Development of Data-Driven Representation Methods for Microstructures of Inorganic Thin Films with Two-Dimensional X-Ray Diffraction

2次元X線回折測定データを用いた無機半導体薄膜微細構造の特徴量抽出手法の開発

February, 2022

Akihiro YAMASHITA
山下　晶洸

Development of Data-Driven Representation Methods for Microstructures of Inorganic Thin Films with Two-Dimensional X-Ray Diffraction

2次元X線回折測定データを用いた無機半導体薄膜微細構造の特徴量抽出手法の開発

February, 2022

Waseda University Graduate School of Advanced Science and Engineering

Department of Advanced Science and Engineering, Research on Life Science and Medical Bioscience

Akihiro YAMASHITA
山下　晶洸

Contents

1. General Introduction 1

1.1 Introduction 1

1.2 Terminology of this Thesis 4

1.3 Outline of this Thesis 7

References 8

2. Feature Extraction of Two-Dimensional X-Ray Diffraction Images with Non-Negative Matrix Factorization 13

2.1 Introduction 13

2.2 Method 14

2.2.1 Dataset and Handling 14

2.2.2 NMF 16

2.2.3 Feature Extraction of New Data 17

2.3 Results and Discussion 18

2.3.1 Analysing Factor Images 18

2.3.2 Discussing hyperparameters 22

2.3.3 Relationship among Feature Vectors and Compositions 31

2.3.4 Relationship among Feature Vectors and Fabrication Conditions 33

2.3.5 Feature Extraction of New Data 35

2.4 Conclusion 37

References 37

3. Continuous Representation of Microstructures of Thin Films Fabricated under Multiple Conditions 41

3.1 Introduction 41

3.2 Method 42

3.2.1 Variational AutoEncoder 42

3.2.2 Visualization of NMF Factor Axes in the Latent Space 43

3.3 Results and Discussion 43

3.3.1 Evaluation of Feature Extraction by NMF 43

3.3.2 Coordinate System of Latent Space and Process Improvement 46

3.3.3 New Data in Latent Space 49

3.4 Conclusion 50

References 51

4. Fabrication Analyses of Indium Gallium Oxide Thin Films with Machine Learning 53

4.1 Introduction 53

4.2 Method 54

4.3 Results and Discussion 54

4.4 Conclusion 61

References 62

5. Density-Based Analysis of Two-Dimensional X-Ray Diffraction Images 63

5.1 Introduction 63

5.2 Method 65

5.2.1 2D-XRD Images 65
5.2.2 Conversion of 2D-XRD Images into Scatter Plots 65
5.2.3 Density-Based Analyses with DBSCAN and OPTICS 65
5.2.4 Hyperparameters of DBSCAN 67
5.2.5 Graph Representation of 2D-XRD Images 68
5.3 Results and Discussion 69
5.3.1 Density Analyses of 2D-XRD Images with OPTICS 69
5.3.2 Signal/Noise Separation with DBSCAN 73
5.3.3 Evaluation of measurement time 79
5.3.4 Remarks for Practical Applications 82
5.4 Future Work 82
5.5 Conclusion 84
References 85

6. Conclusions 87
6.1 Conclusion of the research 87
6.2 Outlook 88
6.3 Industrial Perspective 89

Appendix A 91
A.1 Feature Vector Overview with a Correlation Matrix 91
A.2 Supplemental Figures 95

Appendix B 107
B.1 Supplemental Figures 107

Acknowledgements 108

Chapter 1

General Introduction

1.1 Introduction

Demand for advanced materials under the banners of green transformation, Industry 4.0, or Society 5.0 have been growing at accelerated speed, yet the pace of material discovery from conventional research has not matched the growth in demand. One promising solution to meet the acceleration in demand is data-driven materials research. This emerging research field has been advanced by national projects such as the Materials Genome Initiative (USA), the Novel Materials Discovery (Europe), and the Material Research by Information Integration Initiative (Japan). These projects have cultivated multiple research areas in data-driven materials research [1]. Furthermore, they have promoted open databases of *ab initio* calculations managed under Findable, Accessible, Interoperable, Reusable (FAIR) treatments [2] such as DICE (formerly named Materials Data Repository [3]), Materials Project [4], or QM9 [5]. These databases have enabled numerous laboratories to enter the data-driven materials research field.

In addition to national projects, high-throughput experiments with fast data-acquisition, such as robot-automated experiments [6] or combinatorial synthesis [7–12], have produced large volumes of data. Moreover, high temporal and spatial resolution equipment can produce measurement data that are too large for feasible analysis using conventional methods [13,14]. Thus, to utilize these rapidly produced data, this thesis focuses on methods for inputting raw measurement data into machine learning models.

Applications of machine learning models to raw measurement data have been investigated in automated analysis frameworks (e.g., [15–21]). Because their research focus is automated analysis, few predict novel materials or unmeasured properties from raw measurement data [20,21]. A practical reason for this both inside and outside the automated analysis framework is that few raw measurement databases exist. In open databases, generally, the size of a single data point is smaller than 1 MB, whereas, in materials science, it is often over 1 MB (for example, if an image of transmission electron microscopy is recorded at a resolution of 1024×1024 pixels, then the data size will be over 1 MB). This means that a large database of raw measurement data can range from 10 GB to over 1 TB in size, which is too large for easy Internet distribution. Although such databases are not currently common, data-driven research using raw measurement data is nevertheless important for utilizing data produced by high-throughput experiments.

This thesis focuses on measurement data from two-dimensional X-ray diffraction (2D-XRD) [22], an appropriate technique for characterizing the microstructures of thin films. 2D-XRD measurements capture cross-sectional images of the diffraction cone and detection plane, where the diffraction angle (2θ) indicates the periodicities of atomic positions and the χ angle indicates the orientations of the periodicities (Figure 1.1). The χ angle is an advantage of 2D-XRD measurements compared to conventional XRD measurements, as the orientation of the periodicities is essential for understanding the microstructures of thin films. The properties of a thin film are determined by its microstructure such as dislocation of atoms or mosaic crystallinity as well as the element composition [21,23]. This makes difficult to predict properties of thin films by either *ab initio* calculation or predictive models. However, accurately characterizing the microstructures in thin films is time-consuming, which is one reason why data-driven research is not well developed for thin films as compared to other materials such as alloys or polymers [1,23]. Consequently, this thesis focuses on data-driven research on inorganic thin films using 2D-XRD measurements.

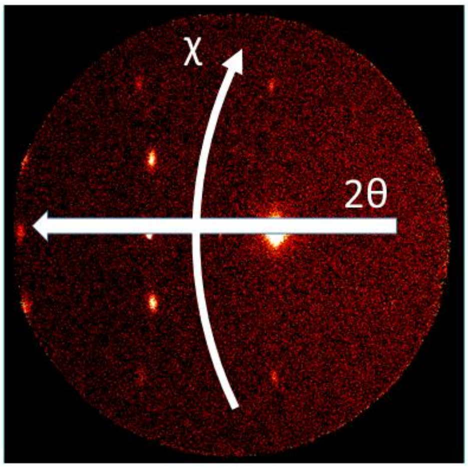

Figure 1.1: Example 2D-XRD image. Horizontal axis represents diffraction angle (2θ), indicating periodicities of atomic positions in sample. Arc-like axis represents χ angle, indicating orientations of periodicities.

Although design of predictive models is generally considered to be a main task in data-driven research, a representation of data, or feature, is actually the main task because it is the major factor that determines the performance of machine learning models [24]. Ideally, machine learning models would understand hidden mechanisms of their tasks using only the given datasets. However, currently machine learning models cannot achieve their tasks without human assistance; thus, the human design of features is needed to discover the hidden mechanisms and improve the accuracy of model predictions. Features have been intensively investigated in the field of representation learning. Neural networks have been a transformational method for this field because they can learn appropriate features based on tasks and given datasets [24]. However, this advantage is enabled by big data; although big data are becoming available to some areas of materials science, they do not yet cover all areas. Therefore, this thesis aims to address this problem by investigating features for limited volume of 2D-XRD images of thin films that are rapidly produced during high-throughput experiments.

1.2 Terminology of this Thesis

This section briefly explains terminology that may be unfamiliar to readers from either a materials science or computer science background. More detailed explanations can be found in the literature referenced and chapter specific terminology in the relevant chapter.

Materials science

- Composition spread: A combinatorial library that has a continuous gradient in composition [8–10,12]. Composition spread samples are important in high-throughput experiments because they serve a high density of data points on a single wafer. Figure 1.2 shows a schematic of a fabrication of the composition spread $A_{1-x}B_x$ by pulsed laser deposition, where the chemical formula $A_{1-x}B_x$ represents the target composition of the fabrication. The moving mask enables the concentration of the deposited materials to be changed as intended, and the addition of a target enables the fabrication of ternary composition spreads [10]. Because of the varying parameters between materials (diffusion speed on the substrate, nucleation speed, crystal-growth rate, etc.), an actual sample may be a mixed phase or solid solution depending on position x.

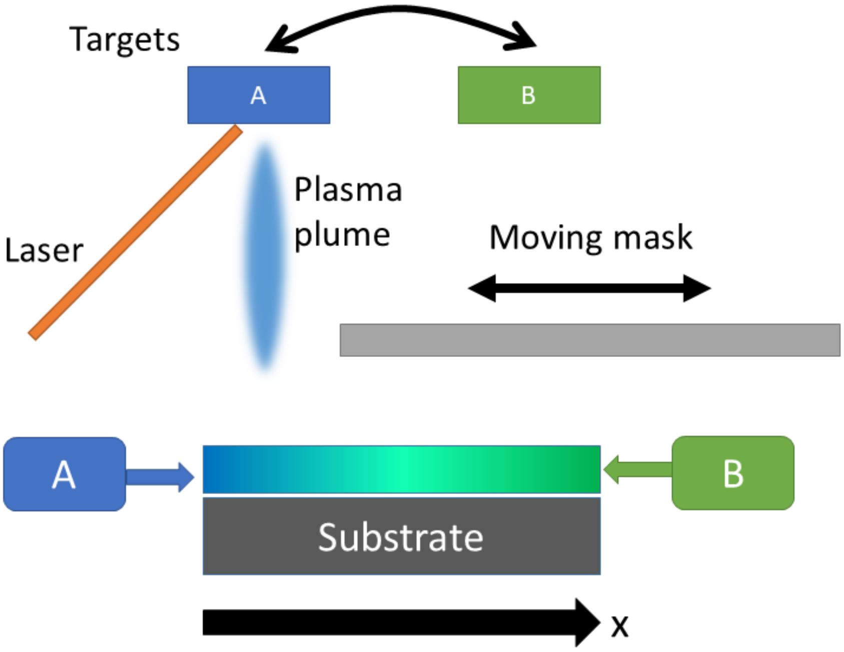

Figure 1.2 Schematic of fabrication of a composition spread of $A_{1-x}B_x$ $(0 \leq x \leq 1)$.

- Orientation of periodicities of atomic positions: Deposited atoms on the substrate constitute a structure depending on their preference and the deposition environment. The structure with no periodicity of atomic positions is called amorphous. Unless perfect amorphous, thin films contain structures with periodic atomic positions, or crystal. A thin film with a single crystal is called an epitaxial film. Most films contain multiple crystals and called polycrystalline films. If orientations of crystals are random, then the 2D-XRD image of the thin film possess broader diffraction patterns over χ angle. If orientations are almost the same, then the film is highly oriented and its 2D-XRD image possess spotty diffraction patterns.

Computer science

- Feature: An informative and lower-dimensional representation of data for machine learning models. If the feature is well designed, any models will predict accurately. How to represent is commonly determined by a data-driven manner with an unsupervised learning technique. The conversion of data into features is called feature extraction or dimensionality reduction. The choice or design of the unsupervised learning technique for the feature extraction is referred to as feature engineering. A feature is sometimes termed a "descriptor" in data-driven materials science.
- Latent variable and latent space: A latent variable is an intermediate representation of input data in a machine learning model. A latent space is where latent variables exist. The latent variables of neural network models are often used as features of the original data [21,23].
- Vector (data representation): One data point is represented by a single vector in computer programs; thus, a dataset is represented by a matrix (Figure 1.3). Some program languages refer to vectors as arrays, but to maintain consistency with mathematics, vector is the chosen term used in this thesis. As Julia was the programming language used for this thesis, all data points have been represented as column vectors. Julia is a column-major language [25], where column major means that data are stored in the column direction in the computer memory. Other program languages such as Python are row-major languages. The dimensionality of data is defined as the number of parameters of the data.

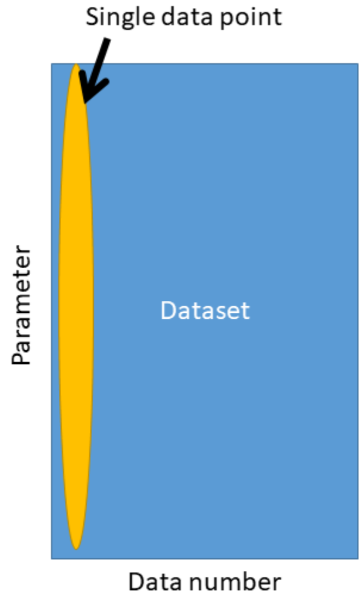

Figure 1.3 Schematic showing data storage in computer programs.

1.3 Outline of this Thesis

This thesis has six chapters. Chapter 1 introduces the background of data-driven materials research and its related problems, and describes the research focus of the thesis.

In Chapter 2, non-negative matrix factorization (NMF) is evaluated as a feature extraction method for 2D-XRD images. The dataset used contained 2D-XRD images from multiple samples fabricated under differing conditions. The NMF result is evaluated to determine whether it is an appropriate feature for representing differences in crystallinity under different fabrication conditions.

In Chapter 3, deep learning models called variational autoencoders (VAEs) are trained to evaluate whether the features extracted by NMF can improve the performance of the models. The latent spaces of the models are analyzed in terms of their relationships to the NMF features and to the sample fabrication conditions.

In Chapter 4, the fabrication conditions of indium gallium oxide thin films are analyzed in a latent space. The methods discussed in Chapters 2 and 3 are evaluated as visualization methods for the fabrication conditions of thin films.

In Chapter 5, the signal density of the 2D-XRD images is evaluated as another possible feature. The use of the density feature for optimizing the 2D-XRD measurement time is discussed. A graph representation of the 2D-XRD images is also discussed as another feature candidate.

Chapter 6 summarizes the findings of this thesis, and describes the implications and business impacts of this work.

References

[1] J. J. de Pablo, N. E. Jackson, M. A. Webb, L.-Q. Chen, J. E. Moore, D. Morgan, R. Jacobs, T. Pollock, D. G. Schlom, E. S. Toberer, J. Analytis, I. Dabo, D. M. DeLongchamp, G. A. Fiete, G. M. Grason, G. Hautier, Y. Mo, K. Rajan, E. J. Reed, E. Rodriguez, V. Stevanovic, J. Suntivich, K. Thornton, and J.-C. Zhao, *New Frontiers for the Materials Genome Initiative*, Npj Computational Materials **5**, 41 (2019).

[2] M. D. Wilkinson, M. Dumontier, Ij. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. ’t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J.

Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, *The FAIR Guiding Principles for Scientific Data Management and Stewardship*, Scientific Data **3**, 160018 (2016).

[3] M. Tanifuji, A. Matsuda, and H. Yoshikawa, *Materials Data Platform - a FAIR System for Data-Driven Materials Science*, in *2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI)* (IEEE, 2019), pp. 1021–1022.

[4] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, *Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation*, APL Materials **1**, 011002 (2013).

[5] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, *Quantum Chemistry Structures and Properties of 134 Kilo Molecules*, Scientific Data **1**, 140022 (2014).

[6] B. Burger, P. M. Maffettone, V. v. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick, and A. I. Cooper, *A Mobile Robotic Chemist*, Nature **583**, 237 (2020).

[7] H. Koinuma and I. Takeuchi, *Combinatorial Solid-State Chemistry of Inorganic Materials*, Nature Materials **3**, (2004).

[8] I. Takeuchi, O. O. Famodu, J. C. Read, M. A. Aronova, K.-S. Chang, C. Craciunescu, S. E. Lofland, M. Wuttig, F. C. Wellstood, L. Knauss, and A. Orozco, *Identification of Novel Compositions of Ferromagnetic Shape-Memory Alloys Using Composition Spreads*, Nature Materials **2**, 180 (2003).

[9] I. Ohkubo, H. M. Christen, P. Khalifah, S. Sathyamurthy, H. Y. Zhai, C. M. Rouleau, D. G. Mandrus, and D. H. Lowndes, *Continuous Composition-Spread*

Thin Films of Transition Metal Oxides by Pulsed-Laser Deposition, Applied Surface Science **223**, (2004).

[10] P. Ahmet, Y.-Z. Yoo, K. Hasegawa, H. Koinuma, and T. Chikyow, *Fabrication of Three-Component Composition Spread Thin Film with Controlled Composition and Thickness*, Applied Physics A **79**, 837 (2004).

[11] B. Fleutot, J. B. Miller, and A. J. Gellman, *Apparatus for Deposition of Composition Spread Alloy Films: The Rotatable Shadow Mask*, Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films **30**, (2012).

[12] H. von Wenckstern, Z. Zhang, F. Schmidt, J. Lenzner, H. Hochmuth, and M. Grundmann, *Continuous Composition Spread Using Pulsed-Laser Deposition with a Single Segmented Target*, CrystEngComm **15**, 10020 (2013).

[13] J. P. Horwath, D. N. Zakharov, R. Mégret, and E. A. Stach, *Understanding Important Features of Deep Learning Models for Segmentation of High-Resolution Transmission Electron Microscopy Images*, Npj Computational Materials **6**, 108 (2020).

[14] M. L. Taheri, E. A. Stach, I. Arslan, P. A. Crozier, B. C. Kabius, T. LaGrange, A. M. Minor, S. Takeda, M. Tanase, J. B. Wagner, and R. Sharma, *Current Status and Future Directions for in Situ Transmission Electron Microscopy*, Ultramicroscopy **170**, 86 (2016).

[15] A. G. Kusne, T. Gao, A. Mehta, L. Ke, M. C. Nguyen, K.-M. Ho, V. Antropov, C.-Z. Wang, M. J. Kramer, C. Long, and I. Takeuchi, *On-the-Fly Machine-Learning for High-Throughput Experiments: Search for Rare-Earth-Free Permanent Magnets*, Scientific Reports **4**, 6367 (2015).

[16] C. J. Long, D. Bunker, X. Li, V. L. Karen, and I. Takeuchi, *Rapid Identification of Structural Phases in Combinatorial Thin-Film Libraries Using x-Ray*

Diffraction and Non-Negative Matrix Factorization, Review of Scientific Instruments **80**, 103902 (2009).

[17] S. E. Ament, H. S. Stein, D. Guevarra, L. Zhou, J. A. Haber, D. A. Boyd, M. Umehara, J. M. Gregoire, and C. P. Gomes, *Multi-Component Background Learning Automates Signal Detection for Spectroscopic Data*, Npj Computational Materials **5**, 77 (2019).

[18] Y. Ozaki, Y. Suzuki, T. Hawai, K. Saito, M. Onishi, and K. Ono, *Automated Crystal Structure Analysis Based on Blackbox Optimisation*, Npj Computational Materials **6**, 75 (2020).

[19] 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, and J. M. Gregoire, *Automated Phase Mapping with AgileFD and Its Application to Light Absorber Discovery in the V–Mn–Nb Oxide System*, ACS Combinatorial Science **19**, 37 (2017).

[20] X. Li, Y. Zhang, H. Zhao, C. Burkhart, L. C. Brinson, and W. Chen, *A Transfer Learning Approach for Microstructure Reconstruction and Structure-Property Predictions*, Scientific Reports **8**, 13461 (2018).

[21] L. Banko, Y. Lysogorskiy, D. Grochla, D. Naujoks, R. Drautz, and A. Ludwig, *Predicting Structure Zone Diagrams for Thin Film Synthesis by Generative Machine Learning*, Communications Materials **1**, 15 (2020).

[22] B. B. He, *Two-Dimensional X-Ray Diffraction* (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2018).

[23] K. Alberi, M. B. Nardelli, A. Zakutayev, L. Mitas, S. Curtarolo, A. Jain, M. Fornari, N. Marzari, I. Takeuchi, M. L. Green, M. Kanatzidis, M. F. Toney, S. Butenko, B. Meredig, S. Lany, U. Kattner, A. Davydov, E. S. Toberer, V. Stevanovic, A. Walsh, N.-G. Park, A. Aspuru-Guzik, D. P. Tabor, J. Nelson, J.

Murphy, A. Setlur, J. Gregoire, H. Li, R. Xiao, A. Ludwig, L. W. Martin, A. M. Rappe, S.-H. Wei, and J. Perkins, *The 2019 Materials by Design Roadmap*, Journal of Physics D: Applied Physics **52**, 013001 (2019).

[24] Y. Bengio, A. Courville, and P. Vincent, *Representation Learning: A Review and New Perspectives*, IEEE Transactions on Pattern Analysis and Machine Intelligence **35**, 1798 (2013).

[25] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, *Julia: A Fresh Approach to Numerical Computing*, SIAM Review **59**, 65 (2017).

Chapter 2

Feature Extraction of Two-Dimensional X-Ray Diffraction Images with Non-Negative Matrix Factorization

2.1 Introduction

As described in Chapter 1, this thesis focuses on features of 2D-XRD images to represent microstructures of thin films. Although machine learning models have been applied to raw measurement data in automated analysis frameworks [1–8], few focused on features of raw measurement data. In contrast, in computer science, features of raw data have been well studied and those of image data is a main focus in the research field of computer vision [9–11]. These features and convolutional neural network (CNN) have produced a lot of advances which leads to automated driving and face recognition [12–14]. However, they are hard to be applied to raw measurement data in materials science [15] because of the following three problems.

- Dimensionalities of raw measurement data of materials are very high because of their high resolution and cannot be reduced for the accuracy in research
- Variations of signal patterns in measurement data, especially 2D-XRD images, are small (spot or arc-like only) and has ambiguous borders compared to those in computer vision
- Accuracy in positions of signal patterns are important, but it is difficult to detect them by CNNs

Therefore, this thesis studies features of 2D-XRD images.

Among the machine learning models in the literature, this chapter focuses on non-negative matrix factorization (NMF) [16]. NMF approximates a non-negative dataset matrix V by a product of two non-negative matrices (W and H), that is, $V \approx WH$. An advantage of NMF is its easiness to understand the result because each data point is represented by the linear combination of non-negative factors. NMF in the literature was applied to 1D-XRD datasets either of simulations or samples under identical fabrication conditions [3,6,17]. In this thesis, NMF is reevaluated by an application to 2D-XRD images of multiple samples fabricated under differing conditions. First, the NMF is evaluated whether the extracted features represented the propensities of the dataset. Relationships of the results to fabrication conditions are also analyzed. Finally, inference ability of NMF to a new dataset is evaluated.

2.2 Method

2.2.1 Dataset and Handling

The samples were thin films of In_2O_3, Ga_2O_3, and composition spreads of $(Ga_{1-x}In_x)_2O_3$ fabricated under multiple conditions (Table 2.1). These samples were investigated in another study [18], therefore what propensities should be represented by features are known. The original objective of the fabrication of these samples was to determine the appropriate fabrication conditions for In_2O_3 and Ga_2O_3 crystals and their solid solutions. Because the fabrication conditions were optimized by experts, possible combinations of fabrication conditions were not fully conducted.

Table 2.1 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 intensity. However, no YSZ (111) sample was fabricated at room temperature. The number of data was 512 for PLD samples and 272 for sputtering. This table is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis," Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

Parameters	Values
Fabrication process	Pulsed laser deposition (PLD), RF Sputtering
Substrate type	c-sapphire, yttria-stabilized zirconia (111) [YSZ (111)], YSZ (100), Pt/Ti/SiO_2/Si (100), $SrTiO_3$ (111) [STO (111)], STO (100)
Target	In_2O_3, Bi:In_2O_3, Ga_2O_3, Bi:Ga_2O_3
Laser intensity in PLD [mJ]	20, 40,70
O_2 pressure in PLD [Torr.]	1e-3, 1e-4, 1e-5
Radio frequency power in sputtering [W]	20, 35, 50
O_2/Ar ratio in sputtering [%]	5, 20, 40
Substrate Temperature [°C]	Room temperature, 300, 400, 450, 500, 600, 650, 700

The 2D-XRD images were obtained using a Discover D8 system and Vantec 500 (Bruker AXS). A typical image in this study contained diffraction signals of the substrate and the thin film, as well as background noise. On the measurements of the composition spreads $(Ga_{1-x}In_x)_2O_3$, 11 measurement points were assigned in accordance with their composition gradients so that each point represented an increase of x by 0.1 step from 0 to 1. The number of 2D-XRD images in this study was 512 for PLD samples and 272 for sputtered samples. The original image size of the 2D-XRD images was 2048 × 2048 pixels. All figures of the 2D-XRD images in this chapter were corrected at $\gamma = 5$ to improve readability.

