Feature Space of XRD Patterns Constructed by an Autoencoder

Keishu Utimula,* Masao Yano, Hiroyuki Kimoto, Kenta Hongo, Kousuke Nakano, and Ryo Maezono

X-ray diffraction (XRD) is commonly used to analyze systematic variations occurring in compounds to tune their material properties. Machine learning can be used to extract such significant systematics among a series of observed peak patterns. The feature space concept, in the context of autoencoders, can be the platform for performing such extractions, where each peak pattern is projected into a space to extract the systematics. Herein, an autoencoder is trained to learn to detect the systematics driven by atomic substitutions within a single phase without structural transitions. The feature space constructed by the trained autoencoder classifies the substitution compositions of XRD patterns satisfactorily. The compositions interpolated in the feature space are in good agreement with those of an XRD pattern projected to a point. Subsequently, the autoencoder generates a virtual XRD pattern from an interpolated point in the feature space. When the feature space is effectively optimized by enough training data, the autoencoder predicts an XRD pattern with a concentration, which is difficult to be described using the possible resolution of the supercell method of ab initio calculations.

1. Introduction

Materials informatics (MI) is becoming prevalent as an application of big data science for the discovery of materials.[1,2] Because XRD analysis is one of the commonly used methods in most material development fields, there have considerable efforts toward the application of machine learning for extracting significant characteristics from observed XRD peak patterns. One of the most common applications of XRD has been the automatic identification of the space group of a crystal structure through XRD peak patterns. The techniques used for this purpose include dynamical time warping (DTW),[3–6] non-negative matrix factorization (NMF),[7] stochastic neighbor embedding with t-distribution (t-SNE),[8] and agile combinatorial factor decomposition (AgileFD),[9,10] but majority of previous studies have used neural networks (NNs) and deep learning.[11–20] The targets range not only for space group identification but also for the identifications of mixtures,[21,22] chemical compositions within a single phase,[4] etc.

In this study, we target the identification within a single phase. There is a demand for tuning the properties of a crystal through atomic substitutions while maintaining its space group.[23] The compositional change of the doped elements leads to slight changes in XRD peak positions, which can be detected to identify the composition of dopants. Studies using DTW to detect the peak shifts have been conducted, showing successful results in clustering XRD patterns according to the composition.[4] Compared with the case of recognizing the peak pattern across phases, the detection of a slight peak-position shift within a single phase, producing almost the same pattern, is more challenging; however, it can be achieved through DTW.[4] It is important to know the peak that is responsible for these slight shifts. Owing to the atomic substitutions, the peaks corresponding to some bonding lengths cannot shift significantly, while others cannot be shifted entirely. Even with significant intensities, such peaks with slight shift in their positions would be irrelevant allowing compositional identification. We considered the possibility to extract such characteristics in peak shifts using feature spaces of machine learning.

An autoencoder is used to perform feature extraction using machine learning. The extraction is performed using dimensionality compression based on a neural-network framework. Such a framework has also been applied to noise reduction,[24] anomaly detection,[25] and novelty detection.[26]
relating to materials science, the identification of phase transitions in spin models using an autoencoder has been reported.\cite{27}

In addition to identification purposes, recent studies have actively used the information constructed in the feature space. For example, there are applications of assessing reconstruction loss as a measure of novelty prediction,\cite{28} or imaging the distribution in feature space with t-SNE and using it to switch between different dimensional representations (2D images to 1D spectra).\cite{29}

We are therefore interested in how the XRD peak relevancy explained above could be captured the feature space of autoencoders. The features are found to be captured well in a 2D feature space, from which we can reproduce the original data precisely. Distance on the space can be used to identify the composition of doped compounds. Surprisingly, the distance can also be used to identify relevant peaks for capturing the features. This is quite in contrast with the traditional approach employed by materials science researchers who interpret the patterns using crystallography where the features are captured using multiple concepts such as Bravais lattice estimations (e.g., Rietveld analysis),\cite{30} indexing of each diffraction peak, fitting of lattice constants (e.g., Vegard’s law),\cite{31,32}, and charge density estimations (e.g., MEM analysis).\cite{33–36} Though our result can be technically described as only a dimension compression from 11 900 original dimensions to two dimensions by an autoencoder, it is surprising that a target problem which is traditionally considered from multiple views by humans is now captured well in only two dimensions by machines.

