Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis

Akihiro Yamashita,a,b Takahiro Nagata,c Shinjiro Yagyu,c Toru Asahi,c and Toyohiro Chikyow,a,b

ABSTRACT
We built a workflow for the fabrication analysis of thin films by applying machine-learning (ML) techniques directly to the measurement data. This will lower the problem in cost of synthesizing and analyzing samples to improve the fabrication conditions. The workflow combines two ML techniques: non-negative matrix factorization (NMF) and variational autoencoder (VAE). The measurement data were two-dimensional X-ray diffraction of indium-gallium oxide system thin films. The thin films were fabricated by physical vapor techniques under multiple conditions. First, the workflow was applied to the data of the thin films fabricated through pulsed laser deposition as a proof of concept. We found that our workflow extracted features that represented crystallinity differences in addition to substrate differences. Second, VAE was analyzed to determine whether it could generate new data from its latent space. The latent space of the VAE, which learned the extracted features, represented the relationship between the fabrication conditions such as laser intensities and crystallinity. Third, the inference ability of the new data fabricated through sputtering was evaluated. The capability of the workflow we confirmed will support researchers in improving fabrication conditions by visually comparing various fabricated samples.

CONTACT Akihiro Yamashita racfcj@asagi.waseda.jp; Takahiro Nagata NAGATA.Takahiro@nims.go.jp Research Center for Functional Materials, National Institute for Materials Science, 1-1, Namiki, Tsukuba, 305-0044, Japan
Supplemental data for this article can be accessed here.
1. **Introduction**

Materials informatics is a promising method for accelerating material discovery. To enforce this method, a large dataset is essential. However, building large datasets in materials science is difficult because of the cost of synthesizing and analyzing samples to build datasets. To circumvent this, one solution is to use simulations such as ab initio calculations [1,2]. Another solution is to perform high-throughput experiments, such as automated laboratories [3] or combinatorial libraries [4,5]. A combinatorial library fabricates a sample under varying fabrication conditions, such as the composition ratio. Furthermore, automated analysis has attracted several researchers’ interest [6–14]. We propose the direct use of measurement data as an input for machine-learning (ML) models to increase the available data and reduce the analysis time. However, the direct use of measurement data has mainly the following three problems. First, raw measurement data are high-dimensional because of its resolution [15]. Second, the variation in the measured patterns is significantly smaller compared to those in the computer vision area. Third, slight differences in such small variance are important. For example, two-dimensional X-ray diffraction (2D-XRD) is a convenient measurement to analyze crystal structures of thin films. However, its dimensionality is high (e.g. $2048 \times 2048$ pixels), possible shapes of diffraction pattern are simple (spot or arc-like) [16], and the exact position of diffraction pattern is important. The three problems are specific to materials science and often overlooked in the computer-vision field. Therefore, we focus on the three problems and demonstrate the solution to directly use measurement data. Among advanced materials, we focus on thin films and attempt to relate their crystal structures with their fabrication conditions using ML techniques.

The application of ML to measurement data is well investigated in automated or on-the-fly analysis context [6–14]. In previous studies, non-negative matrix factorization’s (NMF) [17] advantage has been reported in automated analysis of X-ray diffraction (XRD) spectra from simulation or composition spreads under single fabrication condition [9,12,14]. In this study, we proceed the study into a representation learning context as well as a reevaluation of NMF applicability to 2D-XRD images from samples under multiple fabrication conditions. Representation learning is a research field to investigate handcrafted variables (called features or descriptors) to represent data in an informative manner, which heavily affect performance of ML methods [18]. In materials science, the available data is limited; therefore, research on features is important for advanced ML techniques such as deep learning (DL) [19]. Although features of the crystal structure have been reported [20–22], these features require prior information such as space group or lattice constants, which implies that the features are not applicable to on-the-fly analysis. In this study, we reevaluated NMF as a feature extraction method of 2D-XRD images to improve the performance of DL models, as well as on-the-fly analysis.

We built a workflow that combines NMF and variational autoencoder (VAE) [Figure 1 (a)]. The dataset contained 784 data from thin films of In$_2$O$_3$, Ga$_2$O$_3$, and the composition spreads of (In$_x$Ga$_{1-x}$)$_2$O$_3$ fabricated under multiple conditions (Table 1). Figure 1(b) and (c) show typical 2D-XRD images. The samples in this study were suitable for proof of concept because their material properties were already analyzed in another study [23]. First, we analyzed the NMF results to evaluate how well the extracted features represented the propensities of the 2D-XRD dataset containing multiple-fabrication conditions. For evaluation of features extracted by NMF, we applied VAE [24]. VAE is a deep generative model and reported as a promising model in materials science [25–28]. In the evaluation, two VAE models were trained; one directly learned from 2D-XRD images, and the other learned extracted features. The relationship between the fabrication conditions and the latent space learned by the VAE was analyzed. Finally, we evaluated the capability of the workflow against new data using a pseudoinverse matrix.