Data handling, such as resizing or conversion, and calculations of NMF were conducted using the programming language JuliaLang [26] and its packages. Detailed explanations of the methods and packages are provided below.

2.2.2 NMF

Five hundred and twelve 2D-XRD images of PLD samples were normalized by dividing them with their maximum pixel values, then resized into 512 × 512 pixels to reduce computational time. Subsequently, the images were vectorized, and, consequently, the shape of dataset matrix V was a 262,144 × 512 matrix. The number of factors was set to 10 because this number was assumed to be large enough to represent all images and small enough for human to interpret factors. Each column of W is called a basis vector or an end member in other studies [3,16]; however, here, we called it a factor vector or a factor image. In this chapter, a factor vector is mainly called a factor image because it is represented in the image form. We referred to each column of H as a feature vector because it represents propensities of the original image. Each row of H was recognized as the weight distribution of the corresponding factor because it represents how important the corresponding factor is for each image.

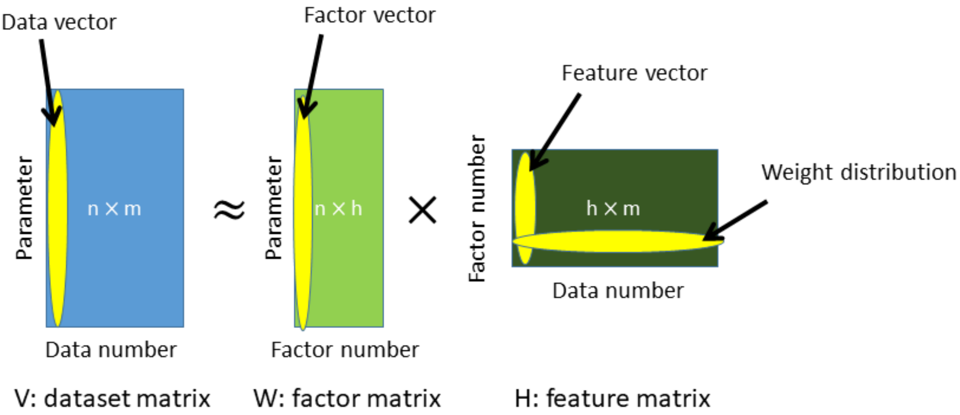

Figure 2.1 Schematic and terminology of NMF. In this section, n = 262,144, m = 512, h = 10.

Initialization and training algorithms were non-negative double singular value decomposition (NNDSVD) [19] and multiplicative update [16], 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, to confirm whether our results were deterministically correct or not. Although more sophisticated algorithms have already been proposed [20], we used the algorithms above because they are basic and easy to understand. The objective function and training times were divergence and 100, respectively.

2.2.3 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 extraction of feature vectors from V*. Consequently, H^* is a set of extracted feature vectors and W^{-1} is a feature extractor. In this chapter, W^{-1} was the Moore–Penrose pseudo-inverse matrix of W.

2.3 Results and Discussion

2.3.1 Analysing Factor Images

First, we applied NMF to 512 2D-XRD images of samples fabricated by PLD with setting the number of factor images to 10. Figure 2.2 shows the factor images of the results. These factor images were calculated ones; therefore, the exact image did not exist in the original dataset.

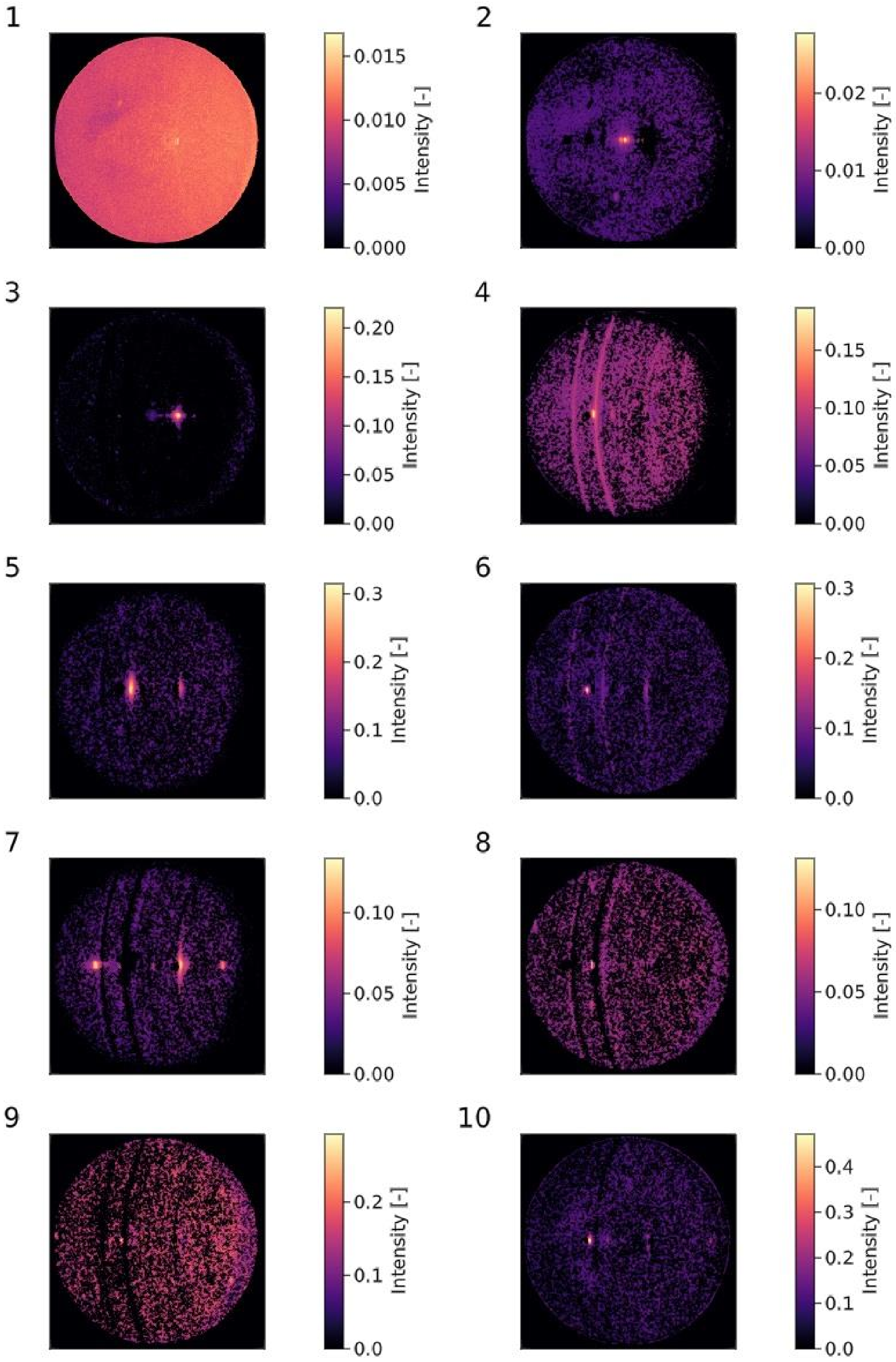
1
0.015
0.010
0.005
0.000
Intensity [-]
2
0.02
0.01
0.00
Intensity [-]
3
0.20
0.15
0.10
0.05
0.00
Intensity [-]
4
0.15
0.10
0.05
0.00
Intensity [-]
5
0.3
0.2
0.1
0.0
Intensity [-]
6
0.3
0.2
0.1
0.0
Intensity [-]
7
0.10
0.05
0.00
Intensity [-]
8
0.10
0.05
0.00
Intensity [-]
9
0.2
0.1
0.0
Intensity [-]
10
0.4
0.3
0.2
0.1
0.0
Intensity [-]

Figure 2.2 Ten factor images. All images are corrected at $\gamma=5$ to improve readability. Factor images with shared color scale is provided in Figure A2. Lighter color represents stronger intensity; in contrast, darker color represents weaker. The corners of images are invalid region because the shape of detector was circular. Heatmap of feature vectors of PLD samples are Figure A3. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis," Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

We interpreted each factor image with expertise and found that NMF learned mainly from positions and shapes of major diffraction peaks (Table 2.2). The first seven factor images were identified, and the remaining factors were categorized as unknown. The unknown factors contained diffraction signals that seemed to represent the c-sapphire substrate. However, referring to their weight distributions, we could not confirm that their signals corresponded to the c-sapphire substrate because their distributions possessed unreasonable weights on samples without the c-sapphire substrate (Figures A9-11). In this study, NMF failed to detect signals of Ga_2O_3, because they were very weak compared to those of the substrates and In_2O_3. Although factor 10 contained slight signals from 401 plane of Ga_2O_3 (vague signal in the most right), we could not conclude that factor 10 represented signals of Ga_2O_3 referring to the weight distribution. We found that both factors 4 and 6 represented the diffraction signal of the c-sapphire substrate, with a slight difference in the diffraction angle. This difference was 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 orientation of In_2O_3 (222), which is an advantage compared to the NMF application to 1D detectors. These results suggest that NMF is applicable to datasets of 2D-XRD images.

Table 2.2 List of representing diffraction signals of the factors. All factors were analyzed by identifying positions and shapes of diffraction peak and referring to fabrication conditions. This table is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis," Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

Factor number	Representing diffraction signal
1	Background noise
2	Diffraction peak at 35° [YSZ (100) or In_2O_3 (400)]
3	YSZ (111) substrate or High oriented In_2O_3 (222) on c-Sapphire substrate
4	Contamination or c-Sapphire substrate
5	Pt (111) electrode or STO (111)
6	c-Sapphire substrate
7	Lower oriented In_2O_3 (222)
8	Unknown
9	Unknown (Peak shift?)
10	Unknown [Weak signal of Ga_2O_3 (201) and Ga_2O_3 (401)]

2.3.2 Discussing hyperparameters

Before further analyses of NMF results, we discuss three hyperparameters, the number of factors, the training times and the objective function. We analyzed whether 10 was sufficient number of factors to represent the dataset by calculating errors between the original and the approximated matrices. The used error metrics were the logarithm of posterior probability, mean squared error, and divergence (Figure 2.3). Because the errors did not saturate over the number of factors ranging from 1 to 512, we concluded that there was no typical numbers to represent the dataset. Considering the trade-off between the error and interpretability of factors, 10 was concluded to be a better number of factors in this study. In another study, the number of factors will be determined by referring to errors or knowledge on the focused materials.

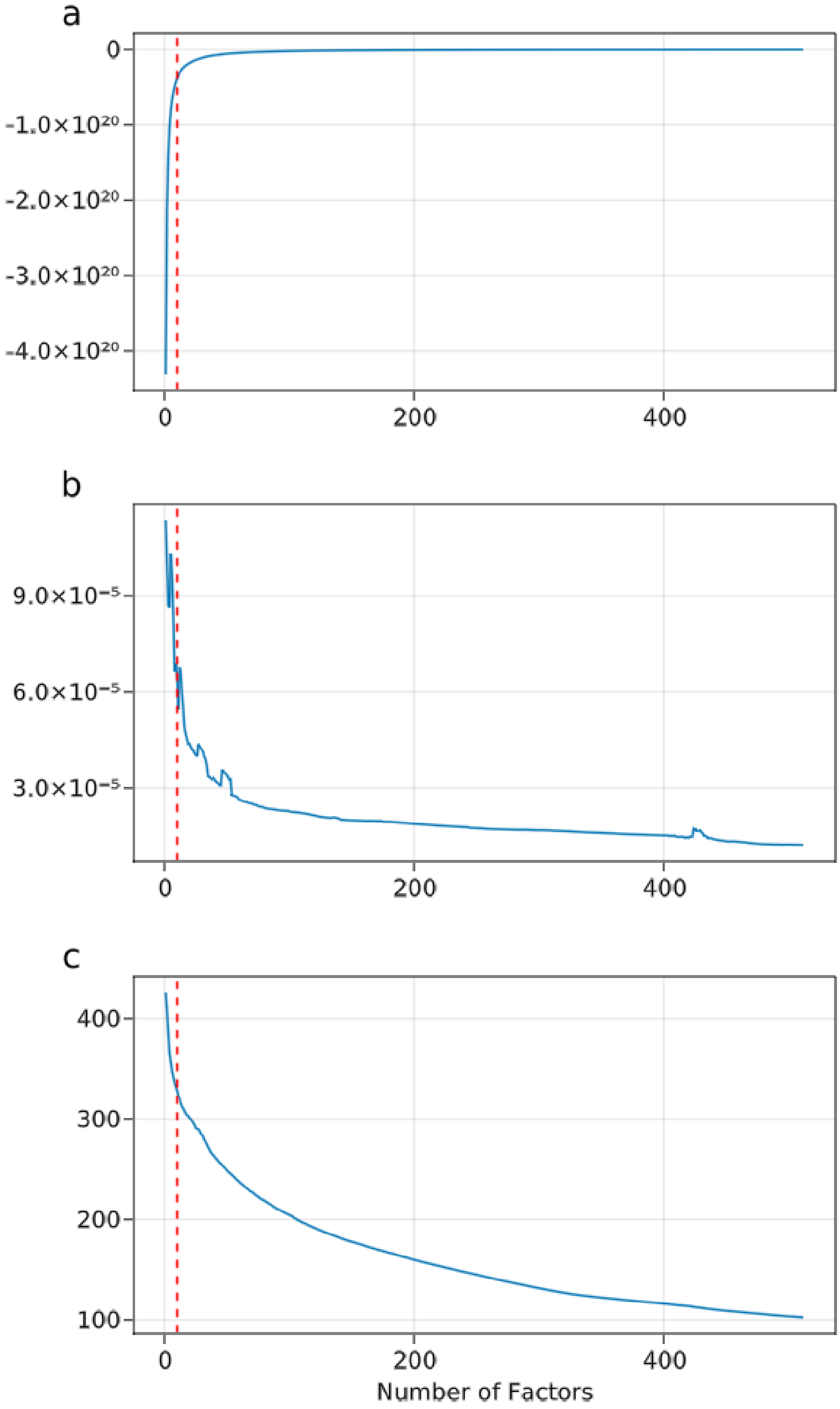

Figure 2.3 Three metrics changed over the number of factors. (a) logarithm of posterior probability, (b) mean squared error, and (c) divergence. The red dashed lines indicates that the number of factors was 10. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for

fabrication analysis," Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

In terms of appearance of the factor images, 10 may not be sufficient number because some factor images contained two or three signals (Figure 2.2). We estimated that this would be solved by increasing the number of factors. To confirm this, NMF was trained with setting the number of factors to 100. Then, we found that increasing the number of factors did not lead to mutually exclusive factor images (Figure 2.4) and some of first seven factors are similar to those of Figure 2.2. This result indicates that a larger number of factors is worse for the interpretability 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 [21], will improve the factor images, and this will be a future work. In this chapter, we continue analyses with 10 factors.

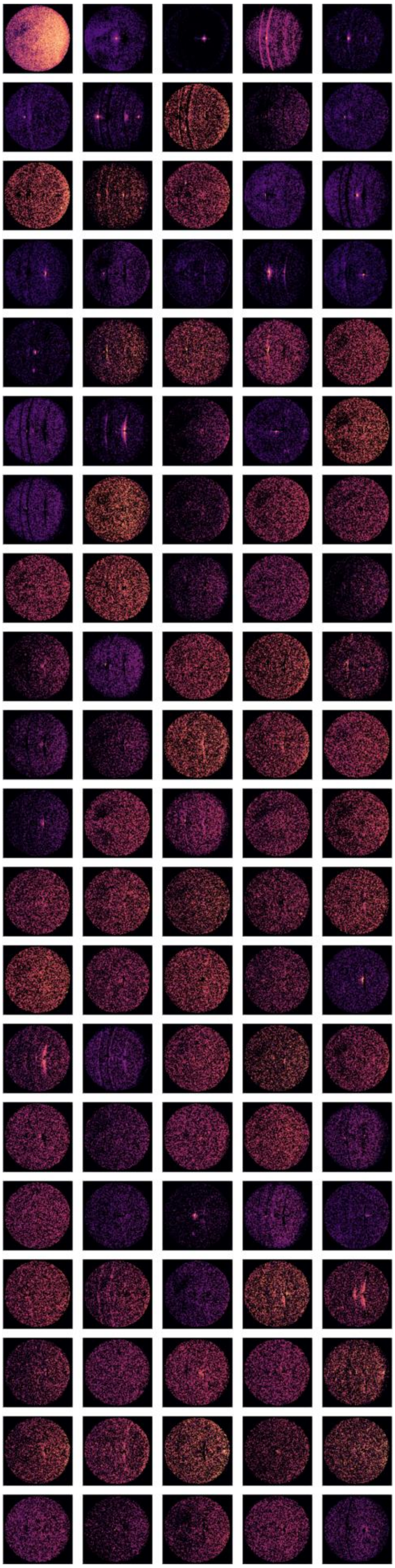

Figure 2.4 Factor images when the number of factors was set to 100. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis," Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

To evaluate whether 100 was enough training times or not, we checked the training process of NMF step by step (Figure 2.5), and found that the training results converged quickly. Therefore, the number of training (i.e., 100 times) was sufficient. Considering the convergence speed, we assume that the initialization result will be a good indicator for estimating the proper number of factors.

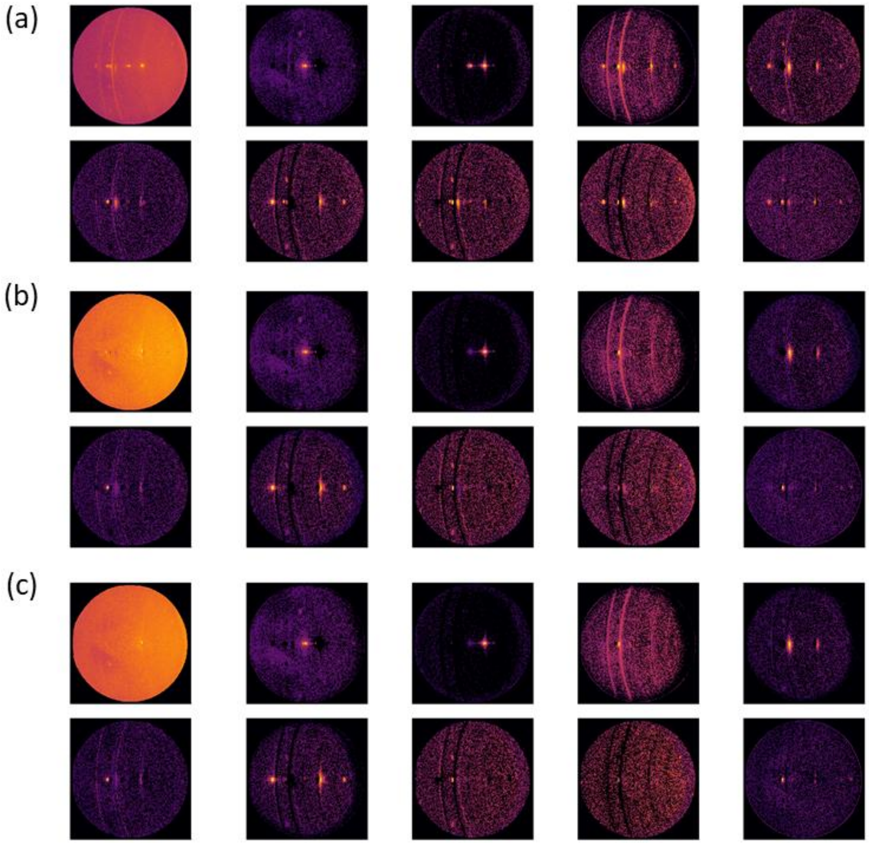

Figure 2.5 Factor images (a) at initialization, (b) after 10 updates, and (c) after 99 updates. This figure is a slightly modified reproduction under Creative Commons

License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis,” Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

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 mean squared error (MSE). Comparing the factor images and weight distributions, we noticed that the factor images were almost similar to the correspondences, but weight distributions were differed significantly. Here, we focus on factors 5 and 6 (Figure 2.6). Figure A4–A11 show the results of the other factors. Factor 5 represented the signal of the Pt electrodes on the Si substrate (Table 2.2). In this study, only two composition-spread samples were attached with the Pt electrodes (data number 480–490 and 491–501). The weight distribution of the divergence result was consistent. In contrast, that of the MSE result had unreasonable non-zero weights under data number 400. Factor 6 denoted the signal peak of the c-sapphire substrate, therefore the weight should be zero over data number 228 (over this number, all samples were on other substrates than c-sapphire). The divergence result was consistent; however, the MSE result had non-zero weights over data number 228. These differences can be discussed in terms of generative models.

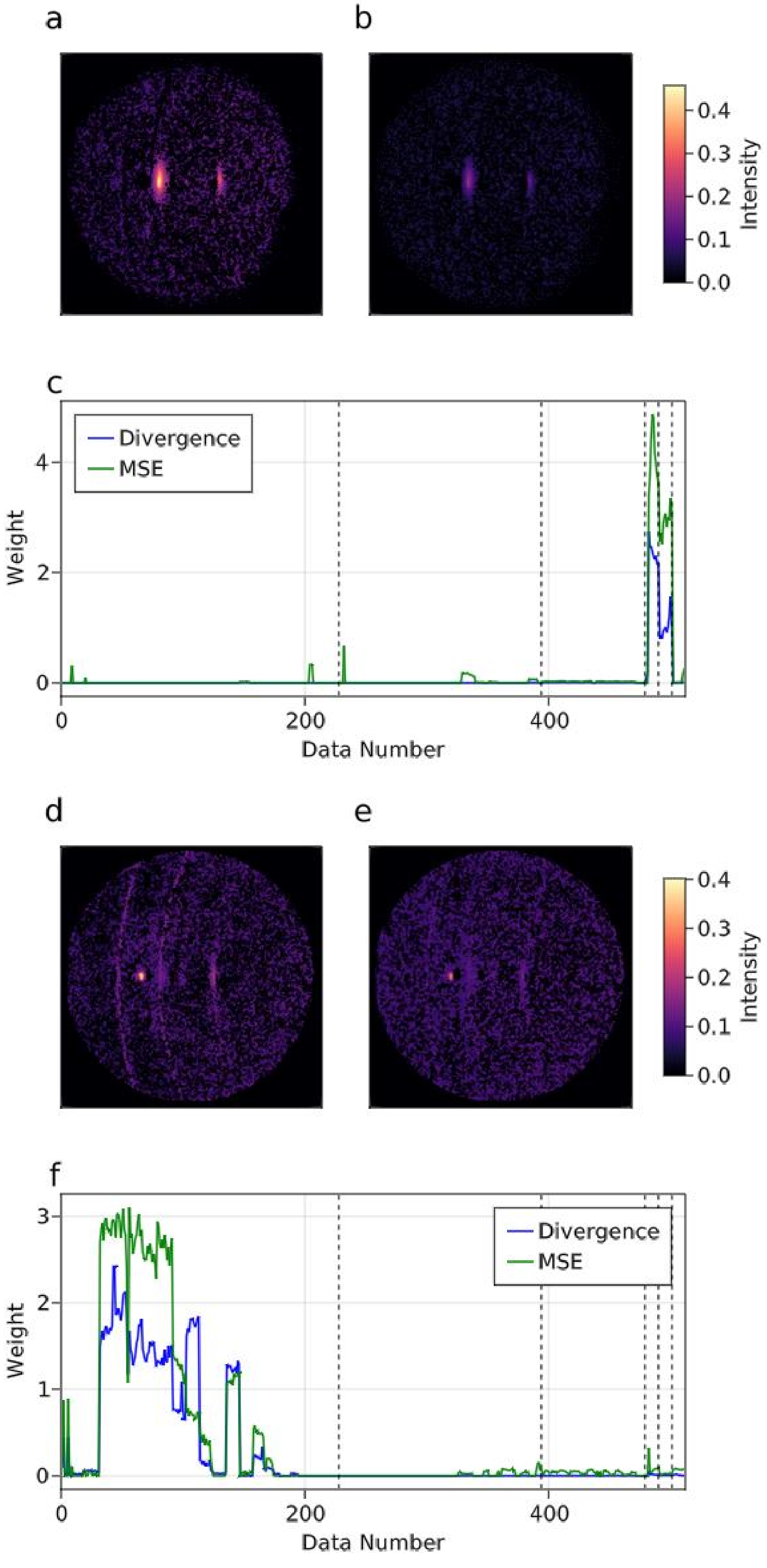
a
b
0.4
0.3
0.2
0.1
0.0
Intensity
c
Divergence
MSE
Weight
4
2
0
0
200
400
Data Number
d
e
0.4
0.3
0.2
0.1
0.0
Intensity
f
Divergence
MSE
Weight
3
2
1
0
0
200
400
Data Number

Figure 2.6 Comparison of objective functions. (a), (d), and blue lines in (c) and (f) are the results when the 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. Data points were ordered by their substrate types as follows; c-sapphire, YSZ (111), YSZ (100), Pt/Ti/SiO2/Si (100), STO (111) and STO (100). The borders of the substrate types are indicated by dashed lines. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis," Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

In a generative model framework, the choice of divergence or MSE as the objective function corresponds to the assumption $\mathrm{V_{ij}} \sim Poisson((WH)_{ij})$, or $\mathrm{V_{ij}} \sim Normal((WH)_{ij})$, respectively [16,22]. Considering the diffraction mechanism [23], the Poisson distribution is suitable for the mechanism. This should be the reason why the result of divergence was better than that of the MSE. Referring to the distribution of the diffraction intensity, it seemed more similar to be an exponential distribution rather than a Poisson distribution because it exhibited a strong peak at 0 intensity and long tail over 0 (Figure 2.7). Therefore, applying the corresponding training algorithm will improve the results. So far, we have confirmed that three hyperparameters are appropriate for this study. Hereafter, the analyses were based on the divergence results.

