ARTICLE OPEN

Automated estimation of materials parameter from X-ray absorption and electron energy-loss spectra with similarity measures

Yuta Suzuki^{1,2}, Hideitsu Hino³, Masato Kotsugi^{1,4} and Kanta Ono^{2,4}

Materials informatics has significantly accelerated the discovery and analysis of materials in the past decade. One of the key contributors to accelerated materials discovery is the use of on-the-fly data analysis with high-throughput experiments, which has given rise to the need for accelerated and accurate automated estimation of the properties of materials. In this regard, spectroscopic data are widely used for materials discovery because these data include essential information about materials. An important requirement for the realisation of the automated estimation of materials parameters is the selection of a similarity measure, or kernel function. The required measure should be robust in terms of peak shifting, peak broadening, and noise. However, the determination of appropriate similarity measures for spectra and the automated estimation of materials parameters from these spectra currently remain unresolved. We examined major similarity measures to evaluate the similarity of both X-ray absorption and electron energy-loss spectra. The similarity measures show good correspondence with the materials parameter, that is, the crystal-field parameter, in all measures. The Pearson's correlation coefficient was the highest for the robustness against noise and peak broadening. We obtained the regression model for the crystal-field parameter 10 Dq from the similarity of the spectra. The regression model enabled the materials parameter, that is, 10 Dq, to be automatically estimated from the spectra. With regard to research progress in similarity measures, this methodology would make it possible to extract the materials parameter from a large-scale dataset of experimental data.

npj Computational Materials (2019)5:39; <https://doi.org/10.1038/s41524-019-0176-1>

INTRODUCTION

Recent years have seen a considerable improvement in the throughput of the fabrication and characterisation of materials.^{1,2} Using a multi-target sputtering technique, for example, has enabled the methodology of fabricating a sample containing all phases of an alloy to be established.³ In the case of X-ray diffraction in synchrotron radiation facilities, 5000 samples can be measured per day.⁴ However, despite the recognition that the acceleration of the fabrication and characterisation of materials is an important problem that has attracted considerable attention, measured data often continues to be analysed via the old manual way. Because data analysis using this conventional approach could take several days to several months, it may become a bottleneck in the process. The objective of efficient materials research with materials informatics is to eliminate the bottleneck, and to accelerate the research flow consisting of the fabrication and characterisation of materials followed by data analysis.^{5–7}

It is thus important to establish a methodology that automatically and quantitatively extracts the materials parameter from the measured data.^{8,9} This technique allows the on-the-fly data analysis to be completed as part of the online characterisation in that it provides a combined procedure ranging from material fabrication to material discovery, thereby eliminating the

bottleneck. The efficiency of the investigation of a material can be expected to drastically improve if the entire procedure flows smoothly. Therefore, automating and enhancing the speed of data analysis for high-throughput materials research has become increasingly important for the discovery of innovative materials.^{10,11}

Spectroscopy is widely employed to evaluate the properties of materials. An example of a spectroscopic method is X-ray absorption spectroscopy (XAS) and electron energy-loss spectra (EELS), which provide information about the electronic and chemical state of a specific atom. The crystal field parameter 10 Dq is one of the most significant material parameters that can be gained from XAS and EELS. It represents the energy splitting originating from the crystal field and provides an important hint relating to material properties such as magnetism and optical properties.

It is possible to calculate the XAS or EELS spectrum of a 3d transition metal if the value of the materials parameter is given. The calculation for XAS and EELS spectra is usually performed based on atomic multiplet calculations with crystal field multiplet and charge transfer multiplet calculations¹² or first-principles calculations.^{13–15} The spectral shape very complicated and the estimation of materials parameter (crystal field parameter) directly from the spectrum should be ill-posed inverse problem and

¹Department of Materials Science and Technology, Tokyo University of Science, Tokyo, Japan; ²High Energy Accelerator Research Organization, Institute of Materials Structure Science, Tsukuba, Japan; ³The Institute of Statistical Mathematics, Tokyo, Japan and ⁴Center for Materials Research by Information Integration (CMII), Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science (NIMS), Tsukuba 305-0047, Japan
Correspondence: Kanta Ono (kanta.ono@kek.jp)