2. **Methods**

2.1. **Dataset and handling**

The thin films of In$_2$O$_3$, Ga$_2$O$_3$, and composition spreads of (In$_x$Ga$_{1-x}$)$_2$O$_3$ were used. These thin films were fabricated by pulsed laser deposition (PLD) on six substrate types, i.e. c-sapphire, yttria-stabilized zirconia (YSZ) (111), YSZ (100), Pt/Ti/SiO$_2$/Si (100), SrTiO$_3$ (STO) (111), and STO (100). These samples were originally fabricated in a previous study [23]. The original objective of the samples was to determine the appropriate fabrication conditions to synthesize In$_2$O$_3$ or Ga$_2$O$_3$ crystals or their solid solutions by changing the fabrication parameters, such as laser intensities or O$_2$ pressures. Preliminary experiments and trials-and-errors were conducted on the samples by experts. Therefore, the fabrication conditions were neither mutually exclusive nor collectively exhaustive.

The 2D-XRD images represented the slice plane of the diffraction sphere of the samples. The horizontal and arc-like axes represented the diffraction 2θ angle and chi angle, respectively. A typical image here contained diffraction signals of the substrate and thin film, as well as background noise. The peak shifts of the
solid solutions of In$_2$O$_3$ and Ga$_2$O$_3$ were our main concern. The 2D-XRD images were obtained using a Discover D8 system and Vantec 500 (Bruker AXS). During the measurements of the composition spreads, 11 measurement points were assigned such that each point represented a composition change of 10% from 0% to 100%. The number of 2D-XRD signals here were 512, and the original image size of each 2D-XRD signal was 2048 $\times$ 2048 pixels. All images were normalized by each maximum pixel value and represented in the column vector forms in the calculations. All the figures of the 2D-XRD images here were corrected at $\gamma = 5$ to improve readability.
Table 1. List of fabrication processes and conditions. It is noteworthy that the fabrication conditions were tested by experts; therefore, not all possible condition combinations were tested. For example, some c-sapphire samples were fabricated at room temperature with 40 mJ laser. However, no YSZ (111) sample was fabricated at room temperature. The number of data was 512 for PLD samples and 272 for sputtering.

| Parameters               | Values                                      |
|--------------------------|---------------------------------------------|
| Fabrication process      | Pulsed laser deposition (PLD), Sputtering   |
| Substrate type           | c-sapphire, yttria-stabilized zirconia (111) [YSZ (111)], YSZ (100), Pr/Ti/SiO2/Si (100), SrTiO3 (111) [STO (111)], STO (100) |
| Target                   | In2O3, BCl2O3, Ga2O3, Bi2O3, Ga2O3         |
| Laser intensity in PLD [mJ] | 20, 40, 70                                   |
| O2 pressure in PLD [Torr] | 1e-3, 1e-4, 1e-5                             |
| Radio frequency power in sputtering [W] | 20, 35, 50                                   |
| O2/Ar ratio in sputtering [%] | 5, 20, 40                                   |
| Substrate Temperature [°C] | Room temperature, 300, 400, 450, 500, 600, 650, 700 |

Data handling, such as resizing or conversion, and calculations, such as NMF and VAE, were conducted using the programming language JuliaLang [29] and its packages. Detailed explanations of the methods and packages are provided below.

### 2.2. Feature extraction and dimensionality reduction with NMF

Five hundred and twelve 2D-XRD signals from PLD samples were normalized by dividing them with their corresponding maximum values and resized to 512 × 512 pixels to reduce computational time. Thereafter, the signals were vectorized, and, consequently, the shape of dataset V was a 262,144 × 512 matrix. We set the number of factors to 10 because we assumed that this number would be sufficient to represent all data and moderate for human interpretation. Each column of W is called a basis vector or an end member in other studies [11,14]; however, here, we called it a factor vector or factor image. We referred to each column of H as a feature vector because this is a set of coefficients that represent the corresponding data by a linear combination of the factor images. Each row of H was recognized as the weight distribution of the corresponding factor.

Initialization and training algorithms were non-negative double singular value decomposition (NNDSVD) [30] and multiplicative update [17], respectively. We chose these algorithms because both NNDSVD and multiplicative update omit random processes. In addition, these algorithms are implemented in several programming languages such as Julia, Python, MATLAB, and R. Note that some implementations of NNDSVD use randomized singular value decomposition (SVD) for computational efficiency. Therefore, we used NMF.jl version 0.5.1, which was implemented without randomized SVD. Although more sophisticated algorithms [9,31,32] have already been proposed, we used the algorithms explained above because it is basic and easy to understand. The objective function and training times were divergence and 100, respectively.