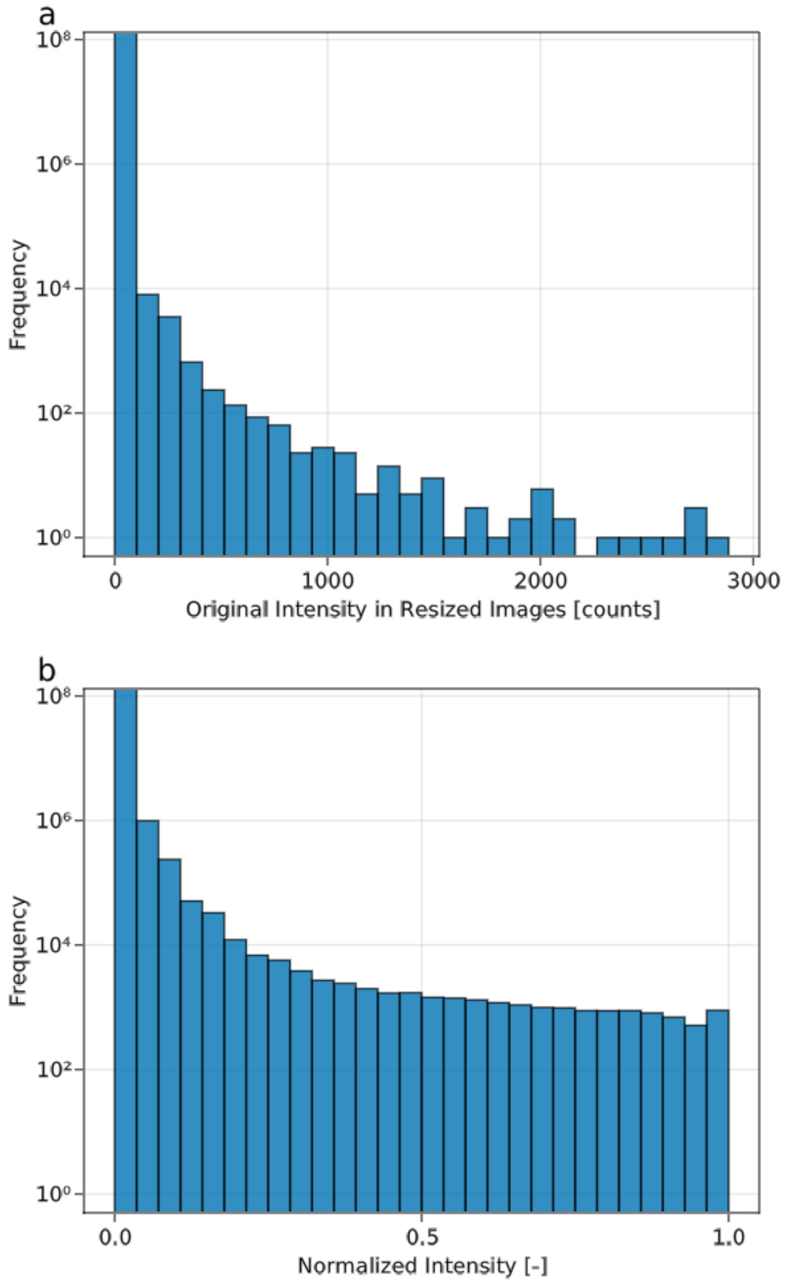

Figure 2.7 Distribution of diffraction intensities of (a) resized images and (b) normalized ones. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis," Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

2.3.3 Relationship among Feature Vectors and Compositions

Figure 2.8 shows the XRD patterns and a heatmap of feature vectors of a $(Ga_{1-x}In_x)_2O_3$ composition-spread sample fabricated on a c-sapphire substrate at 400 °C. Original 2D-XRD images are shown in Figure A12. This sample was reported in another study [18]. This sample contained a solid solution around x = 0.7 - 0.9, and exhibited resulting peak shift around 30.5°. NMF detected this shift and represented it by stronger factor 6 at x = 0.7 and emerging factor 3 from x = 0.8. Factor 6 basically represented the diffraction pattern of the c-sapphire substrate; thus, the values of factor 6 of the sample were above 0.6 over all positions and almost constant under x = 0.7. At x = 0.7, the diffraction signal of the solid solution became stronger and the weight of factor 6 was at its maximum. As shown in Figure 2.2 (6), factor 6 contained slight diffraction pattern of the solid solution of In_2O_3 and Ga_2O_3 around 30.5°. Consequently, factor 6 had stronger intensity at this point. Because the diffraction signal of In_2O_3 (222) was intense at and over x = 0.8, factor 3 had non-trivial values. This result indicates that NMF works as a feature extraction method of 2D-XRD images although its representation may not straightforward for human interpretation.

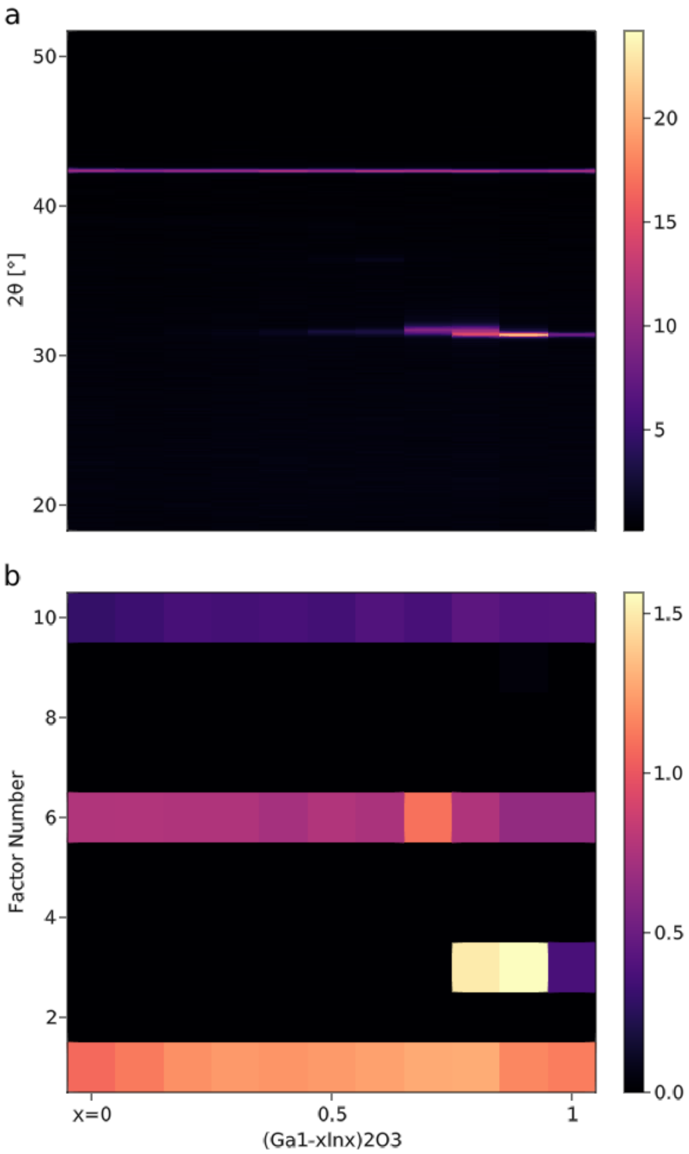

Figure 2.8 (a) Heatmap of the XRD spectral of a $(Ga_{1-x}In_x)_2O_3$ composition spread. The spectral is an integration of the corresponding region in a 2D-XRD image over χ angle. (b) Heatmap of the feature vectors. The x-axes of (a) and (b) are linked and represent composition ratios. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis," Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

2.3.4 Relationship among Feature Vectors and Fabrication Conditions

Figure 2.9 shows the results of the In_2O_3 thin films on the c-sapphire substrates fabricated at substrate temperatures of 300, 400, and 500 °C from the left to the right. Other fabrication conditions were identical. The diffraction signals in the rightmost part of Figure 2.9(a) represent In_2O_3 (222) and evidently the crystal became higher oriented in accordance with the increase in the substrate temperature (the arc-like signal became a spotty signal). Images approximated by NMF [Figure 2.9(b)] represent this tendency through relative intensity changes among the peaks. The leftmost spot in the left and middle images in Figure 2.9(b) was the effect of factor 7 because the factor contained two spots in addition to lower oriented In_2O_3 (222). This is a limitation of the method. The feature vectors represented the change of In_2O_3 (222) [Figure 2.9(c)] as follows: Factor 3 [In_2O_3 (222)] became stronger at 400 °C and 500 °C; 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 in the original 2D-XRD image. We estimate that this sharpness 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 weaker than that of In_2O_3 (222) at these temperatures. Although the arc-like diffraction signal of low-oriented In_2O_3 was insignificant in the approximated image at 300 °C, corresponding factor 7 had a non-trivial value. Further, the feature vectors represented an important propensity at 300 °C that the diffraction signal of In_2O_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 properly recognized diffraction-signal changes according to the fabrication conditions.

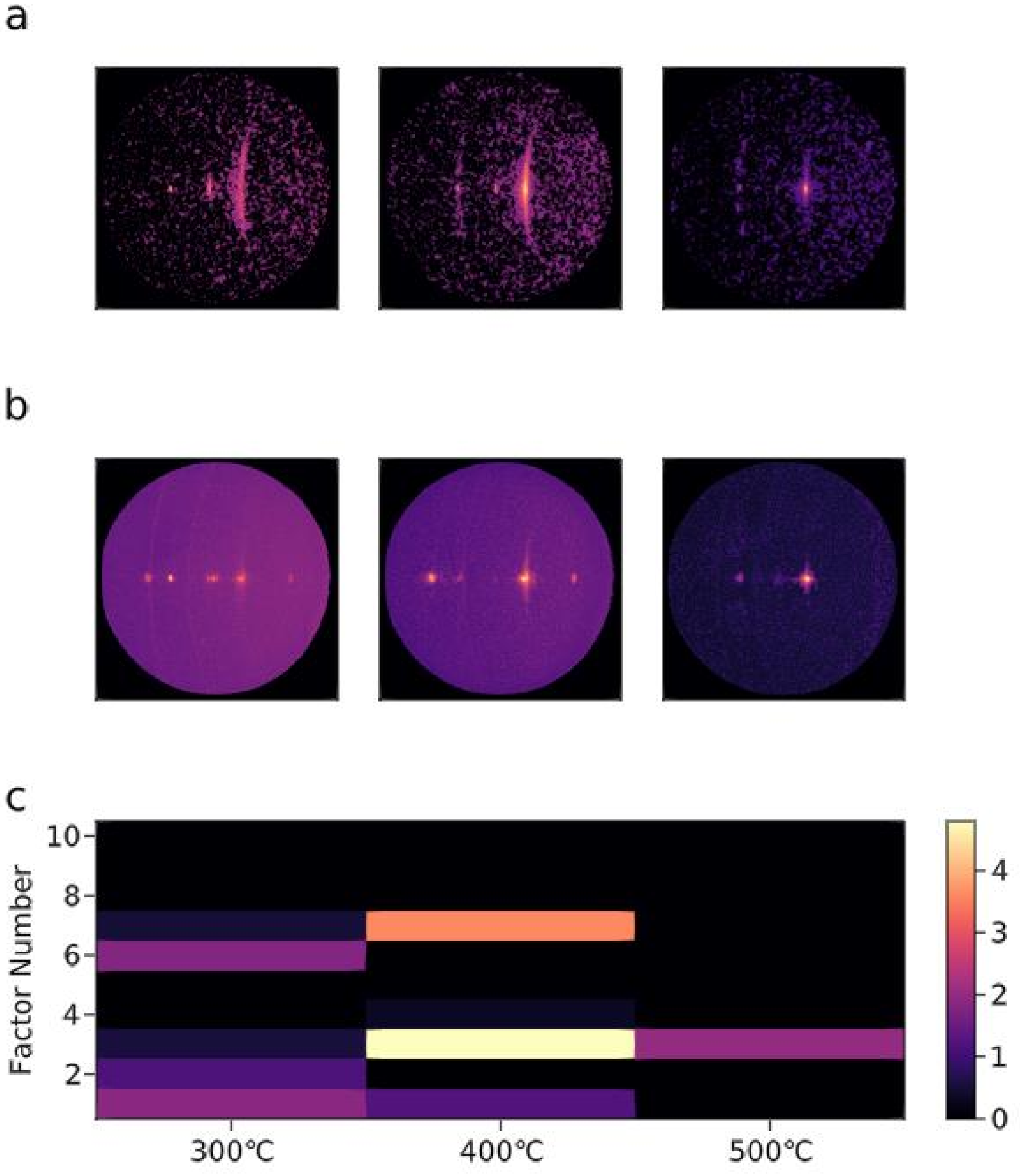

Figure 2.9 Results of three In_2O_3 thin films at substrate temperatures of 300 °C (left), 400 °C (middle), and 500 °C (right). (a) Original 2D-XRD images. (b) 2D-XRD images approximated by NMF. (c) Heatmap of the feature vectors. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis," Science and

Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

2.3.5 Feature Extraction of New Data

Finally, this section evaluates NMF in terms of its inference ability of feature vectors of new data. To test inference ability of the factors, we prepared 2D-XRD images of the thin films of Ga_2O_3, In_2O_3, and $(Ga_{1-x}In_x)_2O_3$ composition spreads fabricated by sputtering. One composition-spread sample was fabricated on an STO (111) substrate, and the others were fabricated on c-sapphire substrates. Because several processes exist to fabricate thin films, confirmation of inference ability to another process is important.

A heatmap of the extracted feature vectors [Figure 2.10 (a)] shows that the diffraction peaks of c-sapphire substrates were represented by factors 4, 8, and 9. Referring to the measurement dates, all samples were estimated to be measured with the Ni-filter and this is why factor 4 was major factor for this dataset. Feature vectors that had significant value in factor 5 (data number 218–228) were from the composition-spread sample fabricated on the STO (111) substrate. Therefore, the heatmap seems reasonable.

To evaluate feature vectors in detail, we compared the results of two composition-spread samples of Bi doped In_2O_3 (doped amount changed from 0% to 15%) [Figure 2.10 (b), (c), and (d)]. One sample [1–11 in Figure 2.10 (b)] contained an intense diffraction peak of In_2O_3 (222), and the shape was slightly arc-like [Figure 2.10 (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_2O_3 (222) crystal [Figure 2.10 (d)]; thus, factor 3 was approximately zero [12–22 in Figure 2.10 (b)]. Note that, in multiplication with the pseudoinverse matrix, non-negativity was not satisfied. These results show that NMF can infer features of new datasets, therefore, NMF will be an appropriate feature extraction method of 2D-XRD images of thin films.

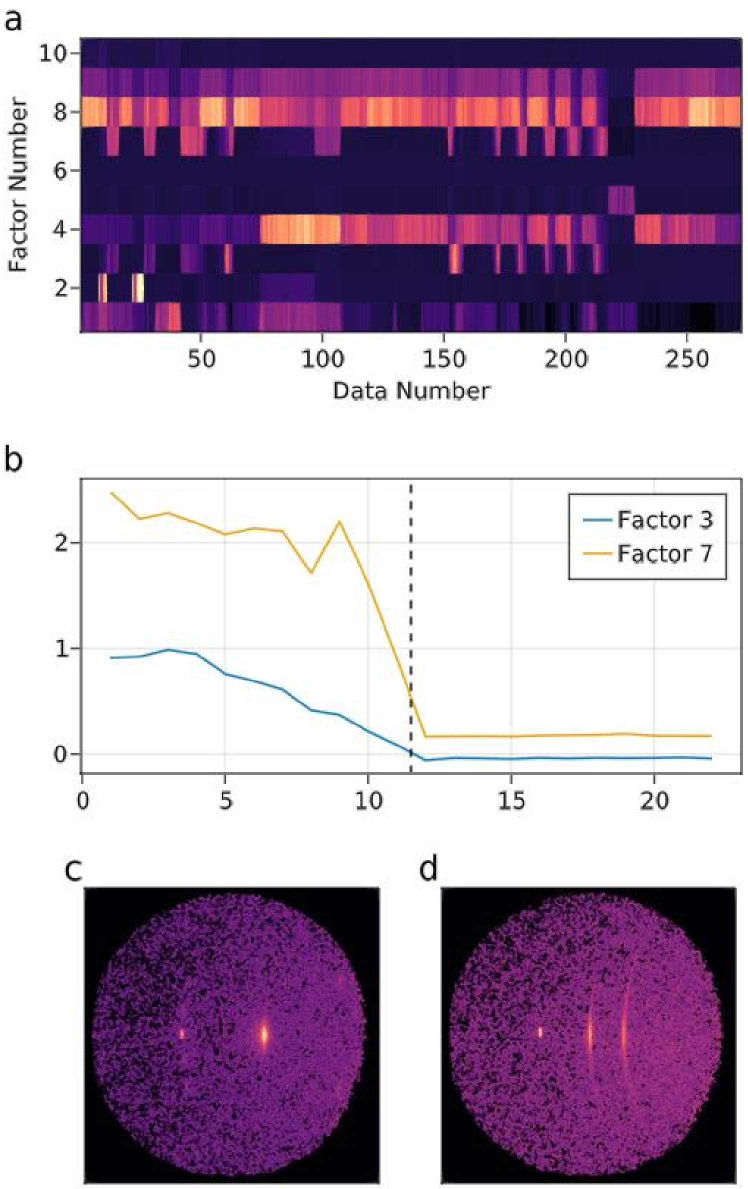

Figure 2.10. (a) Heatmap of the feature vectors of the samples fabricated by sputtering. (b) Changes in factor 3 and factor 7 in two composition-spread samples (1–11 and 12–22). (c) The original 2D-XRD image at 1 in (b). (d) The corresponding image at 12 in (b). This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis," Science and

Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

2.4 Conclusion

In this chapter, we have confirmed that NMF seems an appropriate feature extraction method of 2D-XRD images because of following major five results. (1)NMF learned major diffraction peaks of the dataset based on their diffraction angles and shapes. (2)The output was better under the assumption of Poisson distribution than that of normal distribution. (3)NMF can detect peak shifts in the diffraction angle, although its representation should be improved for better interpretability. (4)NMF also leaned crystalline differences caused by fabrication conditions. (5)The inference of feature vectors of new data by NMF was confirmed to be reasonable. These results indicates that NMF is a candidate of the feature extraction method of 2D-XRD images. Whether the extracted features are appropriate one for deep learning models is evaluated in Chapter 3. Although other structures such as bulk samples than polycrystalline thin films were not tested, NMF will be applicable to 2D-XRD images of those samples.

References

[1] A. G. Kusne, T. Gao, A. Mehta, L. Ke, M. C. Nguyen, K.-M. Ho, V. Antropov, C.-Z. Wang, M. J. Kramer, C. Long, and I. Takeuchi, *On-the-Fly Machine-Learning for High-Throughput Experiments: Search for Rare-Earth-Free Permanent Magnets*, Scientific Reports **4**, 6367 (2015).

[2] C. J. Long, J. Hattrick-Simpers, M. Murakami, R. C. Srivastava, I. Takeuchi, V. L. Karen, and X. Li, *Rapid Structural Mapping of Ternary Metallic Alloy*

Systems Using the Combinatorial Approach and Cluster Analysis, Review of Scientific Instruments **78**, 072217 (2007).

[3] C. J. Long, D. Bunker, X. Li, V. L. Karen, and I. Takeuchi, *Rapid Identification of Structural Phases in Combinatorial Thin-Film Libraries Using x-Ray Diffraction and Non-Negative Matrix Factorization*, Review of Scientific Instruments **80**, 103902 (2009).

[4] S. E. Ament, H. S. Stein, D. Guevarra, L. Zhou, J. A. Haber, D. A. Boyd, M. Umehara, J. M. Gregoire, and C. P. Gomes, *Multi-Component Background Learning Automates Signal Detection for Spectroscopic Data*, Npj Computational Materials **5**, 77 (2019).

[5] Y. Ozaki, Y. Suzuki, T. Hawai, K. Saito, M. Onishi, and K. Ono, *Automated Crystal Structure Analysis Based on Blackbox Optimisation*, Npj Computational Materials **6**, 75 (2020).

[6] 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, and J. M. Gregoire, *Automated Phase Mapping with AgileFD and Its Application to Light Absorber Discovery in the V–Mn–Nb Oxide System*, ACS Combinatorial Science **19**, 37 (2017).

[7] L. Banko, Y. Lysogorskiy, D. Grochla, D. Naujoks, R. Drautz, and A. Ludwig, *Predicting Structure Zone Diagrams for Thin Film Synthesis by Generative Machine Learning*, Communications Materials **1**, 15 (2020).

[8] X. Li, Y. Zhang, H. Zhao, C. Burkhart, L. C. Brinson, and W. Chen, *A Transfer Learning Approach for Microstructure Reconstruction and Structure-Property Predictions*, Scientific Reports **8**, 13461 (2018).

[9] N. Dalal and B. Triggs, *Histograms of Oriented Gradients for Human Detection*, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1 (IEEE, 2005), pp. 886–893.

[10] D. G. Lowe, *Object Recognition from Local Scale-Invariant Features*, in *Proceedings of the Seventh IEEE International Conference on Computer Vision* (IEEE, 1999), pp. 1150–1157 vol.2.

[11] H. Bay, T. Tuytelaars, and L. van Gool, *SURF: Speeded Up Robust Features*, in *Computer Vision – ECCV 2006* (2006), pp. 404–417.

[12] Y. LeCun, Y. Bengio, and G. Hinton, *Deep Learning*, Nature **521**, 436 (2015).

[13] K. He, X. Zhang, S. Ren, and J. Sun, *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*, (2015).

[14] M. Bojarski, D. del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, *End to End Learning for Self-Driving Cars*, (2016).

[15] J. P. Horwath, D. N. Zakharov, R. Mégret, and E. A. Stach, *Understanding Important Features of Deep Learning Models for Segmentation of High-Resolution Transmission Electron Microscopy Images*, Npj Computational Materials **6**, 108 (2020).

[16] D. D. Lee and H. S. Seung, *Learning the Parts of Objects by Non-Negative Matrix Factorization*, Nature **401**, 788 (1999).

[17] V. Stanev, V. v. Vesselinov, A. G. Kusne, G. Antoszewski, I. Takeuchi, and B. S. Alexandrov, *Unsupervised Phase Mapping of X-Ray Diffraction Data by Nonnegative Matrix Factorization Integrated with Custom Clustering*, Npj Computational Materials **4**, 43 (2018).

[18] T. Nagata, T. Hoga, A. Yamashita, T. Asahi, S. Yagyu, and T. Chikyow, *Valence Band Modification of a (Ga x In 1− x) 2 O 3 Solid Solution System Fabricated by Combinatorial Synthesis*, ACS Combinatorial Science **22**, 433 (2020).

[19] C. Boutsidis and E. Gallopoulos, *SVD Based Initialization: A Head Start for Nonnegative Matrix Factorization*, Pattern Recognition **41**, 1350 (2008).

[20] C.-J. Hsieh and I. S. Dhillon, *Fast Coordinate Descent Methods with Variable Selection for Non-Negative Matrix Factorization*, in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’11* (ACM Press, New York, New York, USA, 2011), p. 1064.

[21] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari, *Nonnegative Matrix and Tensor Factorizations* (John Wiley & Sons, Ltd, Chichester, UK, 2009).

[22] T. Virtanen, A. Taylan Cemgil, and S. Godsill, *Bayesian Extensions to Non-Negative Matrix Factorisation for Audio Signal Modelling*, in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, 2008), pp. 1825–1828.

[23] Charles Kittel, *Introduction to Solid State Physics*, 8th ed. (Wiley, 2004).