2. Model and Methodology

We applied our framework to a data set of XRD patterns of magnetic alloys, [Sm\((1-x)\)Zr\(x\)]Fe\(_{12-y}\)Ti\(_y\), with different concentrations \(x\) and \(y\).\cite{4} A fixed concentration still includes multiple possibilities for the inequivalent locations of the crystalline sites.\cite{4} For the ten different concentrations considered here (Table 1), we have recorded 150 XRD patterns corresponding to inequivalent symmetries analyzed with the space-group clustering tool.\cite{37} There is a variation in the number of non-equivalent structures depending on concentration, which causes the imbalance in the training data. In fact, as we will see later in detail, for the case of SmZrFeTi, this imbalance results in low prediction accuracy for this composition as a special case. XRD analysis was performed by using VESTA (RIETAN-FP)\cite{38,39} for the structure relaxed by DFT. In this XRD pattern generation, we used 0.496 Å as the wavelength and isotropic atomic displacement parameter \(B\) of 1.00 Å.\cite{4} The Vienna ab initio simulation package (VASP) was used for DFT with the revised Perdew–Burke–Ernzerhof (RPBE) exchange-correlation functionals. The plane wave cutoff energy was chosen as 400 eV, and the k-points to 5×5×5 Monkhorst-Pack grids were used, which were large enough for energy convergence. The above-mentioned conditions have been found to cause an underestimation of the lattice constants in the present system.\cite{4} Accordingly, the DFT-XRD peak positions were shifted by \(\delta \Phi = 0.1^\circ\), in relation to those from experiments. Because we used a resolution of \(\delta \Phi = 0.01^\circ\) for the input of our NN autoencoder, the above corresponds to the shift with \(\approx 10\) grids in the input edge. If the experimental XRD were to be applied to the present framework, it would be necessary to estimate the above shift in advance and preprocess it with the grid shift by that amount. Alternatively, another NN can be prepared before the input of the current autoencoder, and it can be trained to fill the difference between the experiment and the simulation by using the transfer learning technique.

The data are inputted to a neural network as a 11 900-dimensional vector \(\{I(2\delta)\}_{\delta=0}^{120}\) \((2\delta \approx 0–120\,^\circ, \delta\delta = 0.01\,^\circ)\) at the input layer. Note that the range of \(2\delta\) in our case corresponds to the experimental setting using synchrotron radiation \((\lambda = 0.496\,\text{Å})\). As an autoencoder, the network encodes an input vector, compressing its dimension using hidden layers to two dimensions, and then decodes it out toward the output-layer with the same dimension as that of the input. The parameters in the network are optimized such that the output can identically reproduce the input. For implementation, we used PyTorch with activation functions ReLU\cite{41} for both the encoder and decoder hidden layers. At the final layer of the encoder and decoder, tanh (a nonlinear function) and a linear function are used as the activation function, respectively. Parameters are optimized by the Adam\cite{42} algorithm combined with mean squared error loss function (MSELoss) and L2-norm weight decay error estimations. For training, 90% of the data were used, and the rest were used for testing. We found that the error minimizing network construction (hyper-parameters), which required over ten samples for numerical stabilization, had three hidden layers, a minibatch of size 10, and 1000 epochs.

The network compresses the dimensions of the 11 900-dimensional data at the input into 128 → 64 → 12 dimensions via three hidden layers, and finally creates the 2D feature space at the last edge of the encoder (namely, at the middle of the whole autoencoder). Figure 1 shows the feature space, on which 150 XRD patterns are projected to each point. The initial distribution of the points (panel (a)) is scattered to form clusters as shown in panel (b) as the network learns. We observe that each cluster (as shown by circles in the panel (b)) is formed by patterns with the same concentration (corresponding to the same symbol) of substituents, and hence that the feature space can identify the compositions of the samples.

3. Results and Discussion

Figure 2 compares the input XRD pattern and that reproduced by the autoencoder. These agree with each other, showing that the autoencoder learnt the XRD patterns data correctly. In the

Table 1. Number of inequivalent structures (irreducible structures) in the sense of spatial symmetry for Sm\((1-x)\)Zr\(x\)Fe\(_{12-y}\)Ti\(_y\) to be considered. The numbers in brackets indicate the structures constructed from the \(2 \times 2 \times 2\) supercell (Sm/Zr), while the rest are constructed from the \(2 \times 2 \times 1\) supercell (Fe/Ti).

| \(y/x\) | 0.0 | 0.5 | 1.0 | 1.5 | 2.0 |
|-------|-----|-----|-----|-----|-----|
| 0.000 | 1   | 13  | 22  | 27  | 67  |
| 0.125 | -   | -   | (2) | -   | -   |
| 0.250 | -   | -   | (7) | -   | -   |
| 0.375 | -   | -   | (6) | -   | -   |
| 0.500 | -   | -   | (10) | -  | -   |
Figure 1. Distributions of 150 XRD patterns projected to the 2D feature space created by our autoencoder. Points with the same symbol have the same concentration of atomic substitutions but differ in their location since they are group theoretically inequivalent. As the parameters are optimized (i.e., the learning of the neural network is completed), the initial distribution [panel (a)] forms clusters as shown in panel (b) enclosed by circles.

Figure 2. Comparison of the input XRD pattern and that reproduced by our autoencoder. They coincide well with each other at the resolution of human eyes, demonstrating that the autoencoder learnt the data well. Note that the range of $2\theta$ in our case corresponds to the experimental setting using synchrotron radiation ($\lambda = 0.496\text{Å}$).

following discussions we provide several possible ideas on how to utilize the extracted features of XRD patterns: A) identifying doping concentrations for a given XRD pattern of unknown samples, B) clarifying the irrelevancy of each peak in a pattern in contributing to the features, C) generating artificial XRD patterns for a given concentration as the interpolation over XRD patterns to avoid expensive ab initio analysis, followed by D) a discussion justifying the chosen dimensions of the feature space.