### 2.3. VAE

We modeled two VAE models whose encoders and decoders each had one hidden layer. One VAE model directly learned the 2D-XRD images, which were resized to 64 × 64 pixels owing to the capacity of our computer. The other model learned the feature vectors. The sizes of the hidden layers were approximately the root square of the input sizes: 64 and 3, respectively. The output sizes of the encoders were 2; thus, we visualized the distributions of the latent vectors. The prior distributions of the latent spaces were set to normal distributions with averages and variances of 0 and 1, respectively. Both VAE models were programmed with Flux.jl [33] and trained 100 times.

### 2.4. Visualization of NMF factor axis in the latent space

To analyze the effect of NMF factors in the latent space of VAE, we visualized axes which the NMF factors constituted. To visualize, for example, the axis which factor 1 constitute, we encoded vectors whose first component ranged from 0 to 1, and the other components were 0. We conducted the procedure for all 10 factors. The origin was plotted by encoding a zero vector.

### 2.5. Feature extraction of new data

NMF is an approximation method $V \approx WH$, where $W$ is a set of factors and $H$ is the set of feature vectors. This implies that, with the new dataset $V^*$, multiplying the (pseudo-)inverse matrix of $W$ from the left of $V^*$, i.e. $W^{-1}V^* = H^*$, corresponds to extracting feature vectors of $V^*$. Moreover, $H^*$ is a set of extracted feature vectors. Therefore, $W^{-1}$ was recognized as a feature extractor. We computed the Moore–Penrose pseudo-inverse matrix of $W$. 
3. Results and discussion

3.1. Verifying NMF to 2D-XRD

3.1.1. Analyzing factor vectors

First, we applied NMF to 512 2D-XRD images of samples fabricated by pulsed laser deposition (PLD) with setting the number of factor images to 10. Figure 2 shows the factor images of the results. These images were stored in the columns of $W$ in vector form. Each original image was approximated by a linear combination of these factor images. These factor images were calculated ones; therefore, the exact image did not exist in the original dataset.

We interpreted each factor image with expertise and found that NMF learned mainly from major diffraction peak positions and shapes (Table 2). The first seven factor images were identified, and the remaining factors were categorized as unknown. The unknown

![Figure 2](image.png)

Figure 2. Ten factor images. All images are corrected at $\gamma = 5$ to improve readability and have the same color range. Factor images with shared color scale is provided in Figure S2. Each vectorized image corresponds to each column of the $W$. Lighter color represents stronger intensity; however, darker color represents weaker in. The shape of the valid detector region is circular. Heatmap of feature vectors are Figure S3.
factors contained signals that were similar to those of the c-sapphire substrate. However, in terms of weight distributions, we could not confirm that these signals corresponded to those of the c-sapphire substrate. Herein, the NMF failed to detect signals of Ga2O3, because they were weak. Although factor 10 contained Ga2O3 signals, we concluded that factor 10 was unknown referring to the weight distribution. By analyzing the weight distribution, we found that factors 4 and 6 both represented the diffraction signal of the c-sapphire substrate, with a slight difference. This difference had to be caused by a Ni-filter on the detector to eliminate reflections generated by X-ray Kβ. Some samples were measured with the filter, although we have not tracked the date when the filter was installed. The separation of factors 3 and 7 was based on the crystallinity of In2O3 (222), which is an advantage of 2D detector compared to one-dimensional (1D) detectors.

We also analyzed whether 10 factors were sufficient or not by calculating three metrics: the logarithm of posterior probability, mean squared error (MSE), and divergence (Figure S4). These metrics did not saturate over the number of factors ranging from 1 to 512. However, we concluded that 10 factors were not sufficient considering the metrics and interpretation of the factors.

We checked the training process of the NMF step by step (Supp. Video) and found that the training results converged quickly. Therefore, the number of training sessions in this study (i.e. 100 times) was sufficient. Considering the convergence speed, we assume that the initialization result will be a good indicator for estimating the number of factors.

Ideally, each factor image should represent one significant diffraction signal, but some factor images contained two or three signals (Figure 2). This could be because the number of factors was insufficient compared to the variety of the original dataset. However, we found that increasing the number of factors did not lead to mutually exclusive or collectively exhaustive factor images (Figure S5). Instead, using a larger number of factors could be worse for interpretation because factor images tended to contain meaninglessly separated signals. We assume that this could be because vectorized images lose spatial correlations among signals. Therefore, other methods that preserve spatial correlations, such as non-negative tensor factorization [34], will improve the factor images, but this has not been studied. Considering this, we conclude that NMF is applicable to 2D-XRD image datasets from multiple samples as well as 1D-XRD dataset [9,12,14].