[24] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (AAAI Press, Portland, OR, 1996), pp. 226–231.

[25] A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, *Accelerating Two-Dimensional X-Ray Diffraction Measurement and Analysis with Density-Based Clustering for Thin Films*, Japanese Journal of Applied Physics **60**, SCCG04 (2021).

Chapter 3

Continuous Representation of Microstructures of Thin Films Fabricated under Multiple Conditions

3.1 Introduction

Structure related properties, such as compositions or space groups, are often represented as categorical variables in data-driven materials research. Categorical variables are nominal scale, which means that relationships among variables are mathematically unclear compared to continuous variables, or ratio scale. Continuous representation has more applicable machine learning techniques including gradient-based updates [1] than categorical representation has. Indeed, continuous representation has been reported its advantages in data-driven materials research [2–5]. The continuous representation has other advantages that it can plot graphical maps of datasets for human interpretation and represent numerous data in the fixed dimensionality, which are less described in articles but important for the practical research. The main technique used in the literature is variational autoencoder (VAE) [6], which is a famous generative model. VAE consists of two neural networks, an encoder and a decoder. This model learns so that its latent space (the space where the outputs of the encoder exist) consists of a continuous probability distribution. Thus, an encoded data, or a latent variable, is in continuous representation.

In this chapter, the continuous representation of microstructures of thin films with VAE and 2D-XRD images is discussed. Firstly, features of 2D-XRD images extracted by NMF is evaluated whether it improved the performance of VAE. A feature vector extracted by NMF is referred as a NMF feature in this Chapter. Secondary, the relationships of the latent space to the NMF features and fabrication conditions are analysed. Lastly, the inference ability is evaluated with a dataset of samples fabricated by sputtering. The dataset and the NMF features in this chapter were the same as those in Chapter 2.

3.2 Method

3.2.1 Variational AutoEncoder

We modeled two VAEs [6] whose encoders and decoders each had a single hidden layer. One VAE directly learned the 2D-XRD images, which were resized into 64 × 64 pixels owing to the capacity of our computer. The other VAE learned the features extracted by NMF in Chapter 2. The dimensionalities of the hidden layers were approximately the root square of the input dimensionalities: 64 and 3, respectively. The output dimensionalities of the encoders were 2, thus, we visualized distributions of the latent variables. The prior distributions of the latent spaces were set to normal distributions with averages and variances of 0 and 1, respectively. Both VAEs were programmed with Julialang [7] and Flux.jl [8], and trained 100 times.

3.2.2 Visualization of NMF Factor Axes in the Latent Space

To analyze the effect of NMF factors in the latent space of VAE, we visualized axes which the NMF factors constitute. 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 factors. The origin was plotted by encoding a zero vector.

3.3 Results and Discussion

3.3.1 Evaluation of Feature Extraction by NMF

We compared two VAEs to evaluate whether the feature extraction by NMF improved the performance of a VAE. In this chapter, performance of a VAE is evaluated in terms of a distribution of latent variables. One VAE learned directly from the 2D-XRD images, and the other learned from the NMF features. Figure 3.1 (a) and (b) show the distributions of the latent variables in latent spaces colored based on substrate types. The distribution of the latent variables of direct VAE application apparently failed to separate the samples based on the substrate types. In contrast, the distribution of the latent variables of the NMF features seems better in a point that latent variables of some substrate types were distributed separately. In particular, the latent variables of samples on YSZ (100) and STO (100) substrates were separated from other variables. This is consistent with the fact that these substrates are different from the other substrates in terms of lattice mismatch between In_2O_3 crystals 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

In_2O_3. Differences of substrate types may be a minor cause which separated the latent variables of samples on YSZ (100) or STO (100) substrates from the major linear distribution. Although this dispersion in the horizontal axis can be enlarged by adjustment of coefficients of the loss function of the VAE, we did not adjust in this chapter because of the comparison of the two VAEs. (The fine-tuned results are shown in Chapter 4.) The reason of mixture region of latent variables of the c-sapphire and YSZ (111) substrates is that both the diffraction signal of In_2O_3 (222) on the c-sapphire substrate and the diffraction signal of the YSZ (111) substrate were represented by factor 3. This is because those diffraction signals were located at close diffraction angles of 30.5° and 30.1°, respectively. This slight difference in the diffraction angles could not be differentiated based on the resolution of our measurement setup. Therefore, the mixture region is not problematic. Considering another study whose data volume was 15,000 [9], we conclude that, with high-dimensional data points or a small dataset, feature extraction is important to improve performance of deep learning models.

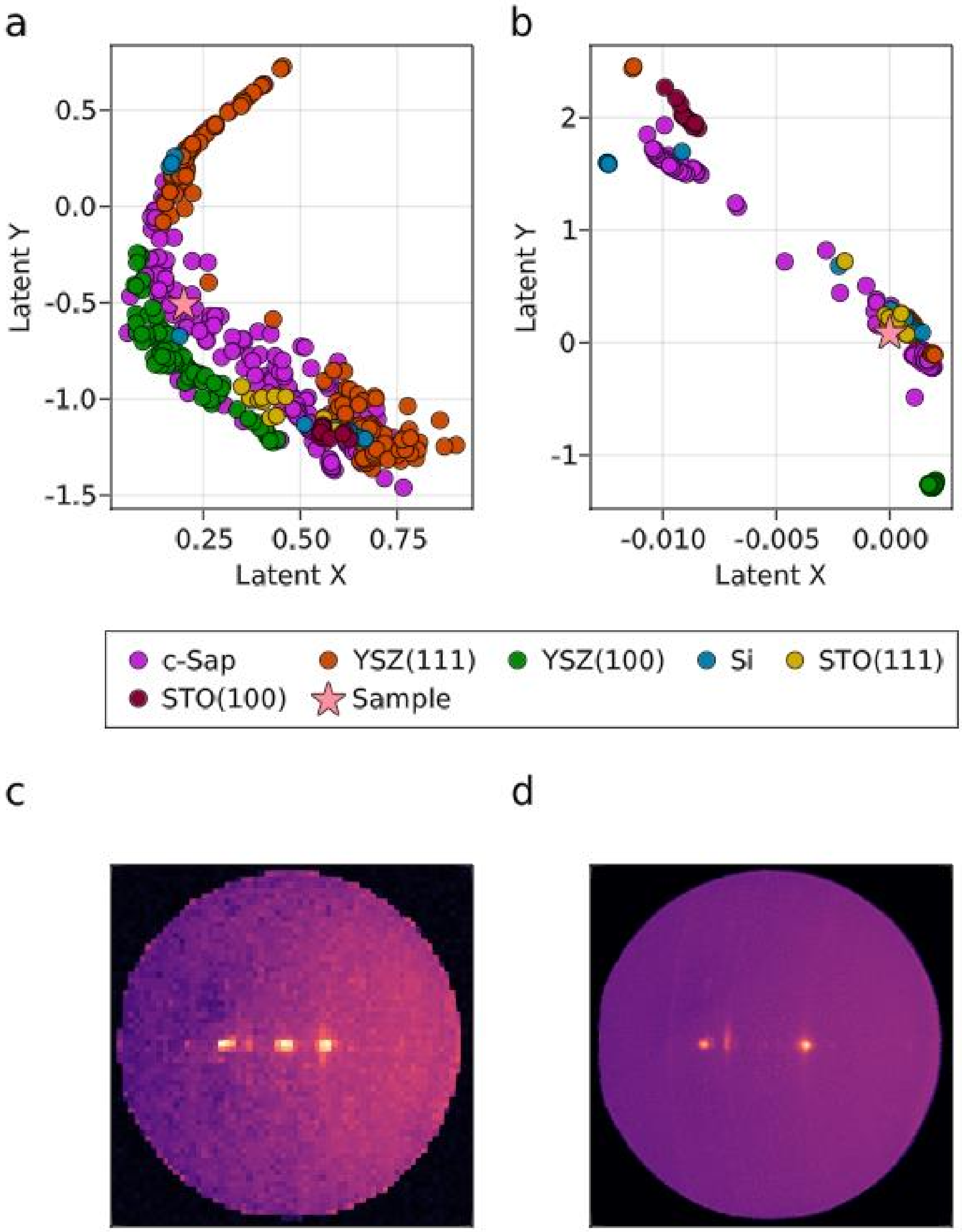

Figure 3.1 Distributions of latent variables of the VAEs that learned (a) the 2D-XRD images directly and (b) the NMF features. (c) and (d) are decoded images from latent variables located at stars in (a) and (b), respectively. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Direct feature

extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis," Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

The major advantage of a VAE is that it can generate new data points through sampling from its probability distribution. To demonstrate this, we sampled single latent variables from each dense region in the latent spaces [stars in Figure 3.1 (a) and (b), respectively], and then decoded them with corresponding decoders [Figure 3.1 (c) and (d)]. The image decoded by the directly learned VAE contained mainly three peaks, which seemed the diffraction signals from the c-sapphire substrate (left), YSZ (100) substrate or In_2O_3 (400) (middle), and In_2O_3 (222) (right). These diffraction patterns corresponded to the distribution where latent variables of c-sapphire and YSZ (100) substrates were close. Note that the reason why background noise seemed intense was the reduction in image size. The image decoded by the VAE which learned the NMF features contained mainly three diffraction peaks, which appeared from the c-sapphire substrate (left), Pt electrode (middle), and In_2O_3 (222) (right). This corresponded the distribution that latent variables of samples on STO (111) with Pt electrode and on c-sapphire substrates existed. These results show a potential to generate new data points to estimate microstructures of unfabricated thin films.

We noticed that the distributions changed their shapes over trials. This can be stabilized by increasing the number of NMF factors or dimensionalities of hidden layers of VAE. The stabilized result is discussed in Chapter 4.

3.3.2 Coordinate System of Latent Space and Process Improvement

Since this section, the results and discussions are on the latent space which learned the NMF features. The relationship between the coordinate system of the latent space and the fabrication conditions is discussed in this section. For the discussion, the axes of the factors in the latent space were visualized by the method in Section 3.2.2 [Figure 3.2(a)]. Each line represents intensities of the XRD patterns shown in Figure 2.1. We noticed that intensities of some factors were not sensitive to the position in the latent space, especially factor 1. This should be because factor 1 represents background noise, which makes a slight difference in the appearance of 2D-XRD images. The axis of factor 4 was drawn to the upper left corner; therefore, some data points of the c-sapphire substrate were distributed in this region [Figure 3.1(b)]. No axes were drawn to the upper right or bottom left. This implies that all the factors were correlated to each other, although factor 2 [YSZ (100)] can be the exception. Considering that factor 7 was longly drawn to the upper left and factor 2 shortly to the bottom left, the discussion above on the narrow dispertion in the horizontal axis is reasonable. Although this suggests high bias in the dataset, the suggestion will not weaken the study because high bias is common situation in all practical research focusing on certain material systems.

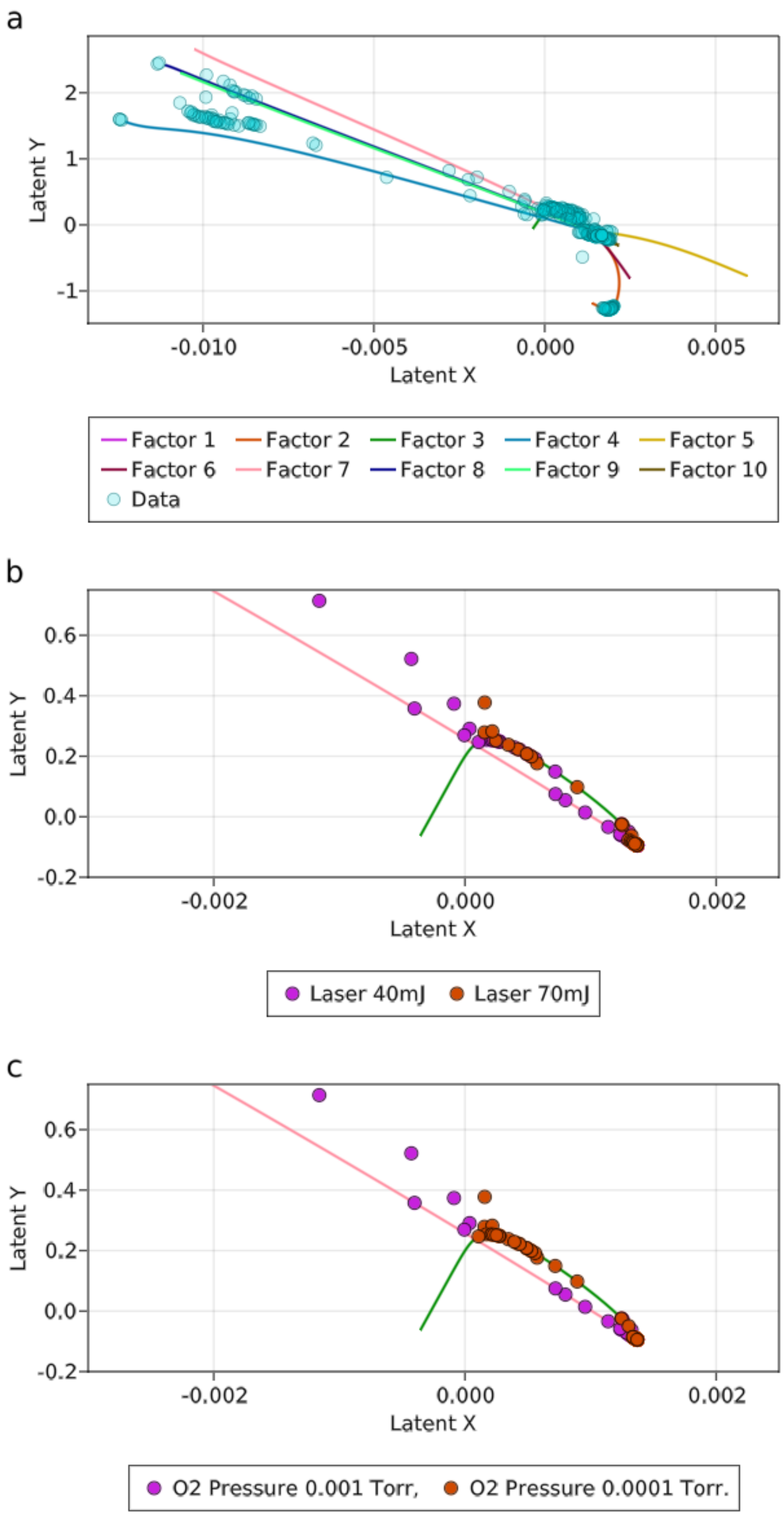

Figure 3.2 (a) Coordinates of the factors in the latent space and the distribution of the latent variables. (b) and (c) are the distribution of the features of the c-sapphire samples, whose components were set to zero except factor 3 and factor 7. The plots are colored

based on the fabrication conditions. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis,” Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

To evaluate the relationship between the fabrication conditions and the crystallinity of In_2O_3 in the latent space, we selected the data points of the samples that were fabricated on c-sapphire substrates. We modified features of them so that all the components except factors 3 and 7 were set to 0. Then, the modified features were encoded and colored based on the fabrication conditions [Figure 3.2(b) and (c)]. Latent variables located on the lines indicate pure phase either of factor 3 or 7 and the others indicate mixed phases of them. The results indicate that In_2O_3 (222) tended to be highly oriented under the fabrication condition that the laser intensity was 70 mJ and O_2 pressure was 0.1 mTorr regardless to other fabrication conditions such as substrate temperature. This was consistent with the knowledge that the higher laser intensity and lower O_2 pressure enhanced the migration of In_2O_3 and its crystal growth. This result suggests that the visualization with VAE support researchers to compare fabrication conditions.

3.3.3 New Data in Latent Space

We encoded the NMF features of samples fabricated by sputtering (same as Chapter 2) to evaluate inference ability of VAE (Figure 3.3). The majority of the latent variables of c-sapphire substrates was distributed in the upper left, because factor 4 was the main factor with these samples. The latent variables of samples on STO (111) were distributed almost close to the correspondence fabricated by PLD. The distribution of samples by sputtering appeared to be wider than that of samples by PLD. This reflects the propensity that sputtering can fabricate more variate microstructures with the material system in this study than PLD can. Therefore, we conclude that the NMF features are appropriate features to represent microstructures of thin films. In addition, the combination of NMF and VAE can visualize propensities of fabrication processes. This is more discussed in Chapter 4.

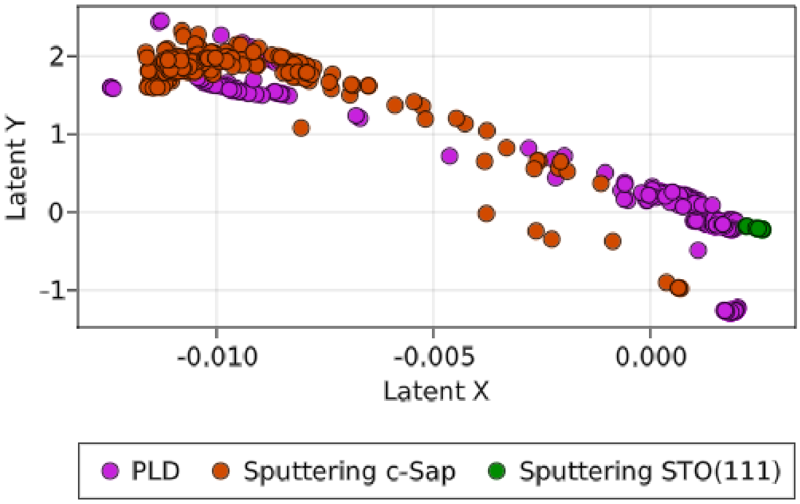

Figure 3.3 The distribution of the data fabricated by sputtering in the latent space. The latent space is the same as Figure 3.2. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis," Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

3.4 Conclusion

In this chapter, we have confirmed that NMF is an appropriate feature extraction method for 2D-XRD images to improve the performance of VAE. The NMF features enabled the VAE to learn that differences in the dataset was caused mainly by crystallinity of In_2O_3

and partially by substrate types. Although we only evaluated VAE, the NMF features will improve performance of other deep learning models. The VAE converted the NMF features into latent continuous variables which visualized latent relationships in the dataset. The visualization of VAE was confirmed to be useful for fabrication analysis and this is more discussed in Chapter 4. We noticed that the decoders were not well optimized in this chapter. Although they were less important than the encoders in this thesis, they should be optimized to generate 2D-XRD images for prediction of microstructures of thin films. Although this chapter lacks this discussion, we confirmed that VAE converted 2D-XRD images into continuous variables.

References

[1] R. Rojas, *Neural Networks* (Springer Berlin Heidelberg, Berlin, Heidelberg, 1996).

[2] I.-H. Lee and K. J. Chang, *Crystal Structure Prediction in a Continuous Representative Space*, Computational Materials Science **194**, 110436 (2021).

[3] J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik, and Y. Jung, *Inverse Design of Solid-State Materials via a Continuous Representation*, Matter **1**, 1370 (2019).

[4] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, *Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules*, ACS Central Science **4**, 268 (2018).

[5] H. S. Stein, D. Guevarra, P. F. Newhouse, E. Soedarmadji, and J. M. Gregoire, *Machine Learning of Optical Properties of Materials – Predicting Spectra from Images and Images from Spectra*, Chemical Science **10**, 47 (2019).

[6] D. P. Kingma and M. Welling, *Auto-Encoding Variational Bayes*, 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings (2013).

[7] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, *Julia: A Fresh Approach to Numerical Computing*, SIAM Review **59**, 65 (2017).

[8] M. Innes, *Flux: Elegant Machine Learning with Julia*, Journal of Open Source Software **3**, 602 (2018).

[9] L. Banko, P. M. Maffettone, D. Naujoks, D. Olds, and A. Ludwig, *Deep Learning for Visualization and Novelty Detection in Large X-Ray Diffraction Datasets*, Npj Computational Materials **7**, 104 (2021).

Chapter 4

Fabrication Analyses of Indium Gallium Oxide Thin Films with Machine Learning

4.1 Introduction

The fabrication of thin films has various conditions to be optimized depending on a process [1]. In addition, suitable processes for research and mass production are different. For research, PLD is a suitable process to fabricate high quality thin films. For mass production, sputtering is a suitable process to fabricate thin films on larger substrates. Conditions to be optimized for PLD and sputtering are different because of differences of fabrication mechanisms. This makes difficult to convey optimized fabrication conditions from research to mass production. To circumvent the difficulty, this chapter discusses a way to compare fabrication processes as well as conditions with the methods in previous chapters.

In this chapter, the methods described in Chapters 2 and 3 are discussed in terms of a fabrication analysis of indium gallium oxide (IGO) thin films. IGO thin films are promising semiconductors for their wider band gaps and tunability of properties. However, they have different crystal structures [In_2O_3: cubic, (β-)Ga_2O_3: monoclinic], which makes difficult to fabricate the solid solution. In addition, crystal structures of their mixture and their properties are still under study [2,3]. These topics will be explored in other studies. This chapter focuses on a relationship of their microstructures and fabrication conditions of PLD and sputtering so that IGO thin films will be moved into mass production stage with ease.

4.2 Method

The dataset contained 773 2D-XRD images, which excluded 11 data with contaminations from the dataset of Chapter 2. The number of factors of NMF was set to 128 to stabilize the distribution of latent variables. Data of samples of PLD (501 data) and sputtering (272 data) was combined from the training of NMF. The dataset was also shuffled to eliminate the dependency of results on the data number. The algorithms were the same as those of Chapter 2. The training time of NMF was 100. The number and dimensionality of hidden layers of encoders and decoders of VAE were one and 64, respectively. The input of VAE was shuffled to eliminate the dependency of results on the data number. The training time of VAE was 100.

4.3 Results and Discussion

Figure 4.1 shows the latent space of the VAE colored based on the substrate types and the fabrication processes. Propensities of the distribution of the latent variables were confirmed to be stable over trials. The latent space consisted of a normal distribution whose parameters were learned from the dataset. Therefore, an anomalous data tend to be distributed in outer region in the latent space because its likelihood should be small. In this study, many latent variables were plotted in the center island which had a small island in its upper right. Latent variables of samples on STO (100) substrates were separately distributed from those of the other substrate types, which implies that they were anomalous in this study. This distribution is consistent with the result of Chapter 3, although those on YSZ (100) were not. The reason why the latent variables of samples on YSZ (100) were not recognized as anomalous is not studied in this thesis for the time constraint. Latent variables of samples fabricated by PLD seem to be distributed broader than those by sputtering. This indicates that PLD is a better process to fabricate anomalous samples, which is consistent experts' intuition that PLD is a suitable process for research. Therefore, the distribution of latent variables seems reasonable in overview.