3.1. Identification of Concentrations

Since in Figure 1b, samples that are close to each other on the feature space have a similar concentration, it is quite likely for samples with unknown concentrations to be projected to a location close to samples with similar concentrations. One could then estimate the unknown concentration using the distance on the feature space.

For such an identification, we paint the feature space as shown in Figure 4, where the color corresponds to the concentration. Based on the color, the location of a given sample that is projected can be determined and the concentration of the sample can be estimated. One would wonder whether the painting can be performed by a clustering technique (unsupervised machine-learning) such as the k-means method. However, we have concluded that the k-means method is never the appropriate choice for the painting: the method relies on the center of gravity of data that forms each cluster region. The data are sorted based on the distance from the center of gravity of each cluster. Such a concept would work well if each cluster forms a simply-connected region; otherwise, points closer to the center might be out of the region. As seen in the solid circled region in Figure 1b, our data can form clusters with the same feature that are not simply-connected (triangle symbols form two regions separate from each other). Rather than using global knowledge over the data such as the center of gravity, as concluded, it is better to use local information in the vicinity of the projected point.

As the simplest way to use such local information, we could use linear interpolation as explained in Figure 3. Suppose the target sample (an XRD pattern) with an unknown property
(concentration in this case) is projected to \( P \), in the vicinity of which we have three data points, \( A \), \( B \), and \( C \) with known properties, \( Q_1 \), etc. When the location of \( P \) is described as
\[
\vec{x}_{AP} = s \cdot \vec{x}_{AB} + t \cdot \vec{x}_{AC}
\]
the quantity for \( P \) is naively estimated using the same fractions as
\[
Q_P = Q_s + s \cdot (Q_B - Q_s) + t \cdot (Q_C - Q_s)
\]
To assign three known points in the vicinity of \( P \) into \( A \), \( B \), and \( C \) we would select them such that the inner products \( (\vec{x}_{AB} \cdot \vec{x}_{AP}) \) and \( (\vec{x}_{AC} \cdot \vec{x}_{AP}) \) are sufficiently large to obtain plausible interpolations (as is shown in Figure 3). This condition may be interpreted as follows: \( P \) should be located inside of the triangle formed by \( A \sim C \), which leads to the condition for the fractions, \( s, t > 0 \) and \( 1 > s + t > 0 \), as implemented the program. From this point of view, we chose three points which give minimum \( s^2 + t^2 \) from 12 points near \( P \).

By sweeping \( P \) over the entire feature space and identifying three of its nearest neighbors, we can estimate the concentration of any point on the space by using the above formalism, resulting in the painted map shown in Figure 4. By using this map, we can identify the concentration of a given sample that is projected to the space. Using test data points (for which the concentration is known), we can examine the performance of the maps predictions by comparing the known concentration to the estimated.

Table 2 lists the results in order of poorest performance measured using the error for the concentration of \( Sm_{x_1}Zr_{x_2}Fe_{x_3}Ti_{x_4} \) defined as
\[
\delta = \sum \left( c_i^{[\text{est}]} - c_i^{[\text{true}]} \right)^2.
\]
where \( A \) and \( E \) are the known concentration and estimate, respectively.

Table 2. [estimation] Estimation performance of test data sample concentrations using our linear estimation shown in order of the poorest performance measured using the error, which is defined in Equation (3) [the worst ten are shown]. The worst performance comes from the fact that there is little learning data for this case as explained in the main text.

| True composition | Estimated composition | Error |
|------------------|-----------------------|-------|
| Sm_{0.02}Zr_{0.02}Fe_{1.01}Ti_{1.00} | Sm_{0.02}Zr_{0.02}Fe_{1.73}Ti_{2.07} | 1.43 \times 10^{-1} |
| Sm_{0.02}Zr_{0.02}Fe_{1.01}Ti_{1.00} | Sm_{0.02}Zr_{0.02}Fe_{1.95}Ti_{1.04} | 3.59 \times 10^{-3} |
| Sm_{0.02}Zr_{0.02}Fe_{1.01}Ti_{1.00} | Sm_{0.02}Zr_{0.02}Fe_{1.52}Ti_{1.48} | 9.85 \times 10^{-4} |
| Sm_{0.02}Zr_{0.02}Fe_{1.01}Ti_{1.00} | Sm_{0.02}Zr_{0.02}Fe_{1.32}Ti_{0.97} | 9.31 \times 10^{-4} |
| Sm_{0.02}Zr_{0.02}Fe_{1.01}Ti_{1.00} | Sm_{0.02}Zr_{0.02}Fe_{1.22}Ti_{0.87} | 9.31 \times 10^{-4} |
| Sm_{0.02}Zr_{0.02}Fe_{1.01}Ti_{1.00} | Sm_{0.02}Zr_{0.02}Fe_{1.32}Ti_{0.97} | 9.85 \times 10^{-4} |
| Sm_{0.02}Zr_{0.02}Fe_{1.01}Ti_{1.00} | Sm_{0.02}Zr_{0.02}Fe_{1.49}Ti_{1.97} | 1.85 \times 10^{-4} |
| Sm_{0.02}Zr_{0.02}Fe_{1.01}Ti_{1.00} | Sm_{0.02}Zr_{0.02}Fe_{1.49}Ti_{1.00} | 7.39 \times 10^{-5} |
| Sm_{0.02}Zr_{0.02}Fe_{1.01}Ti_{1.00} | Sm_{0.02}Zr_{0.02}Fe_{1.49}Ti_{1.00} | 4.80 \times 10^{-5} |
| Sm_{0.02}Zr_{0.02}Fe_{1.01}Ti_{1.00} | Sm_{0.02}Zr_{0.02}Fe_{1.49}Ti_{1.00} | 1.03 \times 10^{-5} |
| Sm_{0.02}Zr_{0.02}Fe_{1.01}Ti_{1.00} | Sm_{0.02}Zr_{0.02}Fe_{1.49}Ti_{1.00} | 4.26 \times 10^{-6} |
| Sm_{0.02}Zr_{0.02}Fe_{1.01}Ti_{1.00} | Sm_{0.02}Zr_{0.02}Fe_{1.49}Ti_{1.00} | 1.20 \times 10^{-6} |