### 3.1.2. Comparison of the objective function

To analyze the effect of the objective function, we compared two NMF results: the objective function of one was divergence and that of the other was the MSE. Comparing the factor images and weight distributions, we noticed that the objective function differed slightly in the factor images, but some weight distributions were differed significantly. Here, we focus on factors 5 and 6 (Figure 3), and Figures S6–S13 show the results of the other factors. Based on our expertise, we concluded that factor 5 represented the signal of the Pt electrodes on the Si substrate (Table 2). Here, only two composition-spread samples were used with the Pt electrodes (Data ID 480–490 and 491–501). The weight distribution of the divergence result was consistent with this fact that the samples with Pt electrodes were only two, but that of the MSE result had unreasonable non-zero values under ID 400. Factor 6 denotes the signal peak of the c-sapphire substrate. Therefore, the weight distribution should be zero over ID 228. The divergence result was consistent with this assumption; however, the MSE result had non-zero values over ID 228.

Regarding the generative model, the choice of divergence or MSE as the objective function corresponds to the assumption V_\| \sim \text{Poisson}\left(\left( WH \right)_\| \right), or V_\| \sim \text{Normal}\left(\left( WH \right)_\| \right), respectively [17,35]. Considering the diffraction mechanism [36], the Poisson distribution is suitable for the mechanism, and we estimate that this is the reason why the result of divergence was better than that of the MSE. However, the distribution of the diffraction intensity seemed more likely to be an exponential distribution rather than a Poisson distribution (Figure S14). Therefore, applying the corresponding training algorithm will improve the results. Hereafter, our analyses were based on the divergence results.

### 3.1.3. Analysis of single samples and fabrication conditions

Figure 4 shows the results of the In2O3 thin films on the c-sapphire substrates at substrate temperatures of 300, 400, and 500 °C from left to right. The other

| Factor number | Representing diffraction signal |
|---------------|---------------------------------|
| 1             | Background noise                |
| 2             | Diffraction peak at 35°         |
| 3             | YSZ (100) or In2O3 (400)       |
| 4             | High oriented In2O3 (222)       |
| 5             | Pt (111)                        |
| 6             | c-Sapphire substrate            |
| 7             | Lower oriented In2O3 (222)      |
| 8             | Unknown                         |
| 9             | Unknown (Peak shift?)           |
| 10            | Unknown [Weak signal of Ga2O3 (201) and Ga2O3 (401)] |
fabrication conditions were identical. The diffraction signals in the rightmost part of Figure 4(a) represent In$_2$O$_3$ (222), and evidently the crystal becomes more oriented in accordance with the increase in substrate temperature (the arc-like signal becomes a spotty signal). Images approximated by NMF [Figure 4(b)] represent this tendency through relative intensity changes among the peaks. Note that the absolute value is meaningless because all the signal intensities were normalized by the maximum of each data point.

Figure 3. Comparison of objective functions. (a), (d), and blue lines are the results of NMF whose objective function was divergence. (b), (e), and green lines are the results of MSE. (a) and (b) are the images of factor 5, and (d) and (e) are those of factor 6. (c) and (f) are the corresponding weight distributions. The borders of the substrate types are represented by black lines.
The leftmost spot in the left and middle images in Figure 4 (b) was the effect of factor 7 because the factor contained two spots in addition to the lower oriented In$_2$O$_3$ (222). This is a limitation of the method and will be solved by increasing number of factors or applying of better training algorithms [9,31,32]. The feature vectors represented the relative change of In$_2$O$_3$ (222) [Figure 4 (c)] as follows: Factor 3 [In$_2$O$_3$ (222)] became stronger at 400 °C and 500 °C; however, factor 1 (background noise) became weaker. Factor 3 was weaker at 500 °C than at 400 °C because the diffraction signal at 500 °C was sharp and resulted in a small peak region in the original 2D-XRD image. We estimate that this small region of the diffraction peak was recognized as ‘weak’ by NMF. Factor 6 (c-sapphire substrate) was approximately zero at 400 °C and 500 °C, which corresponded to the fact that the diffraction intensity of the substrate was relatively weak at these temperatures. Although the arc-like diffraction signal of low-oriented In$_2$O$_3$ was insignificant in the approximated image at 300 °C, corresponding factor 7 had a nontrivial value. Further, the feature vectors represented an important propensity at 300 °C that the diffraction signal of In$_2$O$_3$ (400), which was measured at approximately 35°, almost the center of the 2D-XRD image, by the non-trivial value of factor 2. Considering these results, NMF recognized diffraction signal changes according to the fabrication conditions, although it is sometimes difficult for humans to recognize small changes.