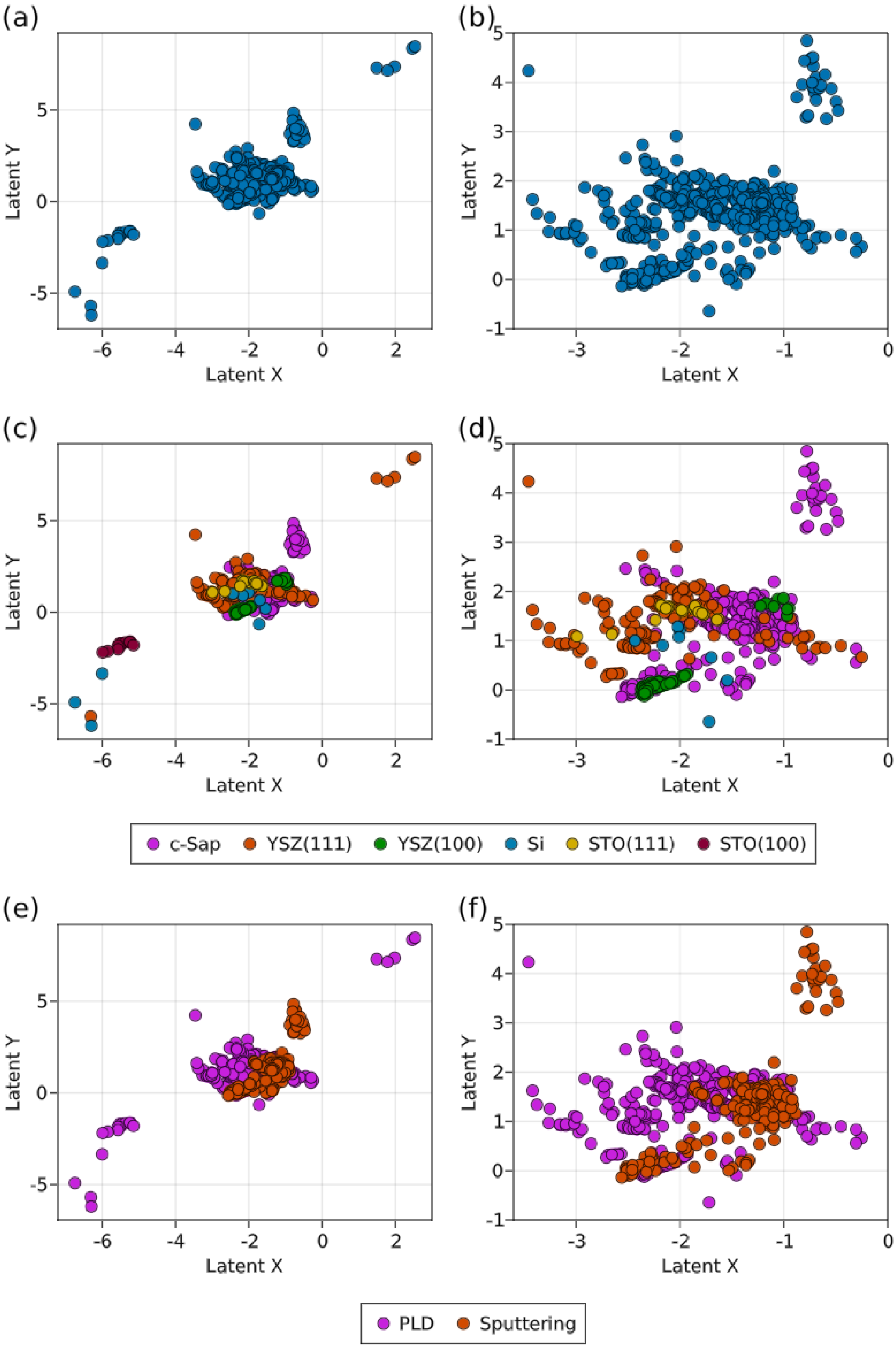
(a)
(b)
(c)
(d)
(e)
(f)
Latent X
Latent Y
-6
-4
-2
0
2
-5
5
-3
-1
1
3
4
● c-Sap
● YSZ(111)
● YSZ(100)
● Si
● STO(111)
● STO(100)
● PLD
● Sputtering

Figure 4.1 The latent space of VAE. (a) Not colored. (c) Colored based on substrate types. (e) Colored based on fabrication processes. (b), (d) and (f) are zoomed plots of their left plots.

The latent variables which were distributed in the small island in the upper right of the central island were 22 2D-XRD images from two Bi doped In_2O_3 thin films. The two thin films were fabricated by co-sputtering so that they possessed dopant gradient from 0% to 15%. In addition to the two thin films, one thin film was fabricated the same conditions with different sputtering intensity. Table 4.1 shows the fabrication conditions of the three thin films. Figure 4.2 shows the zoomed plot of the latent space and their typical 2D-XRD images. Diffraction patterns were c-sapphire substrate (left), In_2O_3 (400) (middle) and In_2O_3 (222). From (b) to (d), diffraction intensity of In_2O_3 (222) became weaker, in contrast, that of In_2O_3 (400) became stronger. Referring to Table 4.1, this difference of microstructures may be an effect of Bi dopant. The VAE seems to recognize that the microstructures which fabricated by sputtering power over 30 W were anomalous. This recognition seems reasonable because 222 plane of In_2O_3 tended to grow better than 400 plane in this study. Further measurements of other properties will reveal more insights.

Table 4.1 Fabrication conditions of three Bi doped In_2O_3 thin films fabricated by co-sputtering.

Sample name	Target A	Target B	Power for Target A [W]	Power for Target B [W]	Gas Ratio (O2/Ar) [%]	Fabrication script	Average Thickness [nm]

Sample α	In_2O_3	Bi:In_2O_3	50	20	20	Co-sputtering	83.2
Sample β	In_2O_3	Bi:In_2O_3	50	35	20	Co-sputtering	85.7
Sample γ	In_2O_3	Bi:In_2O_3	50	50	20	Co-sputtering	76.6

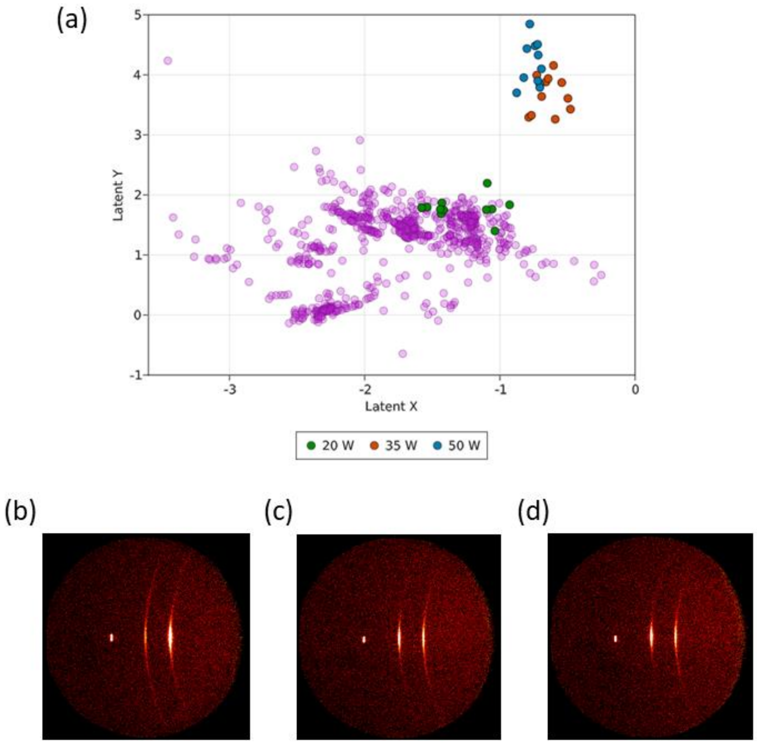

Figure 4.2 (a) the distribution of the three Bi doped In_2O_3 thin films. Colored based on the power of sputtering. (b), (c) and (d) are typical 2D-XRD images of samples α, β and γ, respectively.

Figure 4.3 shows distributions of latent variables of thin films of Ga_2O_3 and In_2O_3 fabricated by PLD and sputtering. Transitions of microstructures of composition spreads $(Ga_{1-x}In_x)_2O_3$ (Samples A, B, C, and D) according to the composition gradient in the latent space are also drawn. Latent variables of thin films of Ga_2O_3 and In_2O_3 fabricated by PLD were closely distributed. In contrast, those fabricated by sputtering were separately distributed. The distribution may reflect that the crystal of Ga_2O_3 did not well grow by PLD in this study. In addition, transitions of composition spreads fabricated by PLD (Samples A and B) seem to have little relationship to distributions of thin films of Ga_2O_3 and In_2O_3. On the other hand, those by sputtering (Samples C and D) seem to continuously move between distributions of thin films of Ga_2O_3 and In_2O_3. Figure 4.4 shows valence band spectra of Sample A measured by X-ray photoelectron spectroscopy and represents the transition from Ga dominant structure to In dominant structure around $x = 0.8$. This result indicates that the large transitions which the four samples exhibited in the latent space (Figure 4.3) reflect transitions of space group from monoclinic (β-Ga_2O_3) to cubic (In_2O_3). Considering the transition, microstructures in the transition area ($0.4 \leq x \leq 0.7$) exhibited by Samples C and D may contain novel structures, although these are not confirmed in this thesis. These results suggest that sputtering is better process to investigate microstructures of the IGO system.

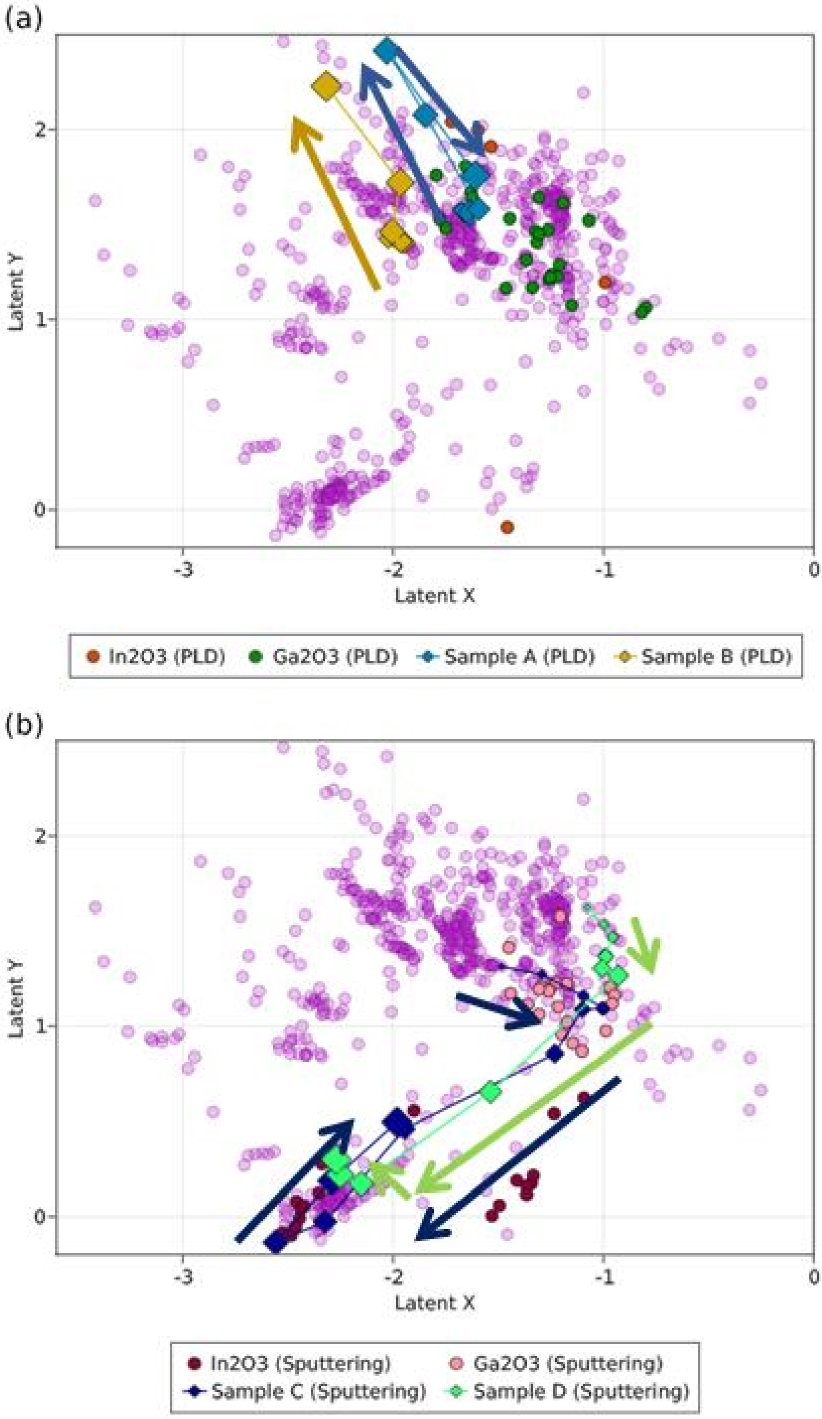

Figure 4.3 (a) Distributions of data points of thin films of Ga_2O_3 and In_2O_3 fabricated by PLD in the latent space. Box markers represent transitions of two composition spreads in the latent space according to the composition gradient. Larger size of markers corresponds larger x of $(Ga_{1-x}In_x)_2O_3$. (b) The corresponding plot of samples fabricated by sputtering.

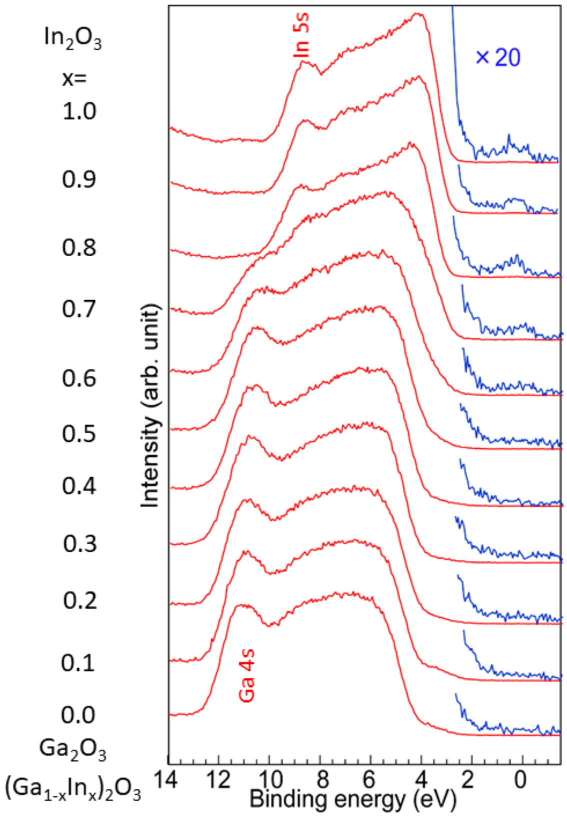

Figure 4.4 Valence band spectra of a (Ga$_{1-x}$In$_x$)$_2$O$_3$ composition spared (Sample A in Figure 4.3). This figure is partially modified from Takahiro Nagata, Takeshi Hoga, Akihiro Yamashita, Toru Asahi, Shinjiro Yagyu, and Toyohiro Chikyow, "Valence Band Modification of a (GaxIn1−x)2O3 Solid Solution System Fabricated by Combinatorial Synthesis," ACS Combinatorial Science, 2020, 22, 9, 433-439 ©2020 American Chemical Society

4.4 Conclusion

The current results of this chapter are just for a proof of concept of the fabrication analysis. Although this chapter is a preliminary study, the VAE has been confirmed that it is an appropriate visualization method for a fabrication analysis. The visualization in this chapter will help non-experts to compare fabrication conditions of thin films. The visualization suggested that sputtering is a better process to fabricate composition spreads of indium gallium oxide to study microstructures of them. Although this chapter compared only two processes, other fabrication processes can be compared in the similar way. The study reported by Banko et. al. [1] will be a good reference to deepen the study of this chapter.

References

[1] L. Banko, Y. Lysogorskiy, D. Grochla, D. Naujoks, R. Drautz, and A. Ludwig, *Predicting Structure Zone Diagrams for Thin Film Synthesis by Generative Machine Learning*, Communications Materials **1**, 15 (2020).

[2] J. E. N. Swallow, R. G. Palgrave, P. A. E. Murgatroyd, A. Regoutz, M. Lorenz, A. Hassa, M. Grundmann, H. von Wenckstern, J. B. Varley, and T. D. Veal, *Indium Gallium Oxide Alloys: Electronic Structure, Optical Gap, Surface Space Charge, and Chemical Trends within Common-Cation Semiconductors*, ACS Applied Materials & Interfaces **13**, 2807 (2021).

[3] J. Sheng, E. J. Park, B. Shong, and J.-S. Park, *Atomic Layer Deposition of an Indium Gallium Oxide Thin Film for Thin-Film Transistor Applications*, ACS Applied Materials & Interfaces **9**, 23934 (2017).

Chapter 5

Density-Based Analysis of Two-Dimensional X-Ray Diffraction Images

This chapter is partially a modified reproduction from Accepted Manuscript version of the article, A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Accelerating two-dimensional X-ray diffraction measurement and analysis with density-based clustering for thin films,” Japanese Journal of Applied Physics, 2021, 60, SCCG04, https://doi.org/10.35848/1347-4065/abf2d8 ©2021 IOP Publishing and Japan Society of Applied Physics

5.1 Introduction

This chapter discusses the signal density of a 2D-XRD image as another feature for machine learning models. In previous chapters, a result of a 2D-XRD measurement was evaluated in the image form, or the heatmap representation. As shown in Figure 5.1, a diffraction pattern is less vague in terms of the signal density compared to the image form. This should be advantage in a measurement of a low signal-to-noise (S/N) ratio sample such as a thin film. For confirmation of this advantage, the signal density was evaluated by a method called ordering points to identify the clustering structure (OPTICS) [1].

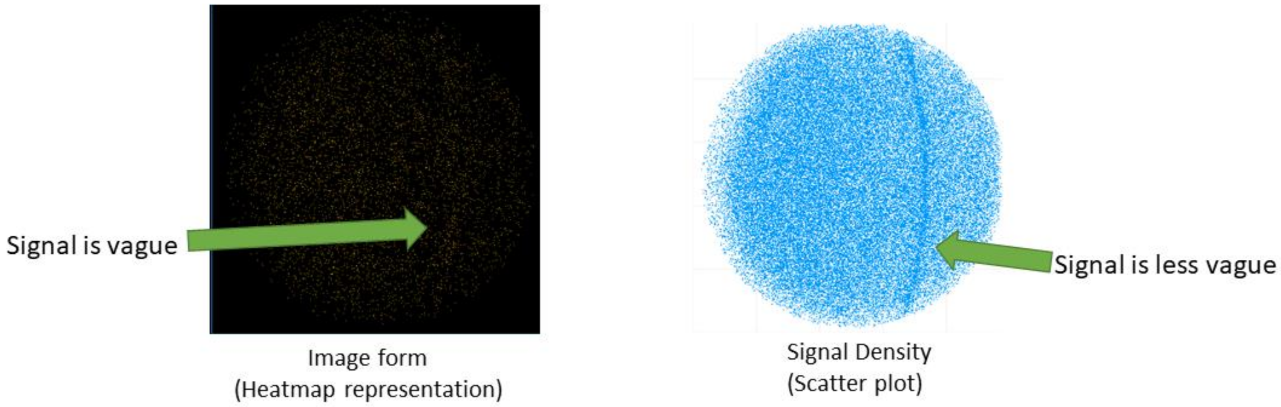

Figure 5.1 Comparison of representation forms of a 2D-XRD measurement with heatmap representation and scatter plot.

This chapter also discusses how to optimize the measurement time of 2D-XRD, which is practically determined based on researchers' experiences, not on the literature [2]. Although un-optimization could be a bottleneck in high-throughput experiments, little research was reported to reduce measurement time with machine learning [3]. This chapter addresses this problem with a density-based analysis. In addition, a method to separate diffraction patterns from noises with density-based clustering is discussed. This method separates a diffraction signal one by one, which makes the fitting of a diffraction peak with a profiling function computationally cheaper.

Firstly, the signal density of 2D-XRD images of bulk and thin film samples are compared with OPTICS in terms of S/N ratio. Then, the signal density is evaluated whether it represented propensities of microstructures of thin films. An appropriate measurement time is also discussed with OPTICS. Thereafter, a procedure was proposed how to separate diffraction patterns from noises with the density-based clustering technique called density-based spatial clustering of applications with noise (DBSCAN) [4]. Finally, a graph representation of a 2D-XRD image with the signal density was studied as a promising feature candidate to represent microstructures of thin films.

5.2 Method

5.2.1 2D-XRD Images

2D-XRD images were captured with Vantec 500 (Bruker AXS) and D8 Discover system, which detected a part of the Debye-Scherrer ring (2θ and χ angles) two-dimensionally. Measured samples were poly crystalline bulk silicon (Poly-Si), doped FeCoMn alloy (FCM-X) thin films, and indium gallium oxide (IGO) thin films [5]. FCM-X ($(FeCoMn)_xX_{(1-x)}$) and IGO ($((Ga_{1-x}In_x)_2O_3)$) were composition spreads fabricated using combinatorial synthesis [6]. All the 2D-XRD images shown in this chapter were applied with gamma correction to improve readability, and the gamma value was 5.

5.2.2 Conversion of 2D-XRD Images into Scatter Plots

2D-XRD images were converted into scatter plots of one unit length square so that they reserved diffraction intensities. For example, if the size of 2D-XRD image was 2048 × 2048 pixels, and signal intensity at (208, 512) pixel was 10, then 10 points were plotted at the coordinate (208/2048, 512/2048) in a scatter plot. This multiple plots mean dense for clustering algorithms.

5.2.3 Density-Based Analyses with DBSCAN and OPTICS

DBSCAN is a clustering method which classifies data points into core points, neighbouring points and noise points. Core points are data points which contains more data points than a criteria in its ε-neighbourhood. Neighbouring points are data points which contains core points and less data points than the criteria in its ε-neighbourhood. Other data points are classified as noise points. The criteria which separates core and neighbouring points is a hyperparameter of DBSCAN. Hyperparameters of DBSCAN in this chapter is explained in Section 5.2.4.

OPTICS is a similar method to DBSCAN, but not a clustering method. OPTICS is a method to determine ε of DBSCAN depending on the criteria. Hyperparameters of OPTICS are minimum points and a maximum of ε. The value of minimum points is the criteria to determine core points of DBSCAN, therefore we set the same value as DBSCAN applications. The other parameter, a maximum of ε, can be infinity because this value is a maximum value to calculate whether a data point is core point or not. In this chapter we set the maximum of ε to 26/2048, which almost corresponds to 0.5° in 2θ angle, to reduce the computational time.

A result of OPTICS is represented in a form of lists, which stores sets of a data-point number, a cluster-ordering and a reachability distance. The list is analysed in a form of a reachability plot, whose horizontal and vertical axes represent cluster-ordering and reachability distance, respectively. Example reachability plot is Figure 5.2. Cluster-ordering is a value related to the position in the original plot. Reachability distance represents the minimum of ε required for the data point to be a core point. Therefore, each valley in a reachability plot represents the dense region in the original plot.

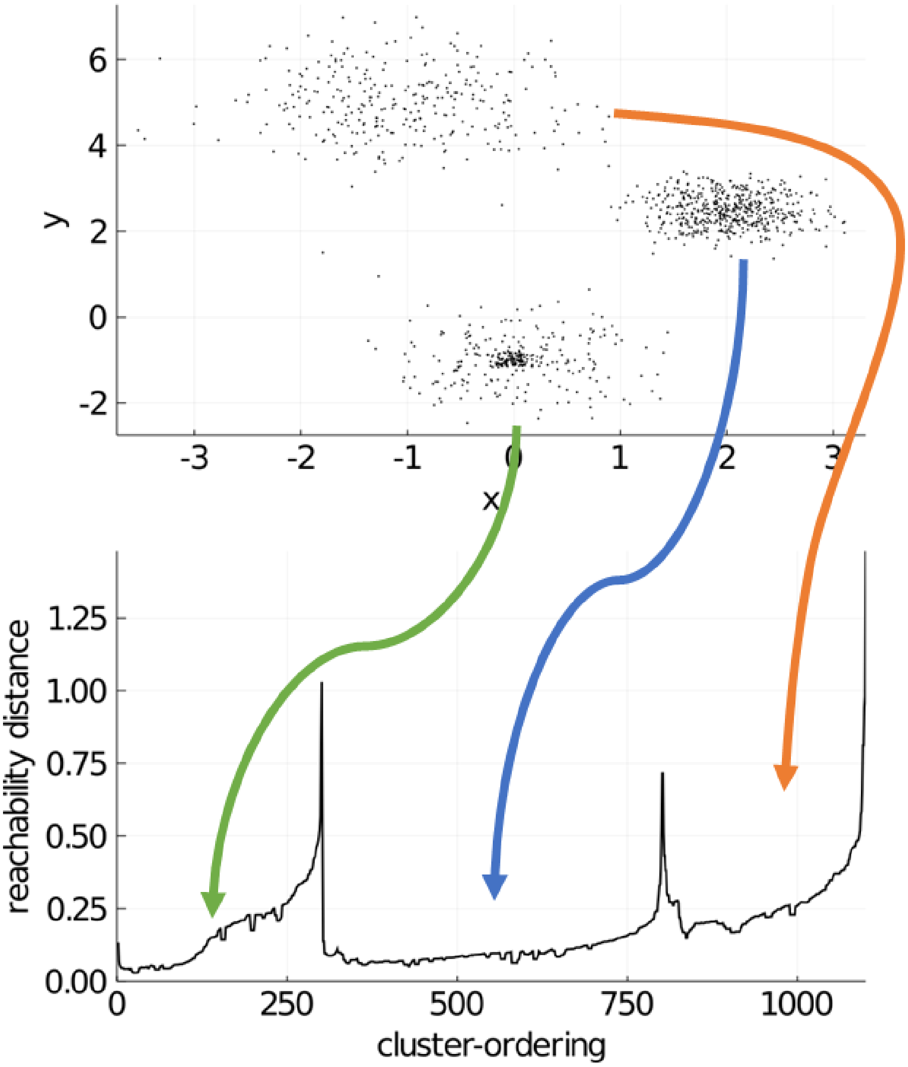

Figure 5.2 Sample scatter plot (top) and its reachability plot (bottom). Each valley represents dense region in the original plot. This figure is a reproduction from Accepted Manuscript version of the article, A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Accelerating two-dimensional X-ray diffraction measurement and analysis with density-based clustering for thin films," Japanese Journal of Applied Physics, 2021, 60, SCCG04, https://doi.org/10.35848/1347-4065/abf2d8 ©2021 IOP Publishing and Japan Society of Applied Physics

5.2.4 Hyperparameters of DBSCAN

Hyperparameters of DBSCAN are minimum points and ε. The value of minimum points was the same as that of OPTICS applications. Each ε was determined by referring to the corresponding reachability plot. We also used minimum cluster size to omit smaller clusters as noise from the analysis. All parameters in this chapter is summarized in Table 5.1. Note that orders of magnitude of minimum points and minimum cluster size are important rather than exact values.