Figure 4. Estimations of the substitutional concentrations for any point on the feature space represented by the color map. By using such maps, we can identify the concentration of a given sample that is projected to the space.
The general consensus is that it is impossible to determine what the quantities on the vertical and horizontal axis are in a compressed 2D feature space. In this study, however, it may be possible to determine what they are to some extent in the following manner: The features in the present XRD case are those characterizing crystal structures tuned by the substituents, for example, the c/a ratio, etc. Generating variations that artificially capture such features through simulations is straightforward. By observing how the locations of the projected points on the feature space are affected by these structural changes, we can extract a trend of what the vertical and horizontal axis represents.

In some compositions, clusters are concentrated in a narrow region, while others are widely spread (e.g., Sm\textsubscript{1.0}Zr\textsubscript{0.0}Fe\textsubscript{0.0}Ti\textsubscript{2.0}). This reflects the variation in the number of irreducible structures\textsuperscript{[4,37]} for each composition, as described below. Even for the same composition, there are several inequivalent structures (irreducible structures) in terms of the space group, depending on which sites are substituted. In this study, XRD patterns are calculated for the relaxed structures obtained by applying DFT geometrical optimizations to each irreducible structure. Therefore, for the same composition, there are several different peak patterns corresponding to the number of irreducible structures with different relaxed geometries. As shown in Table 1, the number of irreducible structures is different for each composition. Therefore, a composition with more variations of irreducible structures gives a wider spread of its cluster in the feature space. Sm\textsubscript{1.0}Zr\textsubscript{0.0}Fe\textsubscript{0.0}Ti\textsubscript{2.0}, which is the most widely spread in the feature space, is the composition with the most variations of irreducible structures, as shown in Table 1. The structures at x = 1.0, 1.5, which have the second most variations also tend to be the most widely spread in the feature space.

In our results, the position in the feature space reflects the substitution composition because the atomic substitutions modulate the lattice constants, and they are reflected in the XRD peak shifts. Considering this mechanism, we notice that the possibility of using machine-learning recognition of XRD patterns to predict physical properties would be limited: It will be difficult to capture the properties in the feature space unless they are closely related to the lattice parameter modulation. If the objective is limited to composition identification from the beginning, it may be more convenient to construct a “composition-axis feature space.” Because the composition of all training data is known as the answer, it would be possible to construct a neural network (NN\textsubscript{\text{-}NN\textsubscript{\text{d}}}) with the XRD pattern (11900 dimensions) as the input, setting the output end to the 2D composition (x, y), and optimizing parameters to reproduce the composition of the answer. Combining its reversed NN\textsubscript{\text{d}} (a neural-network that reproduces the XRD pattern of the training data using the composition of the answer as a 2D input, where “\text{-}e” and “\text{-}d” stand for encoding and decoding, respectively) attached to the output edge of NN\textsubscript{\text{e}}, it would be possible to construct a NN performing the same function as our developed autoencoder. This “NN\textsubscript{\text{-}NN\textsubscript{\text{d}}}” provides the 2D projection space, on which we can easily identify the predicted composition for any given XRD pattern. The “NN\textsubscript{\text{-}NN\textsubscript{\text{d}}}” is, however, different from autoencoders in their purpose. Autoencoders usually learn parameters and extract features heuristically in a way that does not restrict the axis configuration of the feature space. Because our autoencoder heuristically confirmed that the feature space is well-linked to the composition space, it can be argued that the “NN\textsubscript{\text{-}NN\textsubscript{\text{d}}}” configuration can be used for composition identification.

3.2. Identifying Relevant Peaks

Whether all the peaks in an XRD pattern are required to characterize a crystal’s structure, or only a specific bandwidth of \(2\theta\) is sufficient would be a fundamental question in XRD. In this context, we would like to identify how much each peak is relevant incharacterizing the features of XRD. The naivest idea to measure this relevance would be to determine which peaks disappeared after the XRD pattern is reproduced by the autoencoder by comparing the input and output patterns. This approach unfortunately does not work because, as seen in Figure 2, the optimization of our neural network is so successful that the output reproduces the input pattern very well (none of the peaks disappeared).