Figure 5 shows the XRD patterns and a heatmap of feature vector components of an (In$_{x}$Ga$_{1-x}$)$_2$O$_3$ composition-spread sample fabricated on a c-sapphire substrate at 400 °C. This sample was reported in a previous study [23]. This sample contained a solid solution around positions 8 and 9, resulting peak shift around 30.5°. NMF detected this shift and represented it by a stronger factor 6 at position 8 and emerging factor 3 from position 9. Factor 6 basically represented the diffraction pattern of the c-sapphire substrate; thus, the values of factor 6 of the sample were above

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{Results of three In$_2$O$_3$ thin films at substrate temperatures of 300 °C (left), 400 °C (middle), and 500 °C (right). (a) Original 2D-XRD images. (b) 2D-XRD images approximated by NMF. (c) Heatmap of the feature vector components.}
\end{figure}
0.6 over all positions. At position 8, the diffraction signal of the solid solution became stronger and the weight of factor 6 was at its maximum. As shown in Figure 2 (6), factor 6 contained slight diffraction pattern of the solid solution of In$_2$O$_3$ and Ga$_2$O$_3$ around 30.5°. Consequently, factor 6 had stronger intensity at position 8. Because the diffraction signal of In$_2$O$_3$ (222) was intense at and after position 9, factor 3 had significant values. However, at position 7, NMF did not detect a slight peak around 36°.

### 3.2. Combination of NMF and VAE

#### 3.2.1. Feature extraction by NMF

We compared two VAE models to evaluate whether feature extraction by NMF improved the output of the VAE models. One VAE model learned directly from the 2D-XRD images, and the other learned the feature vectors. Figure 6(a) and (b) show the distributions of the latent variables in latent spaces. These distributions were obtained by encoding the corresponding datasets. The latent variable distribution of direct VAE
application apparently failed to separate the samples based on the substrates. In addition, its arc-like distribution appeared to be affected by a manifold of the original images. In contrast, the distribution encoded by the VAE model that learned the feature vectors was better because some points were distributed separately. In particular, the points of YSZ (100) and STO (100) substrates were separated from other points, which was consistent with the fact that these substrates are different from the other substrates in terms of lattice mismatch between In₂O₃ and the substrates. The narrow dispersion in the horizontal axis indicates that 2D-XRD images in the dataset were differentiated mainly by a single cause. We estimate that the cause is the crystallinity of In₂O₃. Differences in substrate types may be a minor cause that separated the latent variables of samples on YSZ (100) or STO (100) substrates from the major linear distribution. Although the dispersion in the horizontal axis can be enlarged by adjustment of coefficients of the loss function of the VAE, we did not adjust because of the comparison of the two VAEs. The mixture region of the c-sapphire and YSZ (111) substrates was due to the diffraction signals of In₂O₃ (222) on the c-sapphire substrate and those of the YSZ (111) substrate were represented by factor 3 and located at close angles of 30.5° and 30.1°. This slight difference could not be differentiated based on the resolution of our measurement setup. Considering the previous research [25], we conclude that, with small dataset or high-dimensional data, feature extraction is important to improve the performance of DL.

Figure 6. Latent variable distributions of the VAE models that learned (a) the 2D-XRD images directly and (b) feature vectors. (c) and (d) are decoded images from the stars in (a) and (b), respectively.
The major advantage of VAE is that it can generate new data through sampling from its learned probability distribution. To demonstrate this, we sampled one point from each dense region [stars in Figure 6 (a) and (b), respectively] and then decoded with the decoders [Figure 6 (c) and (d)]. The image decoded by the directly learned VAE model contained mainly three peaks, which were the diffraction signals from the c-sapphire substrate (left peak), YSZ (100) substrate or In2O3 (400) (middle peak), and In2O3 (222) (right peak). This corresponded to the distribution where the data for the c-sapphire and YSZ (100) substrates were close. Note that the background noise seemed intense because of the reduction in image size. The image decoded by the VAE, which learned the feature vectors, contained mainly three diffraction peaks, which appeared from the c-sapphire substrate, Pt electrode, and In2O3 (222). The output of the decoder was in the form of feature vectors. The decoded image was calculated by multiplying W with the output feature vector.

We noticed that the distributions were unstable and changed their shapes over the trials. This could be due to the over-reduced dimensionality; therefore, increasing the number of factors or the dimensionality of latent vectors would stabilize the distribution. Nevertheless, NMF is a good dimensionality reduction method because the number of factors is the same as the amount of data at the maximum, which is smaller than the dimensionality of the original 2D-XRD images.

The decoder of the directly learned VAE model generated realistic 2D-XRD images. This was because the number of significant signals in the 2D-XRD images used here was significantly smaller than the dimensionality of the 2D-XRD images and relatively close to the dimensionality of the latent variables. This implies that the benefit of feature extraction falls mainly in encoders. However, our objective was to relate the fabrication conditions to the distribution in the latent space, which is discussed in the following section. If we could relate this, we could improve fabrication conditions with visual information. Therefore, decoder was less important here.

### 3.2.2. Coordinate system of latent space and process improvement

In this section, we attempt to relate the coordinate system of the latent space to the fabrication conditions. First, we visualized the axes of the factors in the latent space by the method explained in Section 2.4 [Figure 7 (a)]. Each line represents the intensity of the XRD patterns shown in Figure 2. Figure S15 shows the only lines and their origin. We noticed that the values of some factors were sensitive to the position in the latent space; however, some factors, especially factor 1, were not sensitive to their position in the latent space. This is because factor 1 represents background noise, and the intensity of this factor produces a slight difference in the 2D-XRD signals. The axis of factor 4 was drawn to the upper left corner; therefore, some data of the c-sapphire substrate were distributed in this region [Figure 6 (b)]. No axes were drawn to the upper right or bottom left, which implies that all the factors were correlated to each other, although factor 2 [YSZ (100)] can be the exception.

To evaluate the relationship between the fabrication conditions and the crystallinity of In2O3, we selected the data of the samples that were fabricated on c-sapphire substrates. All the components except factors 3 and 7 of the feature vectors of these data were set to 0. We encoded the modified feature vectors and colored them based on the fabrication conditions [Figure 7 (b) and (c)]. Samples of points located on lines indicate pure phase of factor 3 or 7 and the others indicate mixed phase. The results indicate that In2O3 tended to be highly oriented under the fabrication condition that the laser intensity was 70 mJ and O2 pressure was 0.1 mTorr regardless to other fabrication conditions such as substrate temperature. This was consistent with the fact that the higher laser intensity and lower O2 pressure enhanced the migration of In2O3 and its crystal growth. This result suggests that the visualization with VAE will support to compare fabrication conditions.

The fact that some axes were drawn into similar angles implies that both the dimensionality of the latent space, 2, and the variation of the dataset were small compared to the number of factors. Although this suggests high bias in the dataset, the suggestion will not weaken our study because high bias is common in all practical research focusing certain compositions. The reason for the number of factors being 10 was that we interpreted the factors by humans, and the evaluation of NMF as the feature extraction method of 2D-XRD signals was our main concern. In further studies, the number of factors and the dimensionality of latent variables should be larger, and the fabrication conditions should be related to the coordinate system of the latent space.

Note that relating the fabrication conditions to the intensities of the factors of NMF, excluding VAE, is simpler. We related them to the coordinate system of the latent space for the following two reasons. First, our focus was to evaluate whether the effects of the fabrication conditions on the 2D-XRD signals were preserved in the latent space. Second, we wanted to generate new data under new fabrication conditions without practically fabricating new samples. Therefore, VAE was an appropriate generative model, and relating fabrication conditions to the coordinate system of latent space was inevitable.
3.3. Evaluation of our workflow to new data

Finally, we evaluated NMF in terms of its inference of the feature vectors of the new data. To test inference ability of the factors, we prepared 2D-XRD images of the thin films of Ga$_2$O$_3$, In$_2$O$_3$, and (In$_x$Ga$_{1-x}$)$_2$O$_3$ composition spreads fabricated by sputtering. One composition-spread sample was fabricated on an STO (111) substrate, and the others were fabricated on a c-sapphire substrate. Sputtering is a lower kinetic method; therefore, the signals of samples in 2D-XRD images tended to be larger than those of PLD.

Figure 7. (a) Coordinates of the factors and the distribution of the feature vectors. (b) and (c) are the distribution of the feature vectors of the c-sapphire samples, whose components were set to zero except factor 3 and factor 7. The markers are colored based on the fabrication conditions.
Figure 8. (a) Heatmap of the feature vectors of the samples fabricated by sputtering. (b) Changes in factors 3 and 7 in two composition-spread samples (1–11 and 12–22). (c) The original 2D-XRD image at 1 in (b). (d) The corresponding image at 12 in (b). (e) The distribution of the sputtering data in the latent space. The latent space is the same as in Figure 6(b) and Figure 7.
A heatmap of the extracted features [Figure 8(a)] shows that the diffraction peaks of c-sapphire substrates were represented by factors 4, 8, and 9. It was estimated that all data were measured with the Ni-filter referring to the measurement dates, and factors 8 and 9 contained a slight signal of the substrate. Feature vectors that had significant value in factor 5 (ID 218–228) were from the composition-spread sample fabricated on the STO (111) substrate.

To evaluate this in detail, we compared the results of two composition-spread samples of Bi doped In$_2$O$_3$ (doped amount changed from 0% to 15%) [Figure 8 (b), (c), and (d)]. One sample [1–11 in Figure 8 (b)] contained an intense diffraction peak of In$_2$O$_3$ (222), and the shape was slightly arc-like [Figure 8 (c)]. This was consistent with the feature vector, where factors 3 and 7 were non-zero. In contrast, the other sample had a low-oriented In$_2$O$_3$ (222) crystal [Figure 8 (d)]; thus, factor 3 was approximately zero. Note that, in multiplication with the pseudo-inverse matrix, non-negativity was not satisfied.

Further, we encoded the extracted feature vectors with the encoder that learned the feature vectors of the PLD samples, and Figure 8 (e) shows the results. The majority of the data of c-sapphire was distributed around (~ 0.01, 2), because factor 4, and not factor 6, was the main factor with the sputtered samples. The STO (111) data were distributed almost close to the STO (111) region fabricated by PLD. The sputtering distribution appeared to be wider than that of PLD. This reflects the fact that sputtering can fabricate more variate crystal structures and oriented structures with the n-Ga-O system than PLD. Therefore, we conclude that NMF worked as a feature extraction method against new data, and our workflow visualized the differences in the fabrication methods, PLD, and sputtering.

4. Conclusion

In this study, we confirmed that NMF is an appropriate feature extraction method for 2D-XRD data to improve VAE performance. Although we only evaluated VAE, the extracted features will improve performance of other ML or DL techniques. The results demonstrated in this study were based on basic algorithms and models. Therefore, applying better algorithms and models will improve the results. In addition, the results can be improved by normalizing the data based on background noise level or diffraction peak intensities of substrates. To achieve this, noise separation techniques such as DBSCAN [37,38] can be useful. The number of factors appeared slightly smaller to capture slight changes in 2D-XRD images here; therefore, when applying NMF without factor interpretation, the number of factors should be larger. In this case, the dimensionality of the latent space of the VAE should be of moderate size for interpretation. Even under limited conditions, NMF showed its ability to extract features with significantly reduced dimensionality.

We also found that NMF extracted features of new data using a pseudo-inverse matrix. Feature extraction using a pseudo-inverse matrix will be useful in automated analysis or a search system. Although we generated the matrix using a data-driven method, calculating the pseudo-inverse matrix with reference 2D-XRD data will be more reliable and accurate.

The workflow we presented excluded random processes; therefore, our results were deterministic and reproducible. In addition, the samples we used contained a relatively small variance of fabrication conditions compared to the number of samples in terms of fabrication process optimization. Considering these, our results indicate the robustness of the workflow against error data, and further analysis with high-variant data will reveal more insights. The samples in this study were polycrystalline thin films. Although other structures such as bulk or amorphous were not tested, our workflow will be applicable to those samples with image-like measurement data.

Acknowledgments

This work was supported by the JST-Mirai Program “Development of Materials Design Workflow and Data Library for Materials Foundry,” Grant No. JPMJMI18G5.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability statement

The data that support the findings of this study are available from the corresponding author, T. N., upon reasonable request.

Funding

This work was supported by the Japan Science and Technology Agency [JPMJMI18G5].

ORCID

Akihiro Yamashita http://orcid.org/0000-0001-5975-9044
Takahiro Nagata http://orcid.org/0000-0002-8591-2943
Shinjiro Yagyu http://orcid.org/0000-0002-9825-5719
Toyohiro Chikyow http://orcid.org/0000-0003-3860-4806

References

[1] Ramakrishnan R, Dral PO, Rupp M, et al Quantum chemistry structures and properties of 134 kilo molecules. Sci Data. 2014;1(1):140022.
[2] Ruddigkeit L, van Deursen R, Blum LC, et al. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J Chem Inf Model. 2012;52(11):2864–2875.

[3] Burger B, Maffettone PM, Gusev VV, et al. A mobile robotic chemist. Nature. 2020;583(7815):237–241.

[4] Kukuruznyak DA, Ahmet P, Yamamoto A, et al. Combinatorial fabrication and characterization of Ternary La 2 O 3–Mn 2 O 3–Co 3 O 4 composition spreads. Ipn J Appl Phys. 2005;44(8):6164–6166.

[5] Ahmet P, Yoo Y-Z, Hasegawa K, et al. Fabrication of three-component composition spread thin film with controlled composition and thickness. Appl Phys A. 2004;79(4–6):837–839.

[6] Kusne AG, Gao T, Mehta A, et al. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. Sci Rep. 2014;4:6367.

[7] Ament SE, Stein HS, Guevara D, et al. Multi-component background learning automates signal detection for spectroscopic data. Npj Comput Mater. 2019;5(1):77.

[8] Ning X, Selesnick IW, Duval L. Chromatogram baseline estimation and denoising using sparsity (BEADS). Chemometr Intell Lab Syst. 2014;139:156–167.