Table 5.1 Hyperparameters of DBSCAN in this chapter. This table is a reproduction from Accepted Manuscript version of the article, A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Accelerating two-dimensional X-ray diffraction measurement and analysis with density-based clustering for thin films," Japanese Journal of Applied Physics, 2021, 60, SCCG04, https://doi.org/10.35848/1347-4065/abf2d8 ©2021 IOP Publishing and Japan Society of Applied Physics

Data Name	ε	Minimum points	Minimum cluster size
Poly-Si	20/2048	268	268
FCM-X 180 s	0.003	50	1000
FCM-X 20 s	0.009	50	500
FCM-X 60 s	0.0055	50	3000
FCM-X 900 s	0.0015	50	100000

5.2.5 Graph Representation of 2D-XRD Images

2D-XRD images were first converted into scatter plots. Then Gaussian mixture model (GMM) was applied to the scatter plots with setting the number of kernels to 30. We used sci-kit learn v. 0.20.0, intel-python distribution for the GMM application. A base node was added at the middle point of the right edge of a 2D-XRD image. The all GMM kernels of each image were connected to the base node to constitute a graph of a 2D-XRD image (Figure 5.3). Features of each edge and node were relative position of the connecting nodes and the variance of the GMM kernel, respectively.

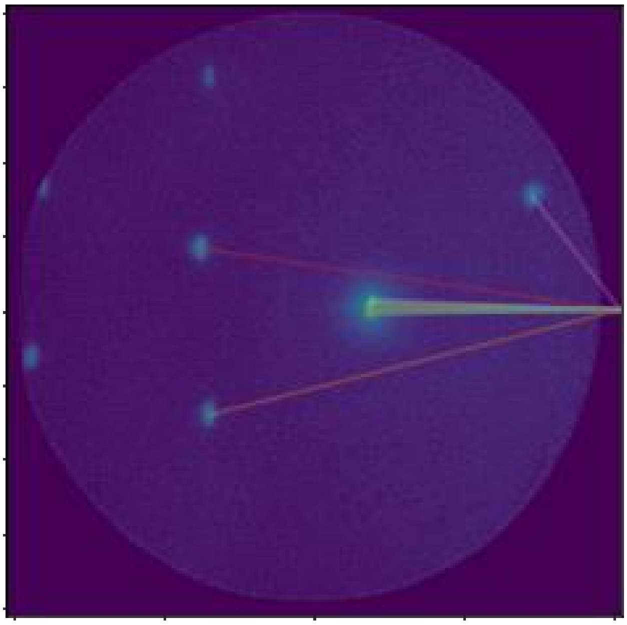

Figure 5.3 Example graph of a 2D-XRD image. Each node represents a main diffraction signal. Strong diffraction signals were allocated multiple kernels.

We modelled variational graph autoencoder with PyTorch and PyTorch Geometric. The number of graph convolutional layers was three. The pooling layer was global mean pool. The number of hidden layers was one and its dimensionality was four. The prior distribution was 2D normal distribution whose mean and variance were 0 and 1, respectively.

5.3 Results and Discussion

5.3.1 Density Analyses of 2D-XRD Images with OPTICS

To investigate propensities of the signal density of a 2D-XRD image, we compared reachability plots of Poly-Si (bulk) and FCM-X (thin film) (Figure 5.4). Note that differences in reachability distances over samples are basically meaningless because average density of noises in a 2D-XRD image may change based on the detection voltage which is automatically adjusted for each measurement. We selected 120 s measurement data of Poly-Si and 60 s measurement data of FCM-X for comparison because the reachability distances of their noises were close values to each other. As shown in Figure 5.4, the valley of FCM-X was narrow and shallow compared with the valleys of Poly-Si. This means that the S/N ratio of a sample is represented with the size and depth of a valley in the reachability plot.

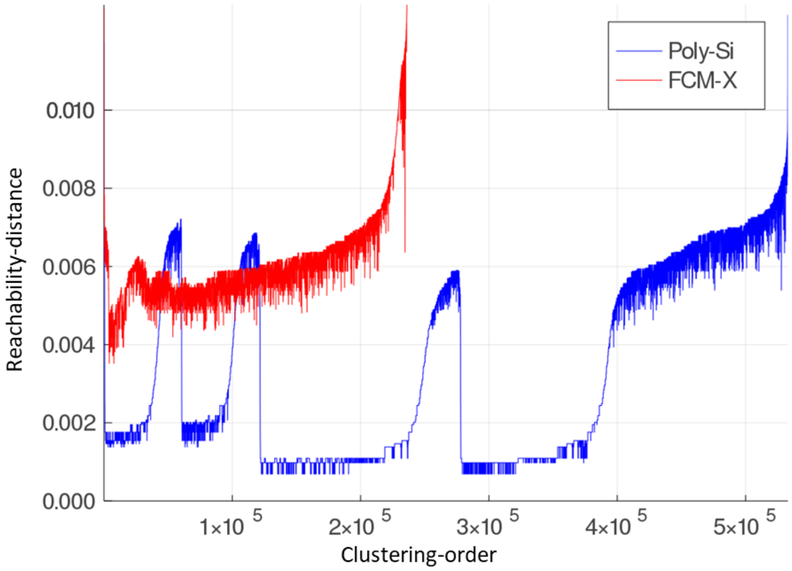

Figure 5.4 Reachability plot comparison of Poly-Si (blue) and FCM-X (red). This figure is a reproduction from Accepted Manuscript version of the article, A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Accelerating two-dimensional X-ray diffraction measurement and analysis with density-based clustering for thin films,” Japanese Journal of Applied Physics, 2021, 60, SCCG04, https://doi.org/10.35848/1347-4065/abf2d8 ©2021 IOP Publishing and Japan Society of Applied Physics

We then applied OPTICS to 2D-XRD images of IGO to investigate the relationship between the signal density and crystallinities of a sample. As shown in the left column of Figure 5.5, the crystal structures of the thin film became higher oriented from (a) to (c). In the same way, the bottom of the valley in the reachability plots in the right column became wider as indicated by the green arrows (the inner plots are zoomed plots). In the conventional method, the crystallinity is characterized with the full width of the half maximum of diffraction peaks over 2θ angle or χ angle using high S/N data. This is another reason why measurement time of XRD of thin films tends to be longer compared to bulk samples. Considering this, an analysis with the signal density will reduce measurement time to characterize the crystallinity. We also confirmed that internal stress (weaker diffraction intensity on the diffraction axis) in thin film was indicated by a small mountain in a valley in reachability plot (Figure 5.6).

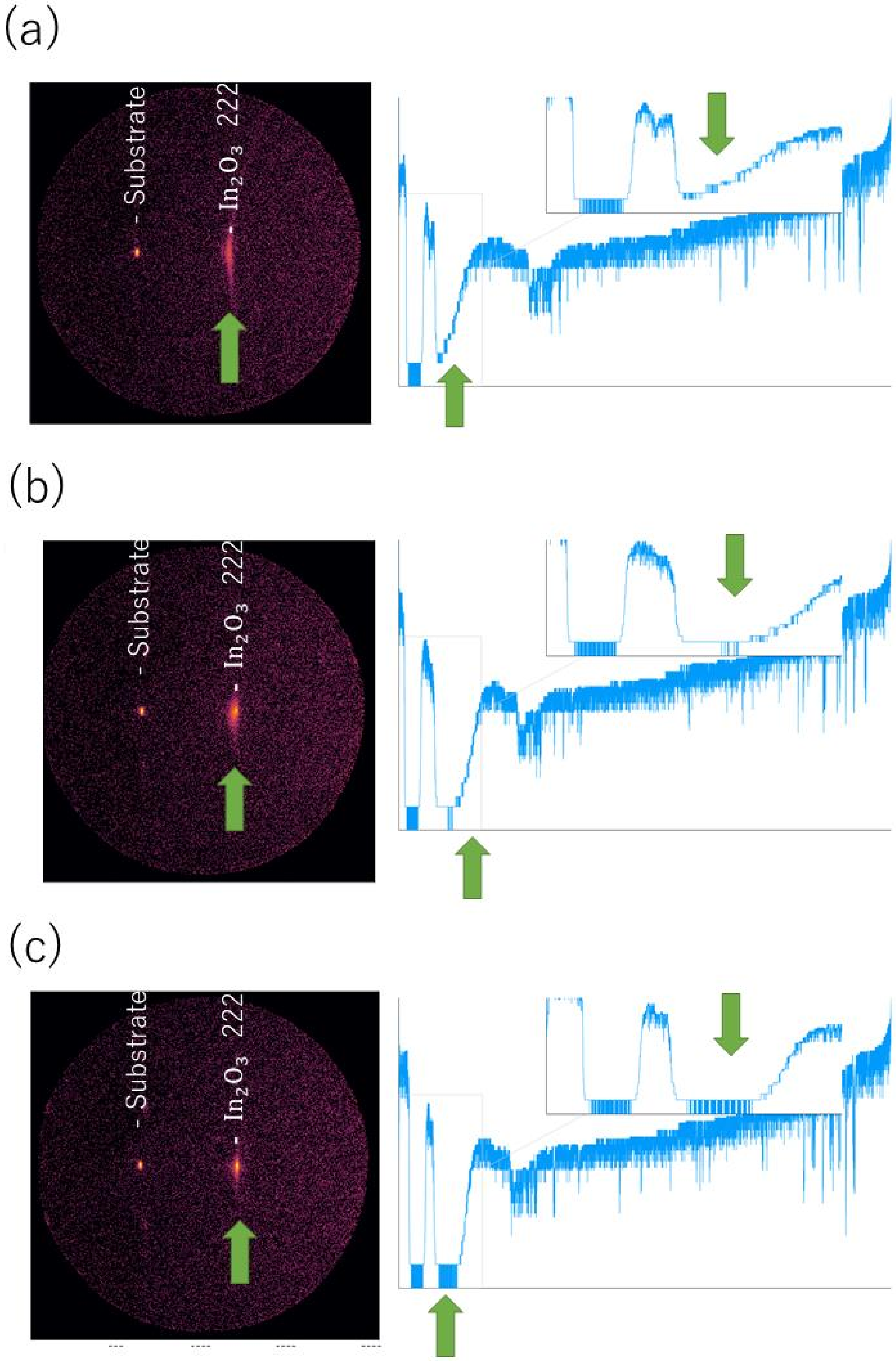

Figure 5.5 Comparison of orientations of In_2O_3 crystals in a composition spread with reachability plots. The original 2D-XRD images are listed in the left column and the corresponding reachability plots are in the right column. Inner plots in the reachability

plots are zoomed plots in the grey box regions. From (a) to (c), composition ratio of Ga_2O_3 decreases and the crystal of In_2O_3 became higher oriented. This figure is a reproduction from Accepted Manuscript version of the article, A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Accelerating two-dimensional X-ray diffraction measurement and analysis with density-based clustering for thin films,” Japanese Journal of Applied Physics, 2021, 60, SCCG04, https://doi.org/10.35848/1347-4065/abf2d8 ©2021 IOP Publishing and Japan Society of Applied Physics

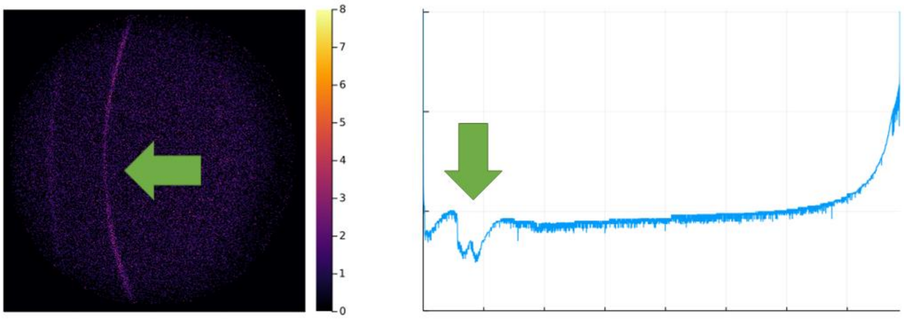

Figure 5.6 Representation of inner stress by a 2D-XRD image and the corresponding reachability plot. This figure is a reproduction from Accepted Manuscript version of the article, A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Accelerating two-dimensional X-ray diffraction measurement and analysis with density-based clustering for thin films,” Japanese Journal of Applied Physics, 2021, 60, SCCG04, https://doi.org/10.35848/1347-4065/abf2d8 ©2021 IOP Publishing and Japan Society of Applied Physics

5.3.2 Signal/Noise Separation with DBSCAN

In the previous section, we confirmed that diffraction patterns are represented as valleys in reachability plots, which suggests that diffraction signals can be separated from noises referring to their signal densities. We propose the procedure to separate diffraction signals from background noises, then fit each diffraction signal with a profiling function (Figure 5.7). Note that, from this section, some 2D-XRD images were flipped for the consistency of the direction of the diffraction angle with 1D-XRD.

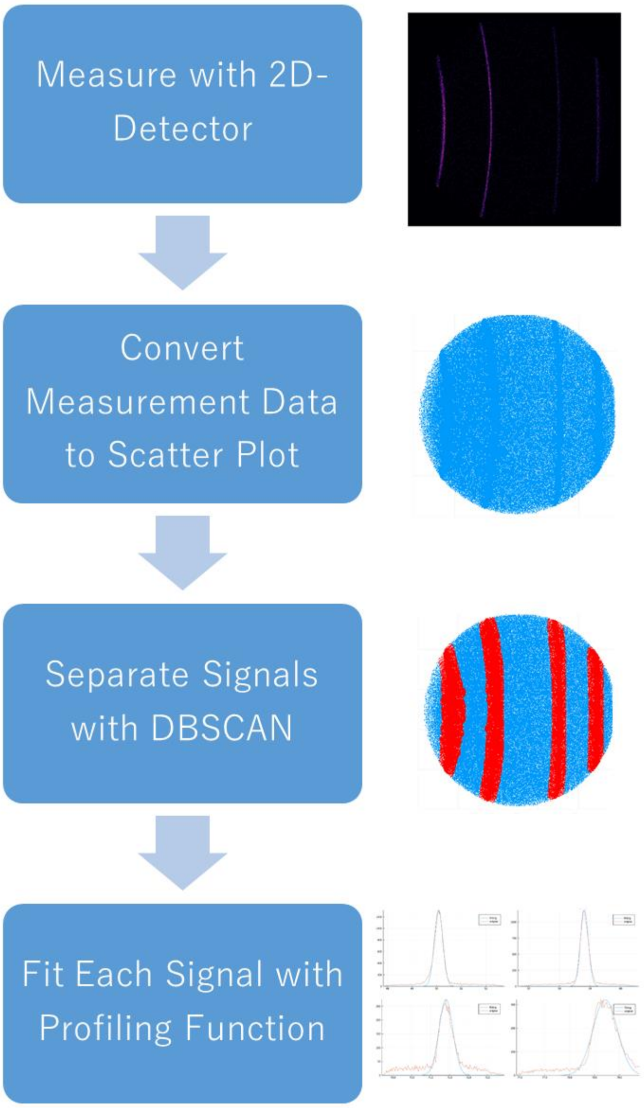
Measure with 2D-Detector
Convert Measurement Data to Scatter Plot
Separate Signals with DBSCAN
Fit Each Signal with Profiling Function

Figure 5.7 Procedure to separate diffraction patterns from background noises with a density-based clustering. This figure is a modified reproduction from Accepted Manuscript version of the article, A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Accelerating two-dimensional X-ray diffraction measurement and analysis with density-based clustering for thin films,” Japanese Journal of Applied Physics, 2021, 60, SCCG04, https://doi.org/10.35848/1347-4065/abf2d8 ©2021 IOP Publishing and Japan Society of Applied Physics

As a proof of concept, we applied our method to 2D-XRD image of Poly-Si. Figure 5.8(a) and (c) show the 2D-XRD image and the XRD chart. All the XRD charts in this study were produced by integration over χ angle ranged from −7.0° to 7.0°. Figure 5.8(b) is the result of DBSCAN application: red and blue points are clustered and noise points, respectively. We integrated each separated signal (red points) over χ angle and fitted each peak with a Gaussian function. The fitted results were concatenated to constitute the XRD chart [Figure 5.8(d)]. Each fitting results are shown in Figure 5.9. Referring to Figure 5.8(c) and (d) we concluded that the method separated signals with appropriate regions.

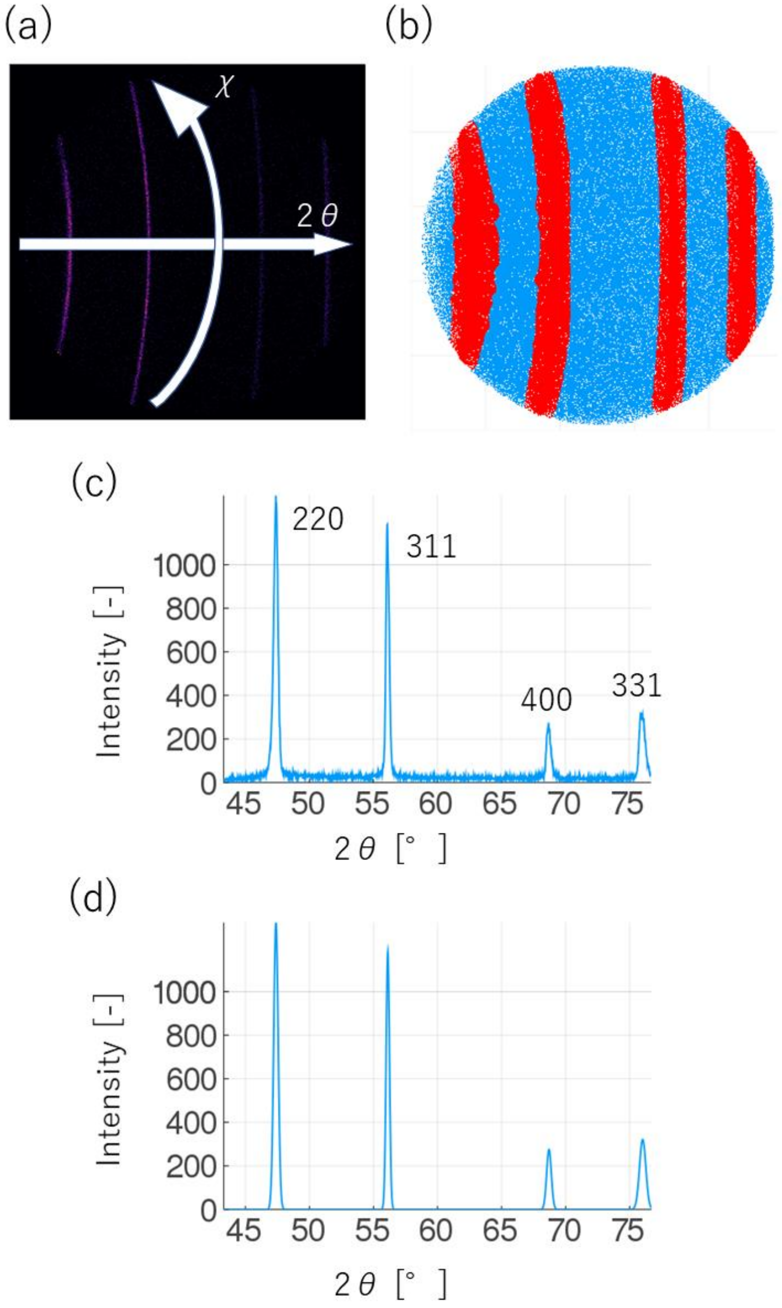

Figure 5.8 Results of Poly-Si. (a) the original 2D-XRD image, (b) a scatter plot of the DBSCAN result, the red points belong to valid clusters, the blue points are noises and omitted from the analysis, (c) the XRD chart of the original 2D-XRD image, (d) XRD chart using our method and fitted with Gaussian functions. This figure is a reproduction from Accepted Manuscript version of the article, A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Accelerating two-dimensional X-ray diffraction measurement and analysis with density-based clustering for thin films," Japanese Journal of Applied

Physics, 2021, 60, SCCG04, https://doi.org/10.35848/1347-4065/abf2d8 ©2021 IOP Publishing and Japan Society of Applied Physics

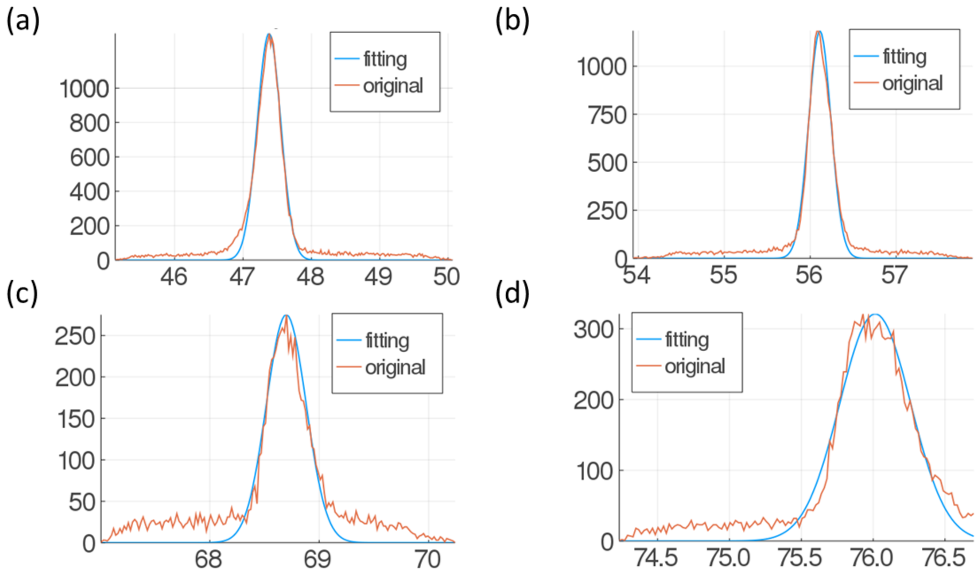

Figure 5.9 Fitting results of Poly-Si diffraction patterns, raw signal (orange) and fitting result (blue). This figure is a reproduction from Accepted Manuscript version of the article, A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Accelerating two-dimensional X-ray diffraction measurement and analysis with density-based clustering for thin films," Japanese Journal of Applied Physics, 2021, 60, SCCG04, https://doi.org/10.35848/1347-4065/abf2d8 ©2021 IOP Publishing and Japan Society of Applied Physics

Next, we applied our method to 2D-XRD of FCM-X to evaluate the method limitation. As shown in Figure 5.10(a), the diffraction pattern of the sample is very weak compared to the noise intensity. Even with this low S/N, our method separated diffraction pattern signals [Figure 5.10(b)]. Furthermore, the density analysis is advantageous compared to the method which separates diffraction pattern signals from noises with thresholds of intensity (Figure 5.11). This is because the diffraction intensity from thin films are at the almost same level to noise signals.