We can instead use the idea that the relevance can be measured by how much the projected location on the feature space is affected when the considered peak is masked. Here, we define a mask vector \(M(2\theta)\) with the same dimensions as the XRD pattern, and its components become zero for \(2\theta \approx 2\theta + \Delta(\Delta = 0.03)\) and one otherwise. By using the mask vector, the masked XRD pattern can then be represented as \(XRD_{\text{m}}(2\theta) := M(2\theta) \circ XRD\), where the \(\circ\) is the Hadamard product. Suppose \(P[XRD]\) is the XRD pattern projected onto the feature space, the displacement on the feature space caused by masking can be evaluated as a normal distance on the feature space: \(\text{Dev}(2\theta) = \|P[XRD_{\text{m}}(2\theta)] - P[XRD]\|\). This would measure how much the masked peak at \(2\theta\) affects the XRD location on the feature space, and hence, corresponds to the “relevance intensity.” The plot of the intensity, \(\text{Dev}(2\theta)\), is shown in Figure 5, superimposed on the XRD pattern. This XRD pattern is for the test data; the composition is Sm\textsubscript{1.0}Zr\textsubscript{0.0}Fe\textsubscript{0.0}Ti\textsubscript{2.0}.

Both the original XRD and the relevance intensity are strongly correlated (namely, the relevance is larger when the XRD intensity gets larger), but at some peaks [e.g., at (a) in Figure 5], they are not correlated. Though the peak at (a) has significant intensity, its relevance is almost zero. One might suspect that this might be because the peak comes from a higher-order plane index, and thus, has some intensity but relevant information is already reflected in the lower-order plane index. We found, however, that the peak (a) has a fundamental of index, (0,0,1), which is not higher-order.

For the existence of peaks that are not relevant, although their peak intensities are large, it is important to recall here that the relevancy is defined by the training dataset. In this study, we trained our NN to detect how peak positions are affected when specific composition ratios such as Sm/Zr and Fe/Ti are modified. Even if the peak has a large intensity, if the peak position hardly changes with respect to the composition in question, then the relevancy will be classified as small. This explains why there are peaks with larger intensities but tiny relevancies.

In this study, the feature space was created to characterize slight differences among the patterns within the single phase, because the learning data belong to the same phase. If the training data is formed from the patterns of different phases, it is likely that minor differences in the patterns within a single phase are regarded as identical, and the autoencoder is trained to extract the differences across the different phases. In such a case, the evaluation criteria for the peak relevancy may change from that...
established in this study. More specifically, the peak indicated by arrow (a) in Figure 5 will have a large relevancy for interphase recognition, while the relevancy evaluated in this study is as small as the one that measures the composition dependence within a single phase.

This insight leads to a discussion on the generality of the method developed in this study. The core structure of the parameters learned in the developed autoencoder can be used as the initial set for any other autoencoder for recognizing an XRD pattern through transfer learning. In addition to the core structure, the collected training dataset would be different depending on the individual purpose of capturing the dependence of the peak patterns. Moreover, the resultant learned relevancy will be tailored to each purpose. Applying the masking strategy developed in this study to each relevancy provides a versatile framework as a tool that tells us which peaks to be focused on to capture the dependence for each purpose.

In this study, relevancy extraction is realized upon the dimensional reduction of the description space to get feature space. Therefore, it will be necessary to discuss whether other dimensional reduction methods such as t-SNE, UMAP (uniform manifold approximation and projection), and PCA (principal component analysis) can realize the same idea. The essence of the masking strategy in this study lies in the projected point deviation in the dimensional compressed projection space, and this idea does not depend on the way the projected space is constructed. If it does, the same idea can be realized by t-SNE/UMAP, which are other methods of providing the projected space. PCA is a method based on eliminating input information with little variation. In the context of PCA, a peak with a higher intensity and lower relevancy is regarded as such input information with a small variance over the training data. In other words, the information that was eliminated as a peak with low relevancy in this study can also be eliminated to the same extent through PCA. An insight into the mechanism of relevancy extraction developed by the current implementation of the autoencoder reveals that similar functions can be achieved by other methods, as discussed before.

We will discuss the contrast with conventional methods of XRD peak-pattern analysis. Conventional methods include “search-match” and “full profile analysis.” Practitioners usually apply the former first, such as comparing observed peak patterns with those existing on the database. The latter is used to identify the structure such that the assumed crystal structure is made to approach the observed XRD using Le Bail analysis [45] or Rietveld analysis. [46] To upgrade these methods, there have been preceding studies conducted to apply data scientific methods such as wavelets and PCA. [47] These preceding methods are used to relate the observed XRD pattern to a single crystal structure, and for this applications, data scientific methods are rapidly developed. In contrast to these, our method relates a series of XRD peak shifts to systematic structural changes triggered by compositional changes. By making the dynamic trend learn from the data, the analysis time is shortened. Even if the conventional one-point analysis has been accelerated by data science, there may be many cases in which a repetitive one-point analysis is not cost-effective for analyzing a dynamic trend. The proposed method can be effective in such cases.