[9] Suram SK, Xue Y, Bai J, et al. Automated phase mapping with AgileFD and its application to light absorber discovery in the V–Mn–Nb oxide system. ACS Comb Sci. 2017;19(1):37–46.

[10] Ziletti A, Kumar D, Scheffler M, et al. Insightful classification of crystal structures using deep learning. Nat Commun. 2018;9(1):2775.

[11] Alfeld M, Wahabzad M, Bauckhage C, et al. Non-negative matrix factorization for the near real-time interpretation of absorption effects in elemental distribution images acquired by X-ray fluorescence imaging. J Synchrotron Radiat. 2016;23:579–589.

[12] Stanév V, Vesselinov VV, Kusne AG, et al. Unsupervised phase mapping of X-ray diffraction data by nonnegative matrix factorization integrated with custom clustering. Npj Comput Mater. 2018;4(1):43.

[13] Long CJ, Hattrick-Simpers J, Murakami M, et al. Rapid structural mapping of ternary metallic alloy systems using the combinatorial approach and cluster analysis. Rev Sci Instrum. 2007;78(7):072217.

[14] Long CJ, Bunker D, Li X, et al. Rapid identification of structural phases in combinatorial thin-film libraries using x-ray diffraction and non-negative matrix factorization. Rev Sci Instrum. 2009;80(10):103902.

[15] Taheri ML, Stach EA, Arslan I, et al. Current status and future directions for in situ transmission electron microscopy. Ultramicroscopy. 2016;161(170):86–95.

[16] He BB. Two-dimensional X-ray diffraction. Hoboken (NJ): Wiley & Sons.; 2018.

[17] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401(6755):788–791.

[18] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35:1798–1828.

[19] Schmidt J, Marques MRG, Botti S, et al. Recent advances and applications of machine learning in solid-state materials science. Npj Comput Mater. 2019;5(1):83.

[20] Bartók AP, Kondor R, Csányi G. On representing chemical environments. Phys Rev B Condens Matter. 2013;87(18):184115.

[21] Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett. 2018;120:145301.

[22] Himanen L, Jäger MOJ, Morooka EV, et al. DScribe: library of descriptors for machine learning in materials science. Comput Phys Commun. 2020;247:106949.

[23] Nagata T, Hoga T, Yamashita A, et al. Valence band modification of a (Ga x In 1–x) 2 O 3 solid solution system fabricated by combinatorial synthesis. ACS Comb Sci. 2020;22(9):433–439.

[24] Kingma DP, and Welling M. Auto-encoding variational Bayes. Proceedings of 2nd International Conference on Learning Representations, ICLR 2014, Banff, Canada. 2013.

[25] Banko L, Maffettone PM, Naujoks D, et al. Deep learning for visualization and novelty detection in large X-ray diffraction datasets. Npj Comput Mater. 2021;7(1):104.

[26] Lee I-H, Chang KJ. Crystal structure prediction in a continuous representative space. Comput Mater Sci. 2021;194:110436.

[27] Noh J, Kim J, Stein HS, et al. Inverse design of solid-state materials via a continuous representation. Matter. 2019;1(5):1370–1384.

[28] Gómez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4(2):268–276.

[29] Bezanson J, Edelman A, Karpinski S, et al. Julia: a fresh approach to numerical computing. SIAM Rev. 2017;59:65–98.

[30] Boutsia C, Gallopoulos E. SVD based initialization: a head start for nonnegative matrix factorization. Pattern Recognit. 2008;41(4):1350–1362.

[31] Hsieh C-J, and Dhillon IS. Fast coordinate descent methods with variable selection for non-negative matrix factorization. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11, San Diego, California. New York (NY): ACM Press; 2011. p. 1064.

[32] Cichocki A, Phan A-H. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. IEICE Trans Fundam Electron Comput Sci. 2009;E92-A:708–721.

[33] Innes M. Flux: elegant machine learning with Julia. J Open Source Softw. 2018;3(25):602.

[34] Cichocki A, Zdunek R, Phan AH, et al. Nonnegative matrix and tensor factorizations. Chichester: Wiley & Sons; 2009.

[35] Virtanen T, Taylan Cemgil A, and Godsell S. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. Proceedings of 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, Las Vegas, Nevada; 2008. p. 1825–1828.

[36] Kittel C. Introduction to solid state physics. 8th ed. Hoboken (NJ): Wiley; 2004.

[37] Ester M, Kriegel H-P, and Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Oregon. Portland, OR: AAAI Press; 1996. p. 226–231.

[38] Yamashita A, Nagata T, Yagyu S, et al. Accelerating two-dimensional X-ray diffraction measurement and analysis with density-based clustering for thin films. Jpn J Appl Phys. 2021;60:SCCGO4.