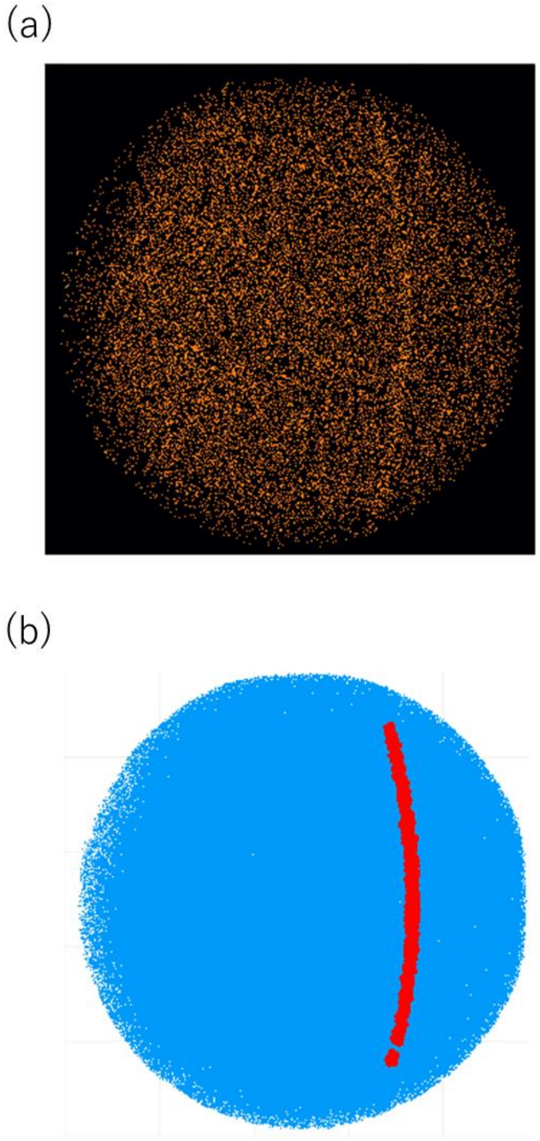

Figure 5.10 (a) 2D-XRD image of FCM-X, the crystal system of the sample is bcc, (b) scatter plot of the DBSCAN result, the red points correspond to the diffraction pattern signal, the blue points are noise signals. This figure is a reproduction from Accepted Manuscript version of the article, A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Accelerating two-dimensional X-ray diffraction measurement and analysis with density-based clustering for thin films," Japanese Journal of Applied Physics, 2021, 60, SCCG04, https://doi.org/10.35848/1347-4065/abf2d8 ©2021 IOP Publishing and Japan Society of Applied Physics

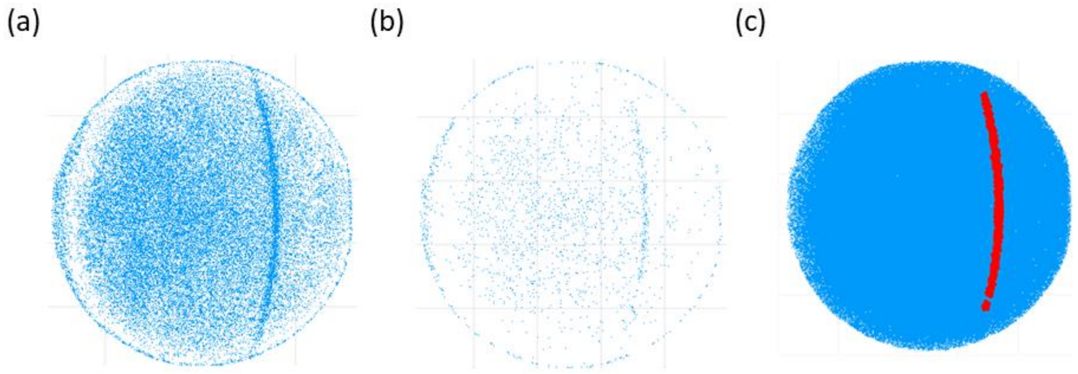

Figure 5.11 Comparison of threshold and density methods. (a) and (b) are separation results by setting thresholds 1 and 2, respectively. (c) is the density result [same as Figure 5.10(b)].

5.3.3 Evaluation of measurement time

To evaluate the appropriate measurement time in terms of the signal density, FCM-X was measured by changing its measurement time from 20 to 900 s and applied the separation method (Figure 5.12). The sample and the measurement position were the same as those in Figure 5.10 and Figure 5.11. In all cases, DBSCAN separated the diffraction pattern signals with parameter adjustments. The parameters are listed in Table 5.1.

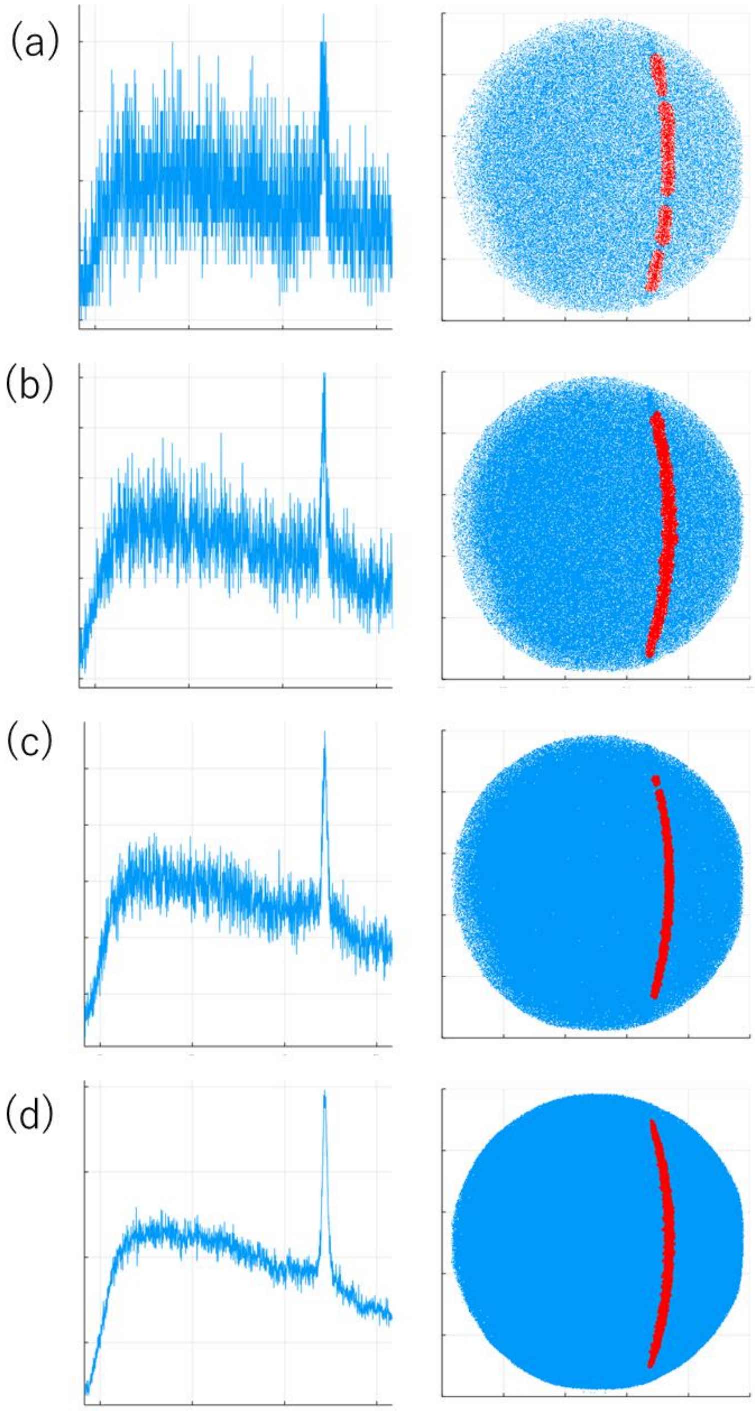

Figure 5.12 Comparison among FCM-X results by changing measurement time from 20 to 900 s, each row corresponds to one measurement, plots in the left column are XRD charts of the original 2DXRD images, scatter plots in the right column are DBSCAN results. Measurement times were (a) 20 s, (b) 60 s, (c) 180 s, and (d) 900 s. This figure is a reproduction from Accepted Manuscript version of the article, A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Accelerating two-dimensional X-ray

Referring to the S/N ratios, 900 s (possibly 180 s for some analyses) was the appropriate measurement time for FCM-X with the conventional XRD analysis or measurements with point detectors. However, in terms of the signal density, 60 s was enough duration to separate the diffraction pattern signals. Because the detection of the diffracted X-ray with point detectors is estimated to follow the Poisson distribution [7], measurements with point detectors consume a lot of time for their counts converging to enough precision. In contrast, our approach is based on the assumption that this convergence is far faster in space compared with in time, and this is why we conclude that 60 s was enough duration.

In terms of time optimization for practical measurements, proper measurement time should be moderate duration to obtain a certain depth of valleys in a reachability plot. Figure 5.13 shows the reachability plots of the result of Figure 5.12. Referring to the reachability plots, shorter measurement time may be better than longer. This is because long measurement time makes noise regions denser and valleys of diffraction patterns relatively shallow. Therefore, proper measurement time will be shorter compared with the conventional methods. The proper metrics of density difference should be discussed for further discussion over measurement time optimization, and this would be the future work.

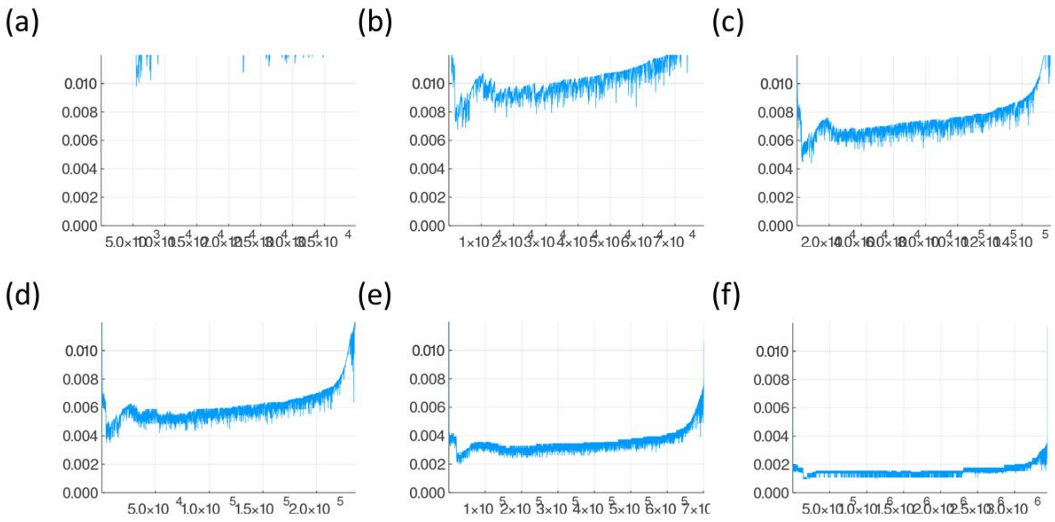

Figure 5.13 Reachability plots of FCM-X, measurement times were (a) 10 s, (b) 20 s, (c) 40 s, (d) 60 s, (e) 180 s, and (f) 900 s. All measured samples are the same as Figure 5.10. This figure is a reproduction from Accepted Manuscript version of the article, A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Accelerating two-dimensional X-ray diffraction measurement and analysis with density-based clustering for thin films," Japanese Journal of Applied Physics, 2021, 60, SCCG04, https://doi.org/10.35848/1347-4065/abf2d8 ©2021 IOP Publishing and Japan Society of Applied Physics

5.3.4 Remarks for Practical Applications

Because OPTICS is a computationally expensive algorithm, researchers should limit the number of applications as well as setting maximum value of ε in the practical study. For example, in the measurement of composition spread, because noises are supposed to be similar intensities over all measurement points, a single application of OPTICS will be enough to determine parameters of DBSCAN.

5.4 Future Work

Graph neural network (GNN) [8,9] is an emerging tool in data-driven materials research [10–13]. GNN is applicable to almost all domains as long as a data point can be represented as a graph (a set of nodes and edges). This section shortly discusses the application of GNN to 2D-XRD images.

Chapter 2 discussed the feature extraction of 2D-XRD images with NMF. Although the NMF features represented propensities of microstructures of thin films, they are hard to represent continuous changes such as peak shift. A graph representation of a 2D-XRD image will be a solution if it represents diffraction signals and their positions as nodes and edges, respectively. To confirm this assumption, we converted 2D-XRD images into graphs and modelled a graph variational autoencoder (GVAE) [14] under basic algorithms without fine tuning of hyperparameters. Figure 5.14 shows the latent space of GVAE. Red and blue points are latent variables of samples fabricated by PLD and sputtering, respectively. The tendency that latent variables of samples fabricated by PLD were distributed broader than those by sputtering is consistent with the results in Chapters 3 and 4. Even with the basic application, GVAE learned propensities of PLD and sputtering. Because the current graph representation failed to capture weak diffraction patterns and broadness of the diffraction patterns over χ angle, further study is required.

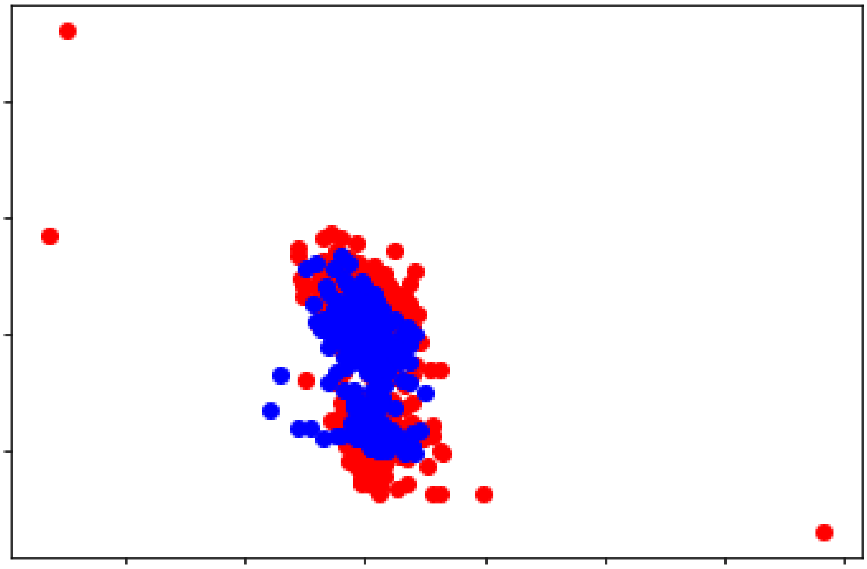

Figure 5.14 Latent space of GVAE. Red and blue points represent samples fabricated by PLD and sputtering, respectively.

5.5 Conclusion

In this chapter, we have confirmed that the signal density of 2D-XRD images represents a lot of information on microstructures of samples. The signal density represents not only differences among diffraction pattern signals and noises, but also other information such as crystallinities or internal stress of crystal structures. A significant advantage of density based representation is that these information can be obtained with short measurement times or very low S/N data, which is beneficial to research on ultra-thin films. In addition, analyses with the signal density omit arbitrariness of the integration range of 2D-XRD images, which is inevitable in the conventional analysis.

We also demonstrated that combining OPTICS and DBSCAN can separate diffraction pattern signals from noises even under short measurement time of thin films. OPTICS and DBSCAN are unsupervised learning, which does not require any big data. Therefore, researchers can apply the method without preparing a large amount of 2D-

XRD images in advance. Although our method requires three hyper parameters, one parameter (minimum points) is not so important, and the other parameters can be estimated using OPTICS. Our approach will work with other density-based clustering methods, such as hierarchical DBSCAN. 2D-XRD images of various samples, e.g., alloy, ceramic, and bulk, can be analysed in the similar way.

We also confirmed that the graph representation of signal density of 2D-XRD images will be better features for deep learning models than NMF features. These results shows that the signal density is a candidate feature of 2D-XRD images to represent microstructures of samples.

References

[1] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, *OPTICS: Ordering Points to Identify the Clustering Structure*, ACM SIGMOD Record **28**, 49 (1999).

[2] M. Steinhart and J. Pleštic, *Possible Improvements in the Precision and Accuracy of Small-Angle X-Ray Scattering Measurements*, Journal of Applied Crystallography **26**, 591 (1993).

[3] K. Saito, M. Yano, H. Hino, T. Shoji, A. Asahara, H. Morita, C. Mitsumata, J. Kohlbrecher, and K. Ono, *Accelerating Small-Angle Scattering Experiments on Anisotropic Samples Using Kernel Density Estimation*, Scientific Reports **9**, 1526 (2019).

[4] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (AAAI Press, Portland, OR, 1996), pp. 226–231.

[5] T. Nagata, T. Hoga, A. Yamashita, T. Asahi, S. Yagyu, and T. Chikyow, *Valence Band Modification of a (Ga x In 1− x) 2 O 3 Solid Solution System Fabricated by Combinatorial Synthesis*, ACS Combinatorial Science **22**, 433 (2020).

[6] H. Koinuma and I. Takeuchi, *Combinatorial Solid-State Chemistry of Inorganic Materials*, Nature Materials **3**, (2004).

[7] B. B. He, *Two-Dimensional X-Ray Diffraction* (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2018).

[8] F. Scarselli, M. Gori, Ah Chung Tsoi, M. Hagenbuchner, and G. Monfardini, *The Graph Neural Network Model*, IEEE Transactions on Neural Networks **20**, (2009).

[9] M. Gori, G. Monfardini, and F. Scarselli, *A New Model for Learning in Graph Domains*, in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks* (IEEE, Montreal, 2005).

[10] T. Xie and J. C. Grossman, *Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties*, Physical Review Letters **120**, 145301 (2018).

[11] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, *Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals*, Chemistry of Materials **31**, (2019).

[12] V. Fung, J. Zhang, E. Juarez, and B. G. Sumpter, *Benchmarking Graph Neural Networks for Materials Chemistry*, Npj Computational Materials **7**, 84 (2021).

[13] K. Hatakeyama-Sato and K. Oyaizu, *Integrating Multiple Materials Science Projects in a Single Neural Network*, Communications Materials **1**, 49 (2020).

[14] T. N. Kipf and M. Welling, *Variational Graph Auto-Encoders*, ArXiv (2016).

Chapter 6

Conclusions

6.1 Conclusion of the research

This thesis studied features of 2D-XRD images to represent microstructures of thin films for machine learning models. Microstructures of thin films are deterministic parameters of material properties; thus, they are indispensable variables for data-driven research on thin films. Although measurement data of microstructures can be rapidly produced by a combination of high throughput experiments and 2D-XRD measurements, the characterization of them is time-consuming. Therefore, this thesis studied how to represent microstructures by data-driven manners to circumvent the problem.

In Chapter 2, NMF is evaluated as a feature extraction method of 2D-XRD images. The NMF features represented crystallinity differences of thin films, which reflected differences of fabrication conditions as well as those of substrate types. Chapter 3 describes that the NMF features improved the performance of VAE so that the model learned distinct borders among some substrate types in its latent space. In Chapter 4, fabrication conditions of indium gallium oxide thin films were visualized with the methods of Chapters 2 and 3. The visualization revealed that structural changes of samples by sputtering and PLD were different, which indicates that sputtering is better process for research on indium gallium oxide thin films than PLD is. In Chapter 5, the signal density of 2D-XRD images is evaluated as another feature candidate. This feature optimized measurement time of 2D-XRD measurements. In addition, graph representation with the density features is a promising feature of 2D-XRD images because this representation will capture continuous shifts of diffraction patterns which the NMF features did not.

These results contribute to utilize outputs of high throughput experiments for data-driven materials research. Because thesis is a case study of 2D-XRD images, further study on feature extraction of other measurements is still required for advances of data-driven materials research.

6.2 Outlook

This thesis mainly focuses propensities of features of 2D-XRD images. Therefore, prediction of some material properties are not discussed. This section describes some details of predictive approaches with the features as follows:

- Prediction of crystallinities under untested fabrication conditions: The prediction model will be a generative model such as VAE. However the input will be a problematic because conditions of fabrication of thin films vary depending on processes such as PLD, sputtering, or metal organic chemical vapour deposition. Concatenation of those conditions will be a sparse and high dimensional input vector. Therefore, NMF will be an appropriate feature extraction method for the input. The output will be NMF or graph features of 2D-XRD images.
- Combination with other data types: Practical materials researches employ various measurements as well as XRD. Concatenating features of those measurements (not raw measurement data) will be an input of a machine learning model.

- Density based analysis for qualification: The separation method of diffraction signals in Chapter 5 will be beneficial to qualifications of products, but must be tested statistically in another study.
- Prediction of materials properties from microstructures: The input will be NMF features or density features discussed in this thesis. The output changes based on the task. The model may be determined by referring to the task or the literature. Another way to determine the model is to simply choose the best accuracy model with packages such as MLJ.jl or PyCaret which train multiple models at a single instruction.

6.3 Industrial Perspective

This thesis proposes methods to extract features from 2D-XRD images of thin films and to visualize them for fabrication analyses. Although the most samples were thin films of indium gallium oxide, the research focus was on methods to treat raw measurement data in data science frameworks. Therefore, the findings in this thesis are applicable to 2D-XRD images of other materials. This section describes contributions of this thesis to materials and semiconductor industries.

Materials industry will be beneficiated from essential parts of this thesis. This thesis explains the method how to input a high-dimensional and small-sized dataset into machine learning models. This is a common situation in almost all sections in materials industry and a server problem in competitive sections such as batteries or pharmaceuticals. As this thesis has described, a solution is feature extraction. However, the importance of it has not well recognized because the feature extraction itself will not discover any novel materials. This thesis helps researchers realize the importance of feature extraction. Considering this, the thesis also will open new consultancy in materials industry.

Semiconductor industry will be practically beneficiated from this thesis. Semiconductor thin films are widely implemented in the modern society, such as logic and memory devices in IT products and power devices and sensors in electric vehicles. These devices are so called final products. This thesis will contribute to more upstream processes, mainly to the deposition process. The main product of this process is multi-layer thin films on a wafer, which are supposed to constitute transistors. These layers are so thin that diffraction intensities in the XRD measurement are very weak, thus, the out-of-plane measurement is not feasible. Although the in-plane measurement is applicable, it loses spatial resolution in the film, thus positions of defects cannot be identified. Therefore, microstructures of them are rarely measured in the production line for now. As discussed in Chapter 5, the density-based method is applicable to such low S/N samples even with the out-of-plane measurement and reduces the measurement time of films to around 1/15. Therefore, the XRD measurement in the production line will be feasible. Feature extraction by NMF (Chapter 2) will be useful in evaluation whether measured points contain defects or not. This evaluation will be calculated within milliseconds by the method in Section 2.2.3. This identifying positions of defects is important to reduce the size of chips to increase the yield of chips per wafer. In addition, the detection of defects in upstream processes is profitable in semiconductor industry because current semiconductor devices are fabricated through large number of complex processes. Therefore, although such thin layers were not measured, this thesis will contribute whole semiconductor industry through the deposition process.

Appendix A

Additional Information for Chapter 2

A.1 Feature Vector Overview with a Correlation Matrix

To verify whether the feature vectors represent propensities of the whole dataset, we computed Pearson's correlation coefficients among the feature vectors and represented the results in a matrix form [Figure A1 (a)]. The results were arranged according to the substrate type and the order was as follows: c-sapphire (data number 1–228), YSZ (111) (229–394), YSZ (100) (395–479), SiO_2 (480–490), STO (111) (491–501), and STO (100) (502–512). Feature vectors on the same substrate type have a relatively strong correlation compared to those on different substrate types. With this figure, we noticed 11 wrongly labeled data, which were from one composition-spread sample on the YSZ (100) substrate but labeled as a c-sapphire sample. Some samples on the c-sapphire substrates (data number 180–228) were weakly correlated with other samples on the same substrate type. This is because these data have a major weight in factor 4, not in factor 6. Some feature vectors of c-sapphire have a stronger correlation with the samples on the YSZ (111) substrate. This may be because the 2D-XRD images measured from the Ga_2O_3 thin films on c-sapphire and YSZ (111) substrates were mainly background noise and the signal of the substrate.

We also found that despite the same substrate composition, the difference in the crystal plane is implied by the correlations of the feature vectors. This finding is consistent with the fact that in this study, the lattice constant of the substrate, and not electric polarity, is the important factor in the fabrication of the thin films. This shows that NMF is a good feature extraction method for 2D-XRD signal datasets containing multiple substrate types.

Considering the results, feature vectors represent mainly substrate-type differences. The histogram of the correlation coefficients [Figure A1 (b)] supports this assumption by showing that the data from the same substrate type have a strong correlation and are mainly over 0.5. In contrast, the other substrate data have a slight correlation and are approximately 0.0. This suggests that the distribution of the feature vectors has some kind of hierarchy; therefore, applying a hierarchical clustering technique will reveal a hierarchy in which the fabrication condition differences are less than the substrate type.