Received: 5 October 2018 Accepted: 7 March 2019

Published online: 29 March 2019

Fig. 1 Statistical learning methodology. First, the spectral dataset is prepared by simulation with various parameters. The similarity values between the reference spectrum and each of the spectra are calculated. This enables the statistical learning model for estimating the materials parameters from the spectra to be constructed. Continuous values are estimated by utilising regression models, whereas a discrete value or category is estimated by employing dimensionality reduction. In this procedure, the physical parameters of each spectrum are paired with a similarity value. Finally, the physical parameters of the measured spectra are estimated with the models

mathematically intractable. Previously, the typical method to evaluate a physical value from a spectrum consisted of visually comparing the measured spectrum with the calculated spectrum.^{16,17}

This suggests that if the similarity measures for spectrum comparison could be established, the task of spectrum analysis could be automated and the materials parameters extracted automatically. In addition, statistical machine-learning methods (e.g. clustering¹⁸ and matrix factorisation¹⁹) could be applied to materials research, whereupon further improvement in the research efficiency could be expected.

In materials informatics, the methodology to analyse the big data obtained from high-throughput experiments and simulations is extremely important when attempting to utilise machine learning.^{8,20–25} Unsupervised learning methods such as clustering extract information based on the relationships among input data; thus, the results may greatly depend on the similarity measure that is used.^{26,27} The selection of appropriate measures is one of the most important steps when applying unsupervised learning methods to evaluate and analyse materials.^{26–29}

The estimation of materials parameters from experimental data via similarity measures has great potential for automated data analysis in materials research. The most important aspect of automated materials parameter estimation is the choice of a similarity measure (kernel functions,³⁰) which is not trivial and varies with the experimental method. The similarity measures should be selected specifically for each combination of materials parameters (e.g. 10 Dq, lattice parameters), measurement method (e.g. XAS, XRD), and material (e.g. transition metal, metal oxide). Data obtained from high-throughput characterisation often include imperfections such as added noise and deteriorating resolution; hence, the similarity measures for these data should be robust against these imperfections. We suggest that good similarity measures should meet the following two requirements: 1. The measure can accurately estimate the materials parameter of interest. 2. The metric is robust against the imperfections of measurement such as noise addition and resolution deterioration.

We depicted the workflow for the estimation of materials parameters from spectra in Fig. 1. In the first step, the spectral dataset for building the statistical models is prepared by simulation or from a large set of experimental data. Then, the spectra are mapped into kernel (similarity) space, using appropriate similarity measures. The similarity between the measured spectrum and the standard spectrum is calculated to estimate the materials parameter from measured spectra. The discrete materials parameters (i.e. the charge of an atom) are estimated with dimensionality reduction (unsupervised learning) and human decision. If needed, this step can be replaced by classification (supervised learning), which does not need a human decision. The continuous materials parameters (i.e. the crystal field parameter) of measured spectra are estimated using the regression model (supervised learning). In the following section, we describe the results of each step in this workflow.

We compared the similarity measures for XAS spectra to determine whether data from high-throughput measurement could be analysed promptly. In this respect, the Euclidean distance (ED) (L2 norm) and Manhattan distance (L1 norm) are widely used in many fields as similarity/distance metrics; however, these metrics may perform poorly as similarity measures between measured data,^{26,27,29} and the appropriate measure of similarity is not trivial.^{26,29,31}

We investigated measures that are robust to noise and peak broadening and are sensitive to changes in the material parameters. We demonstrate that an important material property, such as the crystal field parameter 10Dq, can be estimated automatically and promptly by the constructed regression model based on the similarity measure.

RESULTS

Similarity measures

The spectra of interest are the Mn^{2+} $L_{2,3}$ XAS or EELS spectra of MnO obtained from both calculations and experiments. The similarity measures can be defined by using various distance metrics. A distance is a metric that represents how far apart

objects are. When the distance between vector x and y is written as $d(x, y)$, d is known as the distance function, and the following conditions are satisfied:³²

$$d(x, y) \geq 0, \quad (1)$$

$$x = y \Rightarrow d(x, y) = 0, \quad (2)$$

$$d(x, y) = d(y, x), \quad (3)$$

$$d(x, y) + d(y, z) \geq d(x, z). \quad (4)$$

The similarity s and the distance d are related as $s = 1 - d$, when d is normalised in the range $[0, 1]$. In general, distances are normalised by using their value range; in this work, normalisation was achieved by using the maximum distance estimated by the physical constraints. The crystal field parameter $10Dq$ is the difference between the energy levels originating from the breaking of degeneracies of electron orbital states. The maximum value of $10Dq$ can be extracted from physical properties such as the atomic number and the crystal structure. Including the physical constraints, the value of $10Dq$ can be normalised and included as a metric of which the value is not limited by a maximum AND/OR minimum.

This study evaluated the following distance functions: the ED, city block distance (CD), cosine, Jensen–Shannon divergence (JSD), Pearson's correlation coefficient (PCC), dynamic time warping (DTW), and earth mover's distance (EMD). DTW and EMD require a base measure, and the Manhattan distance was employed in this work.

Let x and y be n -dimensional vectors represented by $x = (x_1, x_2, \dots, x_n)$. The definitions of the metrics that were used are as follows.

The ED and CD are special cases of the Minkowski distance ($p = 2, 1$):

$$d_{\text{Minkowski}}(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}. \quad (5)$$

The cosine metric represents the cosine of the angle of vector x and y in n -dimensional space; it is constant against changes in the length of the vectors because the cosine metric is robust to intensity changes of the whole spectrum:

$$d_{\text{Cosine}}(x, y) = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}. \quad (6)$$

Pearson's product-moment correlation coefficient (PCC) is similar to the cosine metric; it is the cosine between the vector x and y , and their means:

$$d_{\text{PCC}}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (7)$$

JSD is one of the metrics representing the distance between probability distributions, and it is a modification of the Kullback–Leibler divergence (KLD) to satisfy the symmetry rule.³² KLD is a metric of the extent to which one probability distribution diverges from another and is known as the relative entropy:

$$d_{\text{JSD}}(x, y) = \frac{1}{2} D_{\text{KL}}(x, M) + \frac{1}{2} D_{\text{KL}}(y, M), \quad (8)$$

$$D_{\text{KL}}(x, y) = \sum_{i=1}^n x_i \log \frac{x_i}{y_i}, \quad (9)$$

$$M = \frac{1}{2} (x_i + y_i). \quad (10)$$

Here we assume that vectors x and y are normalised to be non-negative and that the summation of elements is one.

DTW, which makes it possible to compare the similarity of distributions that do not have the same length, is utilised especially in voice recognition. The DTW of vector or time-series data x and y is calculated according to the following procedure: first, set the window size; stretch the length of x in each window to minimise the distance to y . The summation of these distances is the DTW of x and y .³³

EMD is a metric related with the optimisation problem of transportation. The Minkowski distance and KLD are bin-to-bin distances and compare the same bins of histograms for which the similarity values decrease with a slight shift of the histograms. EMD is the cross-bin distance, such that it is robust to the shift of the whole histogram.^{32,34,35} Both DTW and EMD are not exact distances, because they do not satisfy the triangular inequality,³⁶ instead, they are designed to preserve a specific characteristic.

Dimensionality reduction and visualisation

Before estimation of materials parameter ($10Dq$) from the spectra, it is important to determine the element and valence of the material. Elements can be differentiated from one another using the photon energy of the location for the peaks in absorption spectra

In many cases, the intrinsic dimension of high-dimensional data is low, and the data is distributed in low dimension manifolds.^{37,38} Based on that idea, we attempted to reduce the dimension of the spectrum by manifold learning and visualise it. Multi-dimensional scaling (MDS) is one of the simplest dimensionality reduction algorithm and is possible to represent high-dimensional data in a low-dimensional space by approximating the distance in the original space.³⁹

In general, there are several intrinsic dimension estimation methods to estimate the optimal number of dimensions,^{40–42} although we do not put emphasis on it in this work. The spectra of Mn, with various valences, and experimentally obtained spectrum of MnO were calculated and represented in two dimensions by MDS. The results are shown in Fig. 2. The numbers in the figure represent the value of the crystal field parameter (eV) multiplied by 10. The valence of Mn was set as 2, 3, and 4+ with the

Fig. 2 Dimensionality reduction and visualisation results for Mn^{2,3,4+} X-ray absorption spectroscopy (XAS) spectra with multi-dimensional scaling (MDS). The numbers in this figure correspond to the value of the crystal field parameter (eV) multiplied by 10. The inset indicates the experimentally obtained XAS spectrum for MnO.¹⁶ The red dots corresponds to the experimentally obtained MnO XAS spectrum

Fig. 3 Similarity as a function of the crystal field parameter 10Dq. The similarity of spectra is defined as the similarity between the target spectra and the reference spectrum, in this case, the simulated spectrum with a 10Dq value of zero

symmetry as Oh. The ED was used as the distance metric for the sake of simplicity. As can be seen from Fig. 2, spectra with different valences are distinctly separated in the data space, and the distance between the spectra and the value of 10Dq correspond.

The automated data analysis for XAS/EELS spectra using dimensionality reduction is validated with the experimentally obtained Mn XAS spectrum of MnO that corresponds to Mn^{2+} and 10Dq = 0.9 eV.¹⁶ The MnO XAS spectra and correspondent dimensionality reduction results (red dot) are plotted in the figure. These spectral data approximated those of Mn^{2+} and 10Dq = 0.9 eV closely. This suggests that the estimation of the physical quantity (i.e. the charge, 10Dq) could be realised by evaluating the distance between the spectra.

Comparison of the similarity measure

We adopt the simplest measure as the similarity measure of the XAS and EELS spectra, although several methods exist according to which to define the similarity measure. We define the similarity of spectra as the similarity between the target spectra and the standard spectrum, in this case, the simulated spectrum with a 10Dq value of zero. The spectra of interest are the Mn^{2+} $L_{2,3}$ XAS or EELS spectra of MnO. We compare the behaviour of each of the similarity measures as a function of the materials parameter 10Dq. Figure 3 shows the similarity of MnO 2p XAS as a function of 10Dq. The similarities are calculated between the simulated spectra by varying the value of 10Dq and the standard spectrum simulated with a 10Dq value of zero. All of the measures except DTW were found to show a one-to-one relationship between the similarity and the materials parameter. As seen in Fig. 3, PCC, cosine, and JSD were insensitive to 10Dq at <1.0 eV. If the estimated 10Dq value is in the insensitive range, coupling another measure could be expected to produce a good result.

Estimation of the materials parameter

We built a regression model to estimate the value of the materials parameter 10Dq from the similarity of the spectra. The trend, according to which the similarity changes, is not trivial against the change in the materials parameter, and we build a regression model from the similarity measure vs. the materials parameter data. A proper regression model is built for each similarity measure with the polynomial function where the degree of the polynomial function is estimated from the Akaike information criterion (AIC).⁴³ The performance of the regression model is sufficient for the estimation of 10Dq from the similarity.

Fig. 4 Comparison of the experimentally obtained MnO X-ray absorption spectroscopy (XAS) spectrum, and the Mn^{2+} XAS spectrum calculated with the estimated value of 10Dq (0.9 eV) from the regression model for Pearson's correlation coefficient (PCC)

The performance of the regression model for experimental data was validated by the experimentally obtained 2p XAS spectrum of MnO.¹⁶ The spectrum of Mn^{2+} reconstructed from the estimated value of 10Dq of 0.9 eV with PCC similarity and the experimentally obtained MnO XAS spectrum are shown in Fig. 4. The figure shows that the spectrum predicted from the similarity measure of PCC corresponds well to the experimentally obtained spectrum. According to the literature,¹⁶ the value of 10Dq estimated by human visual inspection is 0.9 eV, which corresponds well with the estimation from the regression model for PCC.

We compare the performance of the similarity measures on the estimation of the 10Dq value. The DTW measure was not used since the similarity was not determined uniquely from 10Dq. All the similarity measures could estimate the value of 10Dq at ~1.0 eV. Especially, PCC and cosine could correctly estimate the value of 10Dq as 0.9 eV. The calculation time for the estimation is several milliseconds on a general laptop computer and we were able to estimate the materials parameters from more than 10,000 spectra taken by scanning transmission X-ray microscopy in a reasonably short time.

Therefore, it was demonstrated that the crystal field parameter 10Dq can be estimated automatically and promptly by using the appropriate measures.

It should be noted that the appropriate similarity measure could be automatically optimised by distance metric learning, which has been studied recently, and may also contribute to improve the insensitivity.^{44–46} We are currently in the process of the automated determination of appropriate similarity measures for a variety of measurement data from other materials characterisation techniques.

Robustness against noise

In high-throughput measurements, the influence of noise is the most significant factor owing to the short measurement time. Thus, similarity that is robust against noise is indispensable for these measurements.

We hence examined whether the similarity measures are robust against noise. We modelled the noise in the XAS or EELS spectroscopy as Gaussian noise. The noise with the varied variance in the Gaussian distribution was added to the calculated 2p XAS of Mn^{2+} . The similarity with and without Gaussian noise is shown in Fig. 5.

The signal-to-noise (S/N) ratio in Fig. 5 is defined as the ratio between the peak height of the true spectrum and the standard deviation of the noise. Obviously, PCC showed excellent

Fig. 5 Result of robustness against the addition of noise. The noise in the X-ray absorption spectroscopy (XAS) or electron energy-loss spectra (EELS) spectroscopy is modelled as Gaussian noise. The signal-to-noise (S/N) ratio is defined as the ratio between the peak height of the true spectrum and the standard deviation of the noise. **a** Similarity of the spectra with Gaussian noise. Pearson's correlation coefficient (PCC) showed excellent robustness against the addition of noise. **b** Simulated Mn^{2+} XAS spectra with Gaussian noise

robustness against the addition of noise. The results of ED and CD showed the same behaviour.

Using PCC, the similarity of the noisy spectrum with an S/N ratio of 30 was calculated at almost 1.0, whereas it was calculated at below 0.9 with the other measures. Particularly, the result with both ED and CD shows poor robustness against noise, despite the fact that these are commonly used measures. This result suggests that the measurement time can be significantly reduced if an appropriate similarity metric such as PCC is selected.

Robustness against peak broadening

In practical spectroscopy measurements, the energy resolution of the spectroscopy system is one of the most important specifications of the measurement system. The ability to estimate the material parameters with equipment with poor energy resolution may lead to a significant reduction in the cost of an experiment. In this work we established an appropriate similarity measure that is robust against deteriorated energy resolution. We calculated the convolution of XAS spectra and the Gaussian function with varied width. The similarity of the spectra as a function of the width of the Gaussian broadening is shown in Fig. 6. The standard deviation of the Gaussian function, σ was varied in the range from 0.02 to 0.21 eV, and compared to the spectrum with $\sigma = 0.02$ eV, which represents the energy resolution of the measurement system. A good measure requires robustness to peak broadening such that it can be applied to a low-resolution

Fig. 6 Results of robustness against peak broadening. Similarities are plotted as a function of peak broadening (σ). The standard deviation of the Gaussian function, σ , was varied in the range from 0.02 to 0.21 eV, and compared to the spectrum with $\sigma = 0.02$ eV, which represents the energy resolution of the measurement system. It is shown that Pearson's correlation coefficient (PCC), Jensen–Shannon divergence (JSD), and cosine are more robust, whereas Euclidean distance (ED), city block distance (CD), and dynamic time warping (DTW) have poor robustness to broadening

measurement. As shown in the Fig. 6, PCC, JSD, and cosine are more robust, whereas ED, CD, and DTW have poor robustness to broadening.

This result suggests the importance of choosing an appropriate measure that enables the estimation of a materials parameter even from measurement systems with poor energy resolution.

DISCUSSION

PCC shows the best performance among the similarity measures for the estimation of 10 Dq, robustness against spectral broadening, and noise. It should be noted that robustness against noise is a very important property required for a similarity measure.

PCC is considered to be the cosine similarity between the averaged vector and the data vector, and it should be robust against fluctuations in the baseline of the spectra caused by noise. In the case of noisy spectra, the real signal components become smaller when the spectra are normalised. The cosine similarity calculates the angle between vectors and the length of the vectors does not affect the cosine similarity. From this point of view, both PCC and cosine similarity should be robust against noise.

We focused on the estimation of materials parameter from Mn XAS/EELS spectra in this study; however, the proposed approach has an extensibility for a wide range of XAS/EELS spectra. Recently, there is a large open database with 500,000 K-edge X-ray absorption near-edge spectra for more than 40,000 unique materials.⁴⁷ In the next step, we will combine our approach on dimensionality reduction and appropriate similarity measure for XAS with the large XAS spectra dataset to realise automated knowledge discovery from measured XAS/EELS data with high-throughput experiments.

It should be noted that there is no generalisable approach to choose the appropriate measure for an unknown materials parameter at that moment. We think the approach proposed in this study can be automated and applicable for choosing the appropriate measure even for an unknown materials parameter. There is another approach called distance metric learning or similarity learning that we can construct new similarity function or distance metric for an unknown materials parameter with learning from experimental or simulated datasets.

In many cases of high-throughput experiments, the most important information that can be obtained by analysing the acquired data, rather than by acquiring the data, is the material

parameters (e.g. the electronic structure and the lattice parameter). The measurement time should be minimised to the necessary and sufficient conditions to enable the desired parameters to be extracted. For this purpose, it is necessary to coordinate the experimental measurements with the data analysis; however, the development thereof is still in progress.^{8,22} This study led us to identify those measures that are robust to noise and a deterioration of resolution, and that are intended for high-throughput measurement. This result is the basis for the technique that makes it possible to perform on-the-fly extraction of a material parameter from within the measurement. In future, our result is expected to contribute to the realisation of true high-throughput materials discovery, which integrates high-throughput fabrication, characterisation and on-the-fly data analysis.

It is important to point out that we can reduce the measurement point for an experiment with the use of similarity measures and this technique significantly accelerates the characterisation of materials and the automated extraction of material properties, both of which are essential for materials informatics. Now we are working on this problem, and the notable progress was obtained.^{5,48}

METHOD

Simulation of XAS/EELS spectra

We used CTM4XAS for the simulation of XAS/EELS spectra.¹² The dataset for the XAS spectra was prepared by calculating the Mn 2p XAS spectra with Mn^{2,3,4+} configuration by changing the 10 Dq value from 0 to 2.5 eV. Since the spectra of interest are the L_{2,3} XAS or EELS spectra of MnO, we set the symmetry of the crystal field to be octahedral (*O_h*) and tetrahedral (*T_d*), and the other parameters used in this study are identical to those published before.¹⁷ The dataset used in this study is available at <https://doi.org/10.5281/zenodo.2532856>.

Dimensionality reduction

There are numbers of dimensionality reduction or manifold learning algorithms such as Isomap, Locally linear embedding, Laplacian eigenmaps, and *t*-distributed stochastic neighbour embedding.³⁷ Among them, MDS is the simplest algorithm. In order to clarify the validity of dimensionality reduction upon spectroscopy data, we employed the simplest algorithm. The dimensionality reduction was performed with the scikit-learn package for Python.⁴⁹ The code for dimensionality reduction is found in the repository.

Regression model

The similarity or distance between spectra is calculated using R with the package proxy, transport and dtw.⁵⁰ The regression model is built for each similarity measure with the polynomial function where the degree of the polynomial function is estimated from AIC.⁵¹ The training data for regression is the similarity of Mn²⁺ 2p XAS spectra with 10 Dq value from 0 to 2.5 eV in 0.1 eV step. To validate the regression model, we use the experimentally obtained XAS spectrum for MnO, which was scanned from the literature.¹⁶ The best regression model selection is performed by MuMin package in R using automated information-theoretic model selection with AIC. The code for regression model is found in the repository.

DATA AVAILABILITY

The datasets and codes that support the findings of this study can be found at <https://doi.org/10.5281/zenodo.2532856>.

ACKNOWLEDGEMENTS

This work is partly supported by the Elements Strategy Initiative Centre for Magnetic Materials (ESICMM) under the outsourcing project of the Ministry of Education, Culture, Sports, Science, Technology (MEXT). This work is partly supported in part by 'Materials Research by Information Integration' Initiative (MI2I) project of the Support Program for Starting Up Innovation Hub from Japan Science and Technology Agency (JST). H.H. is partly supported by JST CREST grant number JPMJCR1761. Y.S. is

supported by JST, ACT-I, grant Number JPMJPR18UE. K.O. gratefully acknowledges the financial support by Toyota Motor Corporation.

AUTHOR CONTRIBUTIONS

K.O. conceived the idea for the present work. Y.S., K.O. and H.H. carried out the computation. Y.S., K.O. and H.H. wrote the manuscript together. All authors discussed the results.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Computational Materials* website (<https://doi.org/10.1038/s41524-019-0176-1>).

Competing interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

- Lookman, T., Alexander, F. J. & Rajan, K. *Information Science for Materials Discovery and Design*. (Springer International Publishing, Switzerland, 2015).
- Potyrailo, R. et al. Combinatorial and high-throughput screening of materials libraries: review of state of the art. *ACS Comb. Sci.* **13**, 579–633 (2011).
- Koinuma, H. & Takeuchi, I. Combinatorial solid-state chemistry of inorganic materials. *Nat. Mater.* **3**, 429–438 (2004).
- Gregoire, J. M. et al. High-throughput synchrotron X-ray diffraction for combinatorial phase mapping. *J. Synchrotron Rad.* **21**, 1262–1268 (2014).
- Ueno, T. et al. Adaptive design of an X-ray magnetic circular dichroism spectroscopy experiment with Gaussian process modelling. *npj Comput. Mater.* **4**, 4 (2018).
- Green, M. L. et al. Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *Appl. Phys. Rev.* **4**, 011105–18 (2017).
- Hill, J. et al. Materials science with large-scale data and informatics: unlocking new opportunities. *MRS Bull.* **41**, 399–409 (2016).
- Kusne, A. G. et al. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Sci. Rep.* **4**, 191–7 (2014).
- Suram, S. K. et al. Automated phase mapping with AgileFD and its application to light absorber discovery in the V–Mn–Nb oxide system. *ACS Comb. Sci.* **19**, 37–46 (2017).
- Ren, F. et al. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.* **4**, eaq1566 (2018).
- Alberi, K. et al. The 2019 materials by design roadmap. *J. Phys. D* **52**, 013001 (2018).
- Stavitski, E. & de Groot, F. M. F. The CTM4XAS program for EELS and XAS spectral shape analysis of transition metal L edges. *Micron* **41**, 687–694 (2010).
- Shirley, E. L. Ab. Initio Inclusion of electron-hole attraction: application to X-ray absorption and resonant inelastic X-Ray scattering. *Phys. Rev. Lett.* **80**, 794–797 (1998).
- Vinson, J., Rehr, J. J., Kas, J. J. & Shirley, E. L. Bethe-Salpeter equation calculations of core excitation spectra. *Phys. Rev. B* **83**, 115106 (2011).
- Liang, Y. et al. Accurate X-ray spectral predictions: an advanced self-consistent-field approach inspired by many-body perturbation theory. *Phys. Rev. Lett.* **118**, 096402–7 (2017).
- de Groot, F. & Kotani, A. *Core Level Spectroscopy of Solids* (CRC, Boca Raton, 2008).
- de Groot, F. M. F., Fuggle, J. C., Thole, B. T. & Sawatzky, G. A. 2p x-ray absorption of 3d transition-metal compounds: an atomic multiplet description including the crystal field. *Phys. Rev. B* **42**, 5459–5468 (1990).
- Jain, A. K., Murty, M. N. & Flynn, P. J. Data clustering: a review. *ACM Comput. Surv.* **31**, 264–323 (1999).
- Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
- Meredig, B. et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89**, 83–7 (2014).
- Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
- Agrawal, A. & Choudhary, A. Perspective: materials informatics and big data: realization of the “fourth paradigm” of science in materials science. *APL Mater.* **4**, 053208 (2016).
- Zheng, C. et al. Automated generation and ensemble-learned matching of X-ray absorption spectra. *npj Comput. Mater.* **4**, 12 (2018).

24. Kiyohara, S., Miyata, T., Tsuda, K. & Mizoguchi, T. Data-driven approach for the prediction and interpretation of core-electron loss spectroscopy. *Sci. Rep.* **8**, 13548 (2018).
25. Suzuki, Y. et al. Extraction of physical parameters from X-ray spectromicroscopy data using machine learning. *Microsc. Microanal.* **24**, 478–479 (2018).
26. Iwasaki, Y., Kusne, A. G. & Takeuchi, I. Comparison of dissimilarity measures for cluster analysis of X-ray diffraction data from combinatorial libraries. *NPJ Comput. Mater.* **3**, 1–8 (2017).
27. Lerotic, M. et al. Cluster analysis in soft X-ray spectromicroscopy: Finding the patterns in complex specimens. *J. Electron Spectrosc. Relat. Phenom.* **144–147**, 1137–1143 (2005).
28. Shirkhorshidi, A. S., Aghabozorgi, S. & Wah, T. Y. A comparison study on similarity and dissimilarity measures in clustering continuous data. *PLoS ONE* **10**, e0144059–20 (2015).
29. Hernández-Rivera, E., Coleman, S. P. & Tschoop, M. A. Using similarity metrics to quantify differences in high-throughput data sets: application to X-ray diffraction patterns. *ACS Comb. Sci.* **19**, 25–36 (2017).
30. Schölkopf, B. & Smola, A. J. *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond* (MIT, Cambridge, 2001).
31. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. & Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **3**, 419–6 (2013).
32. Deza, M. M. & Deza, E. *Encyclopedia of Distances* (Springer, Berlin, Heidelberg, 2016).
33. Keogh, E. & Ratanamahatana, C. A. Exact indexing of dynamic time warping. *Knowl. Inf. Syst.* **7**, 358–386 (2005).
34. Rubner, Y., Tomasi, C. & Guibas, L. J. in *Sixth International Conference on Computer Vision*, 59–66 (IEEE, Bombay, India, 1998). <https://doi.org/10.1109/iccv.1998.710701>.
35. Rubner, Y., Tomasi, C. & Guibas, L. J. The Earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vision.* **40**, 99–121 (2000).
36. Berndt, D. J. & Clifford, J. Using dynamic time warping to find patterns in time series. In *AAAI-94 workshop on knowledge discovery in databases*, 359–370, Usama M. Fayyad and Ramasamy Uthurusamy Eds. (The AAAI Press, Menlo Park, California, 1994).
37. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, New York, 2006).
38. Ma, Y. & Fu, Y. *Manifold Learning Theory and Applications* (CRC, Boca Raton, 2011).
39. Borg, I. & Groenen, P. *Modern Multidimensional Scaling: Theory and Applications* 2nd edn (Springer, New York, 1997).
40. Hino, H., Fujiki, J., Akaho, S. & Murata, N. Local intrinsic dimension estimation by generalized linear modeling. *Neural Comput.* **29**, 1838–1878 (2017).
41. Hino, H. Intrinsic Dimension Estimation with R. *R J.* **9**, 329–341 (2017).
42. Grassberger, P. & Procaccia, I. Measuring the strangeness of strange attractors. *Phys. D* **9**, 189–208 (1983).
43. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Proc. Second International Symposium on Information Theory* (eds Petrov, B. N. & Csaki, F.) 267–281 (Akademiai Kiado, Budapest, 1973).
44. Weinberger, K. Q., Blitzer, J. & Saul, L. K. in *Advances in Neural Information Processing Systems* (eds Weiss, Y., Ikonf, B. S. O. & Platt, J. C.) Vol. 18, 1473–1480 (MIT, Cambridge, 2006).
45. Xing, E. P., Jordan, M. I., Russell, S. J. & Ng, A. Y. *Distance Metric Learning with Application to Clustering with Side-Information* (MIT, Cambridge, 2003).
46. Davis, J. V., Kulis, B., Jain, P., Sra, S. & Dhillon, I. S. Information-theoretic metric learning. in the *24th International Conference on Machine Learning*. 209–216, Zoubin Ghahramani Ed. (ACM Press, New York, 2007). <https://doi.org/10.1145/1273496.1273523>.
47. Mathew, K. et al. High-throughput computational X-ray absorption spectroscopy. *Sci. Data* **5**, 180151 EP– (2018).
48. Saito, K. et al. Accelerating small-angle scattering experiments on anisotropic samples using kernel density estimation. *Sci. Rep.* **9**, 1526 (2019).
49. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
50. Giorgino, T. Computing and visualizing dynamic time warping alignments in R: the dtw Package. *J. Stat. Softw.* **31**, 1–24 (2009).
51. Burnham, K. P. & Anderson, D. R. *A Practical Information-Theoretic Approach. Model Selection and Multimodel Inference* 2nd edn (Springer, New York, 2002).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019