3.3. XRD Patterns Generated by Autoencoder

XRD provides inferences not only of the lattice constants but also of other valuable information such as strains, crystalline sizes,
Figure 6. A comparison between XRDAE and XRDDFT. As the composition not used in the training, \((x, y) = (0.5, 0.5)\), was chosen. The range of \(2\theta = [1216]\) is shown enlarged (inset is the full range of \(2\theta = [1, 40]\)).

etc. Therefore, learning data that are compared with experimental XRD patterns have the finest composition resolution. Realizing such a fine resolution via ab initio simulations is, however, generally very difficult. The conventional treatment of atomic substitutions using the supercell model requires a very large supercell to represent a tiny percentage of substitutions, which is practically impossible to perform. Hence, we interpolated discrete points (50%, 25%, 12.5%, etc.) that can feasibly be obtained at a realistic cost using ab initio simulations. This can actually be realized through our feature space. Since a composition contour is constructed on the feature space as in Figure 4, we can interpolate to find a point corresponding to a desired composition on the contour. The output pattern generated by the autoencoder projected from the point on the feature space could be a plausible composition XRD pattern that is obtained without costly ab initio simulations. This approach can be applied not only to XRD but also to other spectrum and even to other physical quantities that are evaluated by ab initio simulations. The approach can also be regarded as another remedy for the long-standing difficulties in ab initio calculations, that is, the computational cost of treating small concentrations with large supercell models.

As a way to verify how successfully the interpolation described in is working, we investigated the quality between the virtual XRD pattern (XRDAE) generated by the autoencoder from the interpolated point. A useful technique using virtual data generated by autoencoders has become popular recently in the context of generative adversarial networks (GANs).\(^{48-51}\) For the composition corresponding to interpolation points, we can also generate XRD patterns using DFT (XRDDFT). If the compositional identification by interpolation described in is successful, then the XRDAE and XRDDFT would be expected to agree reasonably well. Figure 6 shows a comparison between XRDAE and XRDDFT for \((x, y) = (0.5, 0.5)\). Though the inset (whole range) gives us the impression that the coincidence is accurate, the main figure (enlarged range for \(2\theta\)) shows some differences to some extent, but the peak features are well captured. Because there are \(\approx 150\) training data points, the agreement will improve as the number of training data points increases.

Because the comparison of XRD depends on the quality of the training data, another convincing fact would be required to show that the interpolation is working properly. Denoting the composition as \(\bar{x} = (x, y)\), we consider the deviation between XRDAE(\(\bar{x}\)) and XRDDFT(\(\bar{x}_0\)).

\[
\Delta(\bar{x}) = \left| \text{XRDAE}(\bar{x}) - \text{XRDDFT}(\bar{x}_0) \right|
\]

where, the difference is defined by the squared summation of the difference of the intensity at each \(2\theta\)." If the interpolation is properly working, then \(\Delta(\bar{x})\) should behave as if it has a minimum at \(\bar{x} = \bar{x}_0\), and increases as \(\bar{x}\) moves away from \(\bar{x}_0\). A plot of \(\Delta(\bar{x})\) evaluated as \(\bar{x}_0 = (0.5, 0.5)\) is shown in Figure 7. The behavior is roughly as expected, supporting that the interpolation is working properly.

3.4. Dimension of the Feature Space

In this study, the original dimension of an XRD pattern is compressed into a 2D feature space. The feature space need not be 2D, but we will show that a 2D space is the optimal choice in the following analysis. Figure 8 shows the dependence of the estimation errors on the choice of the dimensions of the feature space. The error is measured as the squared deviation between the chemical composition of a sample and its prediction.
Figure 7. Log-plot of $\Delta(\vec{x})$ evaluated at the reference point $\vec{x}_0 = (0.5, 0.5)$.

Figure 8. Dependence of the estimation errors on the choice of the feature space dimensions. The error is defined as the squared deviation between the chemical composition of a sample and its prediction from the linear interpolation made in the feature space as explained in Figure 3. The distributions of the mean value of the error obtained by $k$-hold-out cross-validation method (the number of the subset = 10) are shown as box-and-whisker plots with logarithmic scale.

from the linear interpolation made in the feature space as explained in Figure 3. Data shown in Figure 8 are the distribution of the mean value of the error obtained using the $k$-hold-out cross-validation method (with subset = 10) shown as box-and-whisker plots. The plots clearly show that the choice of dimensions is optimal at two dimensions. As a general rule, the number of dimensions increases with the apparent accuracy, but generalizability is reduced by overlearning: a choice of optimal dimensions balances this trade-off. A plausible reason for the 2D space being optimal might be that the crystal structure has...
two degrees of freedom because of the constraints, $a = b$ and $\alpha = \beta = \gamma = 90^\circ$.

4. Conclusion

We developed an autoencoder to construct a feature space that describes XRD patterns. The framework was trained using XRD pattern data to identically reproduce the input peak patterns at the resolution of human eye. Each XRD pattern is projected to a point on the 2D feature space, forming clusters of samples with similar concentrations. The distance in the space can be used to estimate the concentration of any given sample by projecting its XRD pattern to the space. We drew a contour map on the space that describes the concentration by linearly interpolating between points produced from the training data. Using the test data to examine the prediction performance, we confirmed a percentage error less than 0.5% except for cases with little training data. We proposed two applications of the feature space: The first is determining the relevance of each peak in characterizing XRD features. This is done by observing the change in location of a point on the feature space when a peak from the original XRD pattern is masked. Using this method, we found a nontrivial case where a high intensity peak had a low relevance intensity; we could not make a reasonable account for the irrelevancy from the physics of XRD (e.g., higher-order reflections, etc.) The second is interpolating XRD patterns to avoid expensive ab initio simulations, with which it is difficult to simulate small changes in concentrations. The interpolation can be done on the feature space, and hence, the autoencoder can generate an artificial but plausible XRD pattern for the interpolated point with the desired composition. The approach would be regarded as a useful alternative to the computationally expensive ab initio simulations for fine resolutions of concentrations.

Acknowledgements

The computations in this work have been performed using the facilities of Research Center for Advanced Computing Infrastructure at JAIST. R.M. is grateful for financial supports from Reference MEXT-KAKENHI (19H04692 and 16KK0009), from FlagsHIP2020 (project nos. hp190169 and hp190167 at K-computer), from Toyota Motor Corporation, from I-O DATA Foundation, from the Air Force Office of Scientific Research (AFOSR-AOARD/FA2386-17-1-4049; FA2386-19-1-4015), and from JSPS Bilateral Joint Projects (with India DST). The X-ray diffraction (XRD) measurements for the Sm-Fe-Ti system were performed at beamline BL02B2 of SPring-8 under the proposal No. 2017B1634. K.H. is grateful for financial support from MEXT-KAKENHI, Japan (JP19K05029, JP19H05169, JP21K03400, JP21H01998, and JP22H02170), JSPS Bilateral Joint Research Seminars with Korea NRF (JP20S220218802), and the Air Force Office of Scientific Research, United States (Award Numbers: FA2386-20-1-4036).

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

autoencoder, feature extraction, machine learning, materials informatics, X-ray diffraction

Received: August 24, 2022
Published online: December 18, 2022

[1] H. Ikebata, K. Hongo, T. Isomura, R. Maezono, R. Yoshida. J. Comput.-Aided Mol. Des. 2017, 31, 379.
[2] C. Oses, E. Gossett, D. Hicks, F. Rose, M. J. Mehl, E. Perim, I. Takeuchi, S. Sanvito, M. Scheffler, Y. Lederer, O. Levy, C. Toher, S. Curtarolo. J. Chem. Inf. Model. 2018, 58, 2477.
[3] Y. Iwasaki, A. G. Kusne, I. Takeuchi, npj Comput. Mater. 2017, 3, 4.
[4] K. Utimula, R. Hunkao, M. Yano, H. Kimoto, K. Hongo, S. Kawaguchi, S. Suwanna, R. Maezono, Adv. Theory Simul. 2020, 3, 2000039.
[5] R. Bellman, R. Kalaba, IRE Trans. Autom. Control 1959, 4, 1.
[6] S. Salvador, P. Chan, Toward Accurate Dynamic Time Warping in Linear Time and Space, vol. 17, Conf.: KDD workshop on mining temporal and sequential data 2004, pp. 70–80.
[7] C. J. Long, D. Bunker, X. Li, V. L. Karen, I. Takeuchi, Rev. Sci. Instrum. 2009, 80, 103902.
[8] Y. Suzuki, H. Hino, T. Hawai, K. Saito, M. Kotsugi, K. Ono, Sci. Rep. 2020, 10, 21790.
[9] J.-W. Lee, W. B. Park, J. H. Lee, S. P. Singh, K.-S. Sohn, Nat. Commun. 2020, 11, 86.
[10] S. K. Suram, Y. Xue, J. Bai, R. Le Bras, B. Rappazzo, R. Bernstein, J. Bjorck, L. Zhou, R. B. van Dover, C. P. Gomes, J. M. Gregoire, ACS Comb. Sci. 2017, 19, 37.
[11] W. B. Park, J. Chung, J. Jung, K. Sohn, S. P. Singh, M. Pyo, N. Shin, K.-S. Sohn, IJUCJ 2017, 4, 486.
[12] P. M. Vecsei, K. Choo, J. Chang, T. Neupert, Phys. Rev. B 2019, 99, 245120.
[13] L. C. O. Tiong, J. Kim, S. S. Han, D. Kim, npj Comput. Mater. 2020, 6, 196.
[14] H. Wang, Y. Xie, D. Li, H. Deng, Y. Zhao, M. Xin, J. Lin, J. Chem. Inf. Model. 2020, 60, 2004.
[15] A. N. Zaloga, V. V. Stanovov, O. E. Bezrukova, P. S. Dubinin, I. S. Yakimov, Mater. Today Commun. 2020, 25, 10662.
[16] J. A. Aguilar, M. L. Gong, T. Tasdizen, Comput. Mater. Sci. 2020, 173, 107309.
[17] F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. T. P. Hartono, S. Ramasamy, B. L. DeCost, S. I. P. Tian, G. Romano, A. Gilad Kasne, T. Buonasissi, npj Comput. Mater. 2019, 5, 60.
[18] P. M. Maffettone, L. Banko, P. Cui, Y. Lysogorskiy, M. A. Little, D. Olds, A. Ludwig, A. I. Cooper, Nat. Comput. Sci. 2021, 1, 290.
[19] D. Jha, A. G. Kasne, R. Al-Bahrami, N. Nguyen, W.-K. Liao, A. Choudhary, A. Agrawal, in 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, Piscataway, NJ 2019, pp. 1–8.
[20] A. G. Kasne, H. Yu, C. Wu, H. Zhang, J. Hattrick-Simpers, B. DeCost, S. Sarker, C. Oses, C. Toher, S. Curtarolo, A. V. Davydov, R. Agarwal, L. A. Bendersky, M. Li, A. Mehta, I. Takeuchi, Nat. Commun. 2020, 11, 5966.
[21] J.-W. Lee, W. B. Park, M. Kim, S. Pal Singh, M. Pyo, K.-S. Sohn, Inorg. Chem. Front. 2021, 8, 2492.
[22] N. J. Szymanski, C. J. Bartel, Y. Zeng, Q. Tu, G. Ceder, Chem. Mater. 2021, 33, 4204.
[23] T. Yoshida, R. Maezono, K. Hongo. ACS Omega 2020, 5, 13403.
[24] Y. Bengio, A. Courville, P. Vincent, IEEE Trans. Pattern Anal. Mach. Intell. 2013, 35, 1798.
[25] L. Beggell, M. Pfeiffer, B. Bischi, in Machine Learning and Knowledge Discovery in Databases, (Eds: U. Brefeld, E. Fromont, A. Hotho, A.
Knobbe, M. Maathuis, C. Robardet, Springer, Cham 2020, pp. 206–222.
[26] T. Amarbayasgalan, B. Jargalsaikhan, K. H. Ryu, Appl. Sci. 2018, 8, 9.
[27] S. J. Wetzel, Phys. Rev. E 2017, 96, 022140.
[28] L. Baniko, P. M. Maffettone, D. Naujoks, D. Olds, A. Ludwig, npj Comput. Mater. 2021, 7, 104.
[29] H. S. Stein, D. Guevarra, P. F. Newhouse, E. Soedarmadji, J. M. Gregoire, Chem. Sci. 2019, 10, 47.
[30] H. Rietveld, J. Appl. Crystallogr. 1969, 2, 65.
[31] L. Vegard, Zr. Phys. 1921, 5, 17.
[32] A. R. Denton, N. W. Ashcroft, Phys. Rev. A 1991, 43, 3161.
[33] W. Jauch, A. Palmer, Acta Crystallogr. A 1993, 49, 590.
[34] W. Wei, J. Appl. Crystallogr. 1985, 18, 442.
[35] S. F. Gull, A. K. Livesey, D. S. Sivia, Acta Crystallogr. A 1987, 43, 112.
[36] M. Sakata, M. Sato, Acta Crystallogr. A 1990, 46, 263.
[37] G. I. Prayogo, A. Tirelli, K. Utimula, K. Hongo, R. Maezono, K. Nakano, J. Chem. Inf. Model. 2022, 62, 2909.
[38] F. Izumi, K. Momma, Solid State Phenom. 2007, 130, 15.
[39] K. Momma, F. Izumi, J. Appl. Crystallogr. 2011, 44, 1272.
[40] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. in Advances in Neural Information Processing Systems, SPIE, Bellingham, WA 2019, pp. 8024–8035.
[41] X. Glorot, A. Bordes, Y. Bengio, in Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, ACM, New York 2011, pp. 315–323.
[42] D. P. Kingma, J. Ba, arXiv:1412.6980 2014.
[43] J. MacQueen, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California, Berkeley, CA 1967, pp. 281–297.
[44] K. Utimula, K. Nakano, G. I. Prayogo, K. Hongo, R. Maezono, arXiv:1911.08071 2019.
[45] A. Le Bail, H. Duray, J. Fourquet, Mater. Res. Bull. 1988, 23, 447.
[46] H. M. Rietveld, J. Appl. Crystallogr. 1969, 2, 65.
[47] I. Ivanisevic, D. E. Bugay, S. Bates, J. Phys. Chem. B 2005, 109, 7781.
[48] Y. Zhao, B. Deng, J. Huang, H. Lu, X.-S. Hua, in Proceedings of the 25th ACM International Conference on Multimedia, ACM, New York 2017, pp. 244–251.
[49] J. Wang, W. Zhou, J. Tang, Z. Fu, Q. Tian, H. Li, in Proceedings of the 26th ACM International Conference on Multimedia, ACM, New York 2018, pp. 709–717.
[50] J. g. Kwak, H. Ko, Appl. Sci. 2020, 10, 6.
[51] J. Zhu, D. Zhao, B. Zhang, B. Zhou, Int. J. Comput. Vis. 2022, 130, 1259.
[52] J. W. Tukey, Exploratory Data Analysis, Addison-Wesley, Boston, MA 1977.