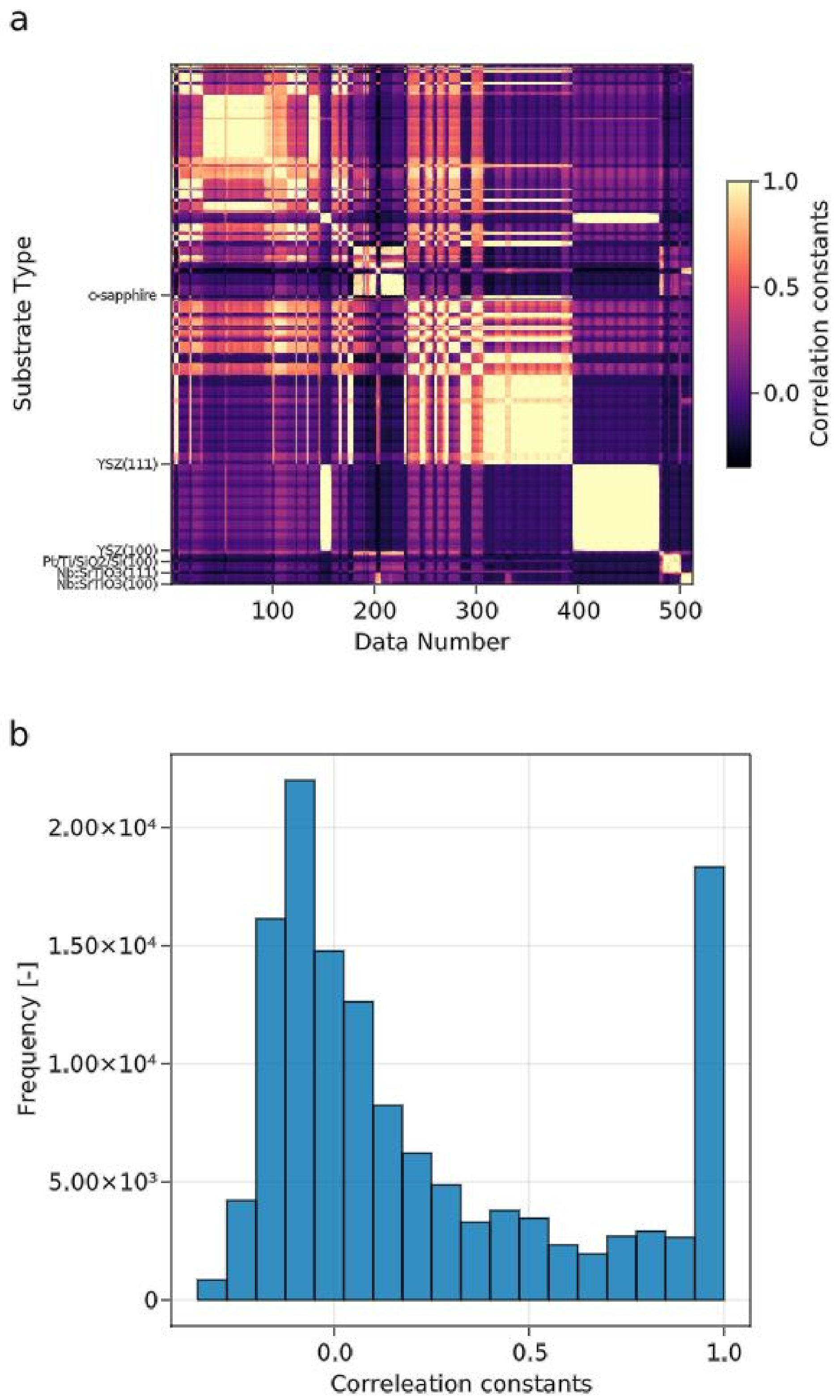

Figure A1. (a) Correlation constants of the feature vectors. Both x-axis and y-axis represents data ID; however, the y-axis indicates the borders of the substrate types. (b) Histogram of correlation constants. This figure is a reproduction under Creative

Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis,” Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

A.2 Supplemental Figures

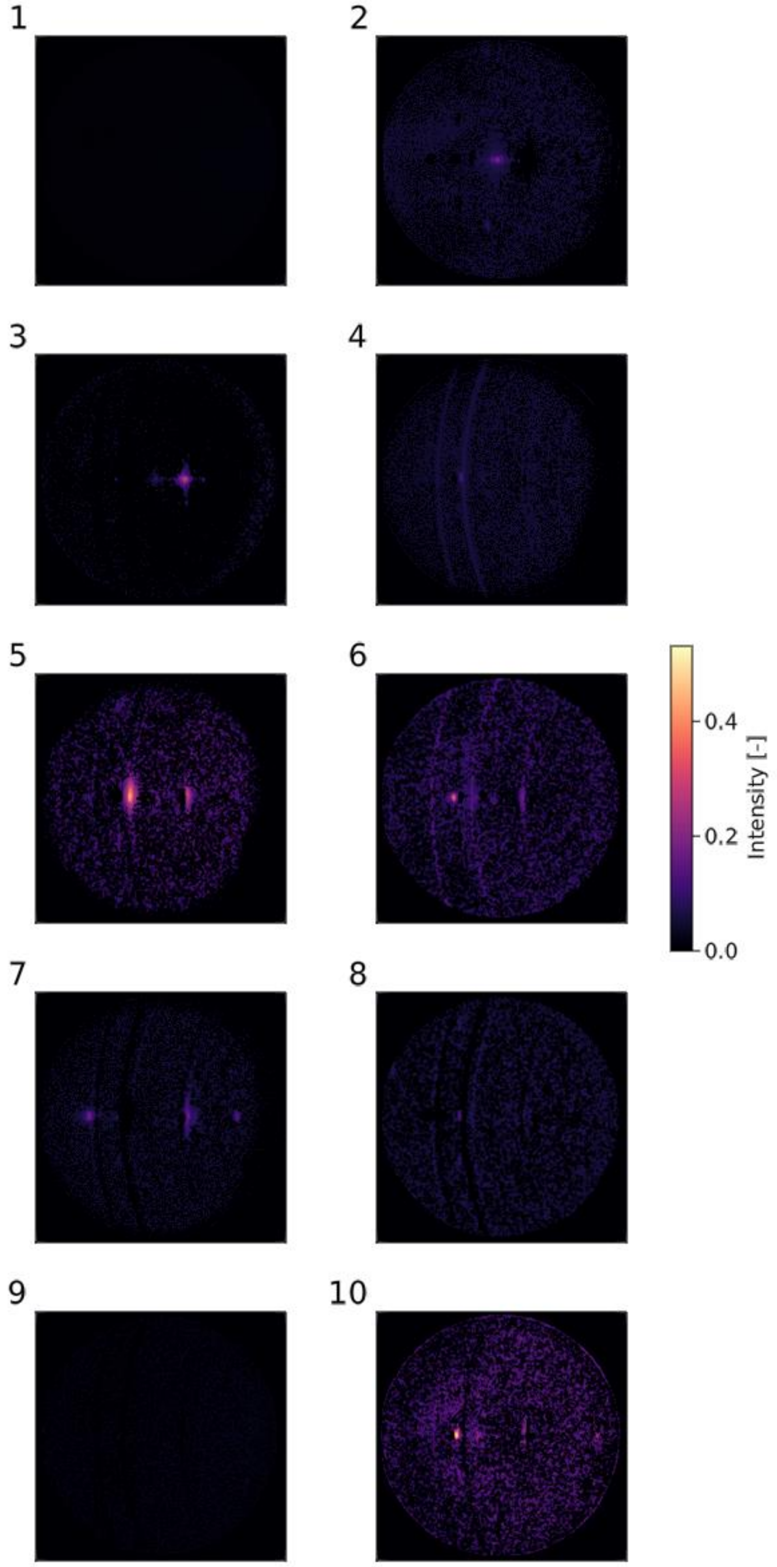

Figure A2. Factor images with the shared colour scale. All images are corrected at γ = 5. This figure is a reproduction under Creative Commons License 4.0 (CC BY)

https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis,” Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

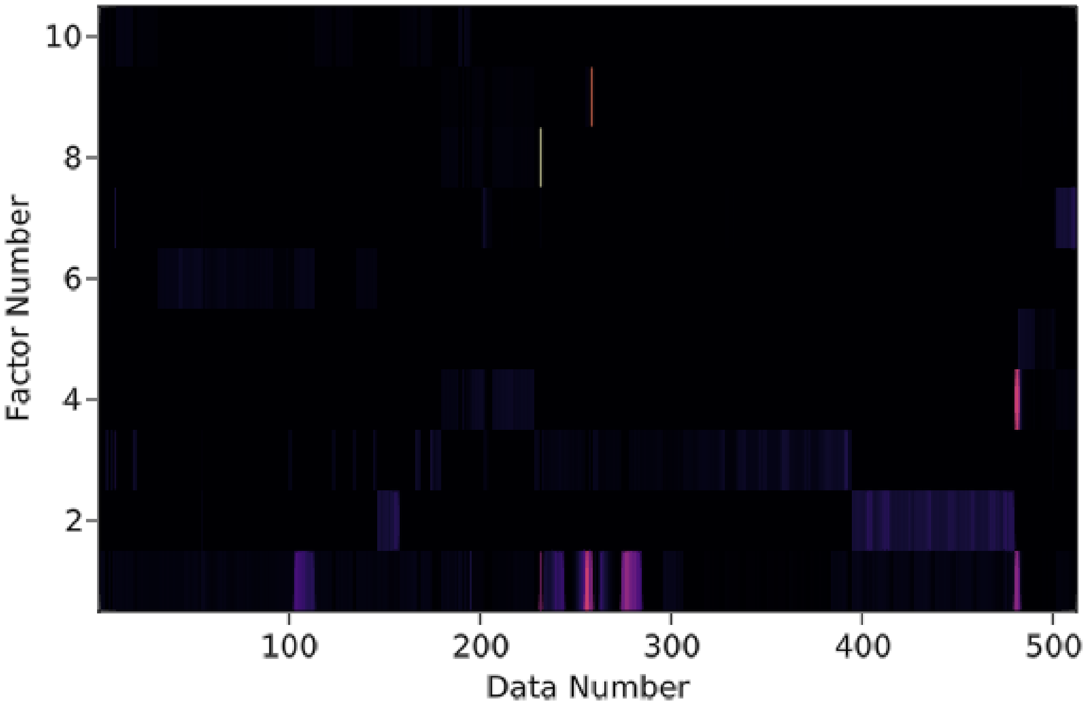

Figure A3. Heatmap of feature vectors of PLD samples. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis,” Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

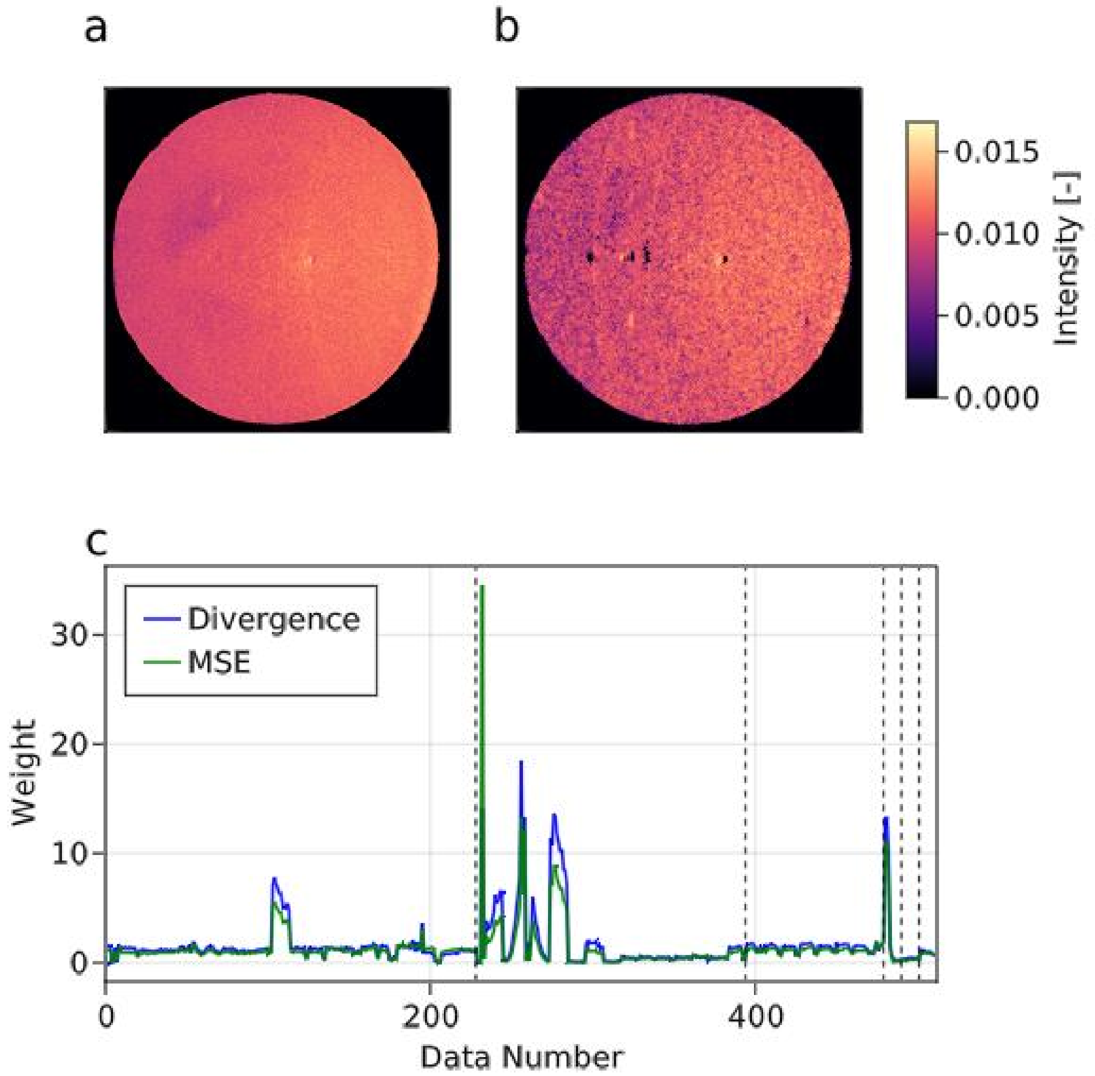

Figure A4. Objective function comparison in the case of the factor 1. (a) and the blue line in (c) is the result of the divergence case. (b) and the green line in (c) is that of the mean squared error (MSE) case. The colour scale of (a) is identical to that of (b). The dashed lines in (c) indicates borders of substrate types. The order of the substrate types is c-sapphire, YSZ (111), YSZ (100), Pt/Ti/SiO$_2$/Si (100), STO (111) and STO (100). This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis,” Science and

Technology of Advanced Materials: Methods, in press,

https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

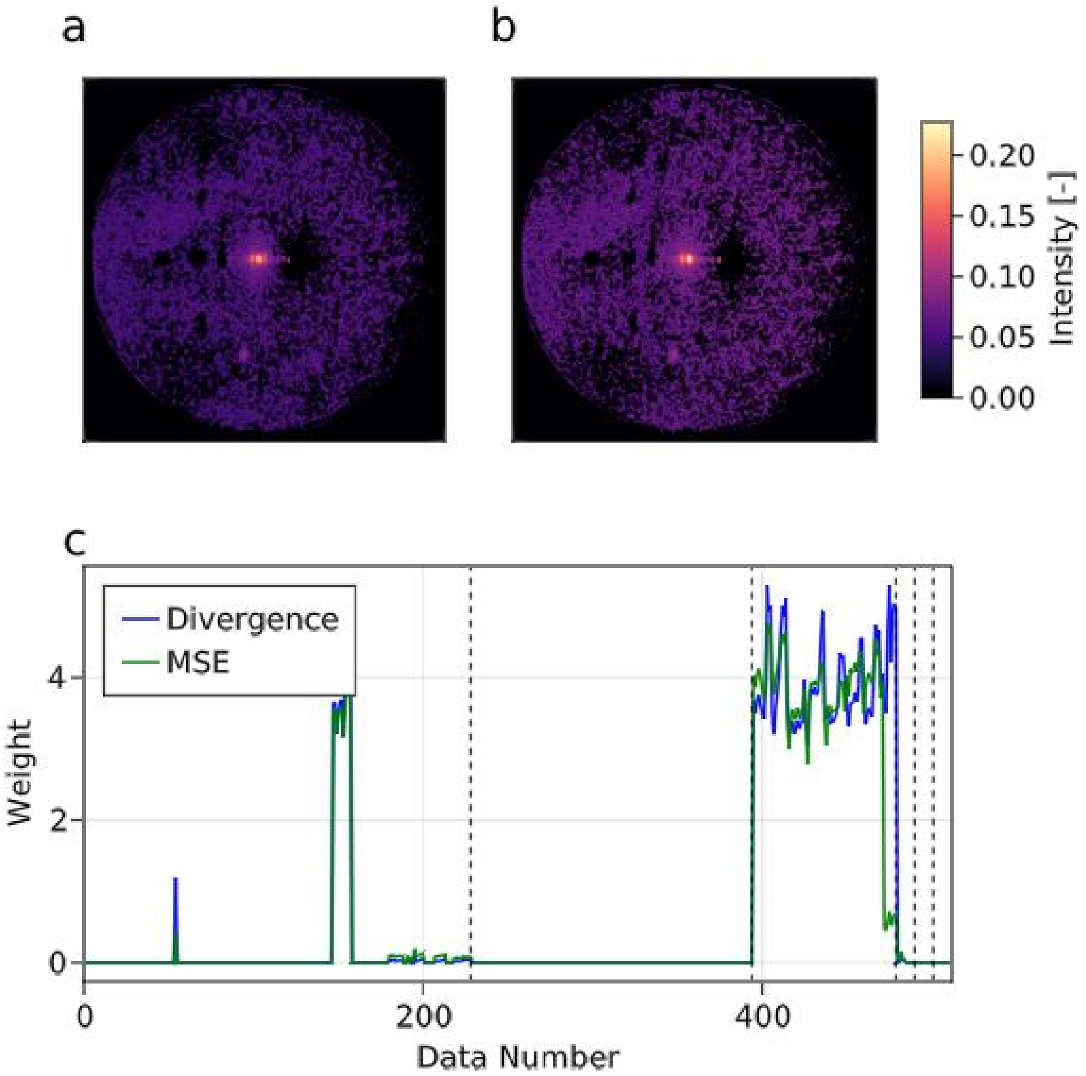

Figure A5. The result of the factor 2. The displaying manner is identical to Figure A4. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis," Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

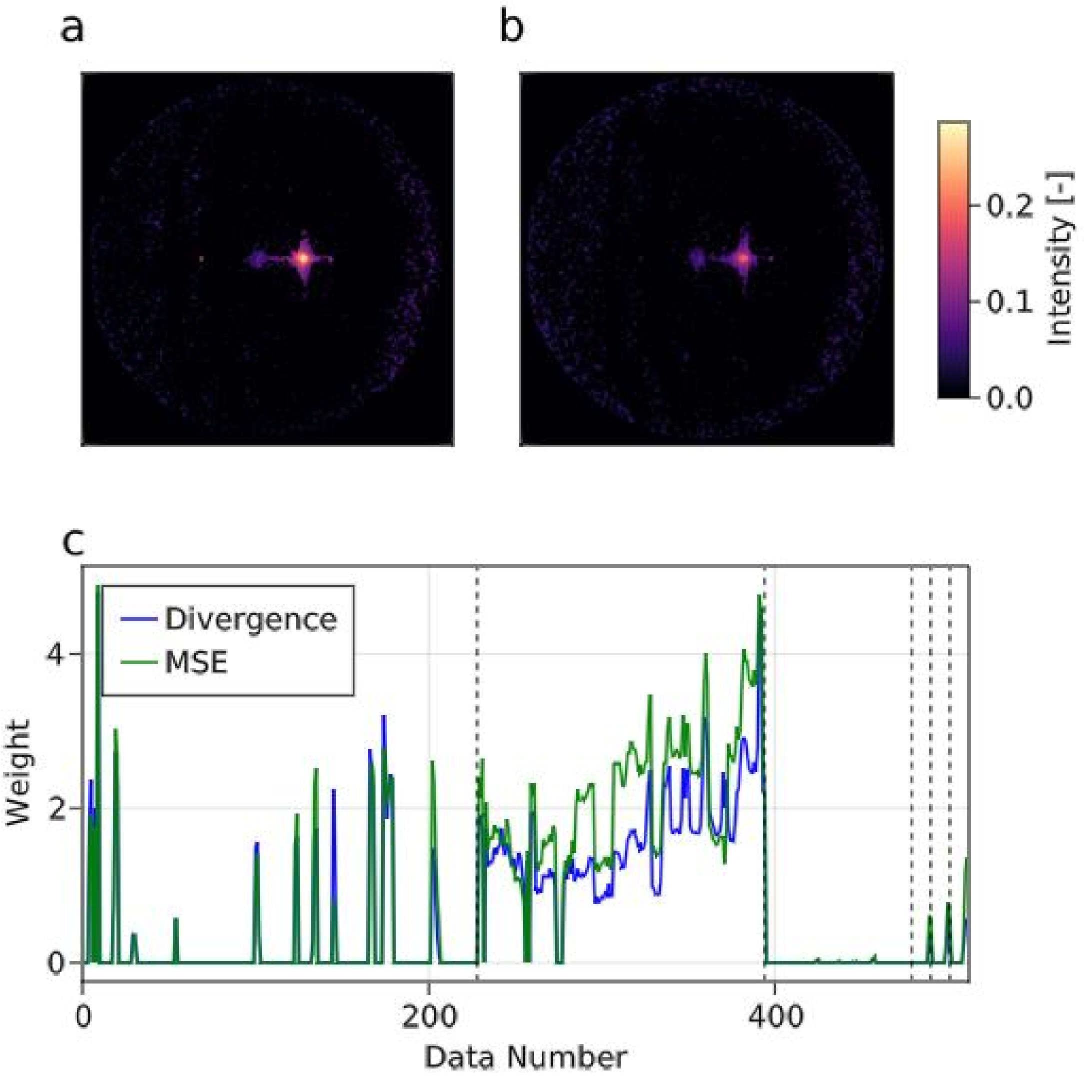

Figure A6. The result of the factor 3. The displaying manner is identical to Figure A4. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis,” Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

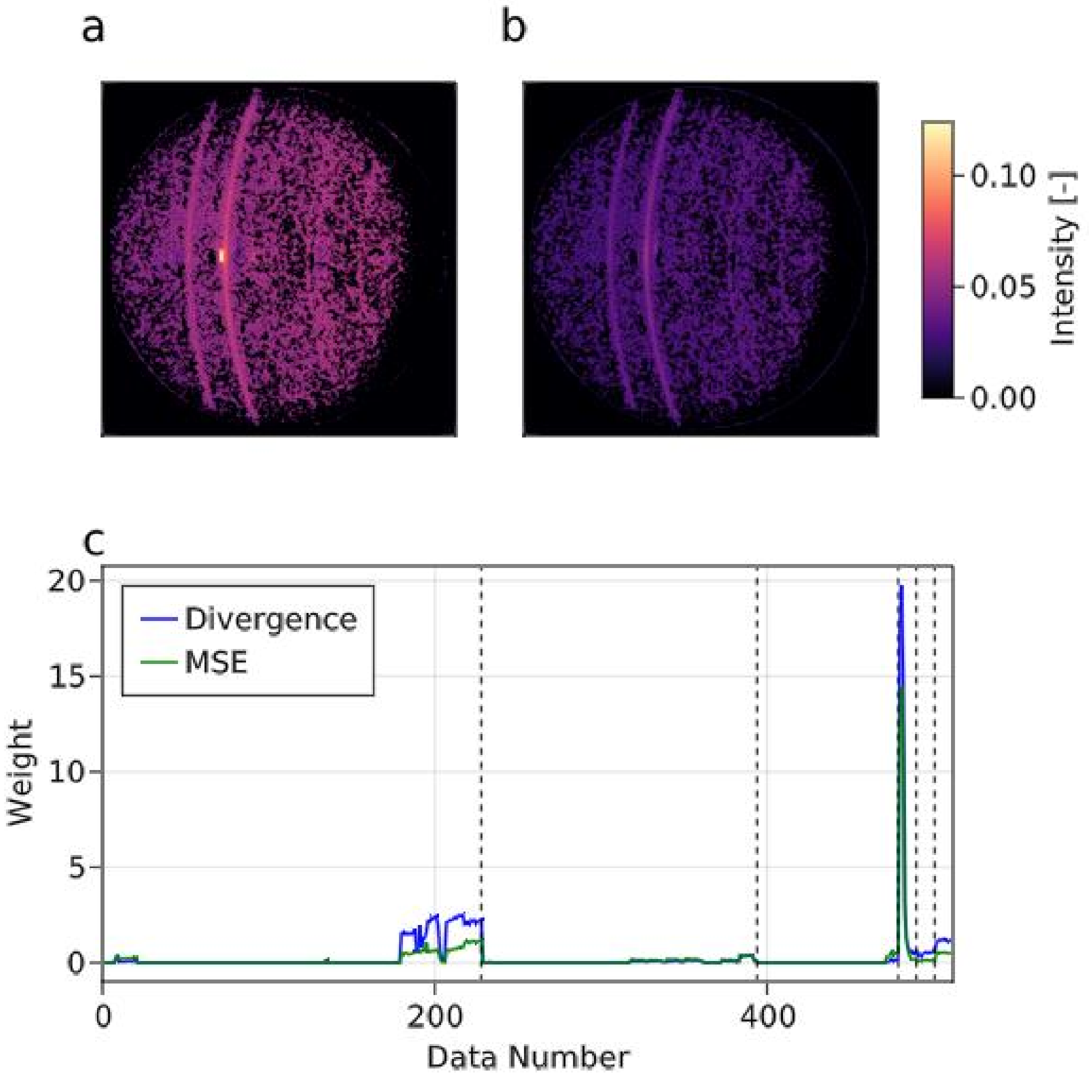

Figure A7. The result of the factor 4. