deformation. C K-edge spectra were calculated for isotropically strained benzene molecules with +2%, +1%, −1%, and −2% deformation (+ and − mean tensile and compressive strain, respectively). The calculated C K-edges are shown in Figure S10a, Supporting Information. Depending on the strain amount, the spectral features gradually change. Based on the spectra, the excitation energy was predicted using Model A and Model B in Figure S10b,c, Supporting Information, respectively. Because the properties of such hypothetical molecules are not present in the database\[25\] the excitation energy simulated in the present study was used for the prediction. All predicted values appeared at the center of the test data, whereas their distribution was not perfectly fitted to the orthogonal gray line of the validity plots, indicating that both Model A and Model B overestimated the excitation energy. However, the order of the predicted values, +2% → +1% → 0 → −1% → −2%, reproduced that of the actual values. This indicates that the present model can correctly capture the spectral differences of hypothetical molecules and can correctly predict the tendency of the chemical shift of the spectrum.

In the aforementioned results, the FNN output data corresponded to only one property, and a prediction model was constructed for each property individually. We attempted to predict all properties simultaneously using one prediction model, namely, multiooutput learning. Figure S11, Supporting Information, shows the prediction accuracies achieved by changing the prediction models and output data. The results of Model A and Model B for the two outputs, i.e., a single property or all 12 properties, were compared. Although single-property learning commonly achieved slightly better prediction accuracy than multiple-property learning, both learning models correctly
Preparation of the Dataset: The dataset used in this study consisted of C K-edge spectra and the properties of organic molecules; the former were the input data and the latter were the output data of the neural network. The 12 properties are shown in Table 1. To obtain a sufficiently large and slightly biased dataset for learning, we constructed the database by simulation. Structural data from 22 155 organic molecules were extracted from the dataset,[25,26] and their C K-edge spectra were calculated.

For the C K-edge simulation, we used the full-potential first-principles plane-wave pseudopotential method with the generalized gradient approximation (CGA) using the CASTEP code[44] for the C K-edge calculations. We selected this full-potential pseudopotential-based method to achieve sufficient accuracy with fast computation. The plane-wave cutoff energy was set to 500 eV. Because the C K-edge simulations required the excited state of the core electrons to be calculated, an on-the-fly pseudopotential based on CASTEP was applied to the excited carbon atoms. The theoretical excitation energy was estimated by a method reported in a previous study[49] in which both ground and excited state simulations were performed separately. In all calculations, a 15 Å cubic supercell was used to avoid interactions between the excited carbon atoms. To obtain spectra from 22 155 molecules, a total of 117 337 C K-edge simulations were performed owing to the presence of carbon atoms with different environments.

The molecular structures were used as reported in a previously reported dataset without further modification,[25] and structural relaxations were not performed. All the calculated C K-edges were broadened by a Gaussian function with a standard deviation of 0.3 eV. The C K-edges of the non-equivalent carbon atoms were subsequently aligned using the calculated transition energy and summed for each molecule to obtain the C K-edge spectrum of each molecule.

These spectra were normalized by dividing the spectral intensity by the number of carbon atoms. In addition to normalization by the number of carbon atoms, we also examined two other normalization methods: normalization using the maximum peak intensity and normalization using the total intensity. The detailed results of the various normalization methods are shown in Figure S11, Supporting Information. After obtaining the C K-edge spectrum of each molecule, all the calculated spectra were shifted such that their thresholds were 0 eV. Excitation energy information was intentionally eliminated, and only the spectral shape was used as the input for the FNN.

Some molecular properties were obtained from the database.[25] These properties were simulated using B3LYP/6-31G(2df,p)-level simulation. Although the accuracy of the respective properties has not been reported, the accuracy of the atomization energy of the database was reported to be 5 kcal mol\(^{-1}\).[25] In general, the accuracy of the simulated HOMO–LUMO gap using such the B3LYP-level approximation is known to be 0.85–1.68 eV.[46] The ratio of the number of sp\(^3\)-hybridized carbon atoms to the total number of carbon atoms in each molecule, the molecular weight, and nine other molecular properties was extracted from the molecular structures and theoretical calculations.

The dataset, including C K-edges and molecular properties, was divided into two subsets at a ratio of 8:2. The latter was used as test data, and the former was divided into training data and validation data for fivefold cross-validation. Ratios of 6:4 and 9:1 were also examined, and it was confirmed that the prediction accuracies were comparable with the prediction accuracy of the 8:2 ratio, as shown in Figure S13, Supporting Information.

Supporting Information
Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements
This study was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) (grant nos. 17H06094, 19H00818, 19H05787 and JPMJCR1993) and by a fund from the Institute of Industrial Science, University of Tokyo (Tenkai15504850104). T.M. acknowledges Prof. Chris J. Pickard at Cambridge University and Prof. Hidekazu Ikeno and Hidenobu Murata at Osaka Prefecture University for their helpful discussions on the simulation of core loss.

Conflict of Interest
The authors declare no conflict of interest.

Data Availability Statement
The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords
chemoinformatics, core-loss spectrum, descriptors, machine learning, molecular properties

Received: May 26, 2021
Revised: September 10, 2021
Published online: October 13, 2021

[1] K. Kimoto, T. Asaka, T. Nagai, M. Saito, Y. Matsu, K. Ishizuka, Nature 2007, 450, 702.
[2] F. S. Hage, G. Radtke, D. M. Kepaptsoglou, M. Lazzeri, Q. M. Rammase, Science 2020, 367, 1124.
[3] O. L. Krivanek, M. F. Chisholm, V. Nicolosi, T. J. Pennycook, G. J. Corbin, N. Deliby, M. F. Murffit, C. S. Own, Z. S. Szilagyi, M. P. Oxley, S. T. Pantelides, S. J. Pennycook, Nature 2010, 464, 571.
[4] S. C. Lin, C. C. Chang, S. Y. Chiu, H. T. Pai, T. Y. Liao, C. S. Liao, W. H. Chiang, M. K. Tsai, H. M. Chen, Nat. Commun. 2020, 11, 1.
[5] C. Bressler, M. Chergui, Chem. Rev. 2004, 104, 1781.
[6] I. Tanaka, T. Mizuguchi, M. Matsu, S. Yoshikoa, H. Adachi, T. Yamamoto, T. Okajima, M. Urnesaki, W. Y. Ching, Y. Inoue, M. Mizuno, H. Araki, Y. Shirai, Nat. Mater. 2003, 2, 541.
[7] R. F. Egerton, Electron Energy-Loss Spectroscopy in the Electron Microscope, Springer, Boston, MA 2011.
[8] J. M. Thomas, G. Sankar, Acc. Chem. Res. 2001, 34, 571.
[9] J. Stohr, NEXAFS Spectroscopy, Springer, Berlin, NY 1992.
[10] Machine Learning in Chemistry, (Ed: H. M. Cartwright), Royal Society of Chemistry, Cambridge 2020.
[11] Machine Learning Meets Quantum Physics, (Eds: K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, K.-R. Müller), Springer International Publishing, Cham 2020.
[12] S. Kiyohara, H. Oda, T. Miyata, T. Mizuguchi, Sci. Adv. 2016, 2, 1600746.
[13] N. C. Frey, M. K. Horton, J. M. Munro, S. M. Griffin, K. A. Persson, V. B. Shenoy, Sci. Adv. 2020, 6, 1.
[14] L. Li, Y. Yang, D. Zhang, Z. G. Ye, S. Jesse, S. V. Kalinin, R. K. Vasudevan, Sci. Adv. 2018, 4, eaaap8672.
[15] C. Carey, T. Boucher, S. Mahadevan, P. Bartholomew, M. D. Dyer, J. Raman Spectros. 2015, 46, 894.
[16] J. Curry, Y. Xie, C. J. Pickard, A. A. Hassanali, J. Chem. Theory Comput. 2016, 12, 765.
[17] K. Varmuza, P. N. Penchev, H. Scsibrany, J. Chem. Inf. Model. Sci. 1998, 38, 420.
[18] S. Kiyohara, M. Tsubaki, K. Liao, T. Mizuguchi, J. Phys. Mater. 2019, 2, 024003.
[19] I. Miyazato, L. Takahashi, K. Takahashi, Mol. Syst. Des. Eng. 2019, 4, 1014.
[20] S. Kiyohara, T. Mizoguchi, J. Phys. Soc. Jpn. 2020, 89, 103001.
[21] J. Timoshenko, D. Lu, Y. Lin, A. I. Frenkel, J. Phys. Chem. Lett. 2017, 8, 5091.
[22] S. Kiyohara, T. Miyata, K. Tsuda, T. Mizoguchi, Sci. Rep. 2018, 8, 13548.
[23] M. R. Carbone, S. Yoo, M. Topsakal, D. Lu, Phys. Rev. Mater. 2019, 3, 33604.
[24] R. Ramakrishnan, M. Hartmann, E. Tapavicza, O. A. Von Lilienfeld, J. Chem. Phys. 2015, 143, 084111.
[25] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, Sci. Data 2014, 1, 140022.
[26] L. Ruddigkeit, R. van Deursen, L. C. Blum, J.-L. Reymond, J. Chem. Inf. Model. 2012, 52, 2864.
[27] L. C. Blum, J. L. Reymond, J. Am. Chem. Soc. 2009, 131, 8732.
[28] T. Fink, H. Bruggesser, J. L. Reymond, Angew. Chem. Int. Ed. 2005, 44, 1504.
[29] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, J. Chem. Phys. 2018, 148, 241722.
[30] M. Tsubaki, T. Mizoguchi, J. Phys. Chem. Lett. 2018, 9, 5733.
[31] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, V. Pande, Chem. Sci. 2018, 9, 513.
[32] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, Nat. Commun. 2017, 8, 13890.
[33] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, in Proc. 34th Int. Conf. Mach. Learn. (Eds.: D. Precup, Y.W. Teh), PMLR, Inter. Convention Centre, Sydney, Australia, 2017, pp. 1263–1272.
[34] A. P. Hitchcock, I. G. Eustatius, J. T. Francis, C. C. Turci, J. Electron. Spectrosc. Relat. Phenom. 1998, 88–91, 77.
[35] A. P. Hitchcock, P. Fischer, A. Gedanken, M. B. Robin, J. Phys. Chem. 1987, 91, 531.
[36] E. Apen, A. P. Hitchcock, J. L. Gland, J. Phys. Chem. 1993, 97, 6859.
[37] G. Onida, L. Reining, A. Rubio, Rev. Mod. Phys. 2002, 74, 601.
[38] I. Ishii, R. McLaren, A. P. Hitchcock, M. B. Robin, J. Chem. Phys. 1987, 87, 4344.
[39] A. P. Hitchcock, C. E. Brion, Chem. Phys. 1979, 37, 319.
[40] H. Shu, H. Zhu, Proc. AAAI Conf. Artif. Intell. 2019, 33, 4943.
[41] S. Kiyohara, M. Tsubaki, T. Mizoguchi, npj Comput. Mater. 2020, 6, 1.
[42] T. Mizoguchi, S. Kiyohara, Microscopy 2020, 69, 92.
[43] D. Kingma, J. Ba, arXiv:1412.6980 [cs.LG] 2014, 1.
[44] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. J. Probert, K. Refson, M. C. Payne, Z. Kristallogr. 2005, 220, 567.
[45] T. Mizoguchi, I. Tanaka, S.-P. Gao, C. J. Pickard, J. Phys. Condens. Matter 2009, 21, 104204.
[46] G. Zhang, C. B. Musgrave, J. Phys. Chem. A 2007, 111, 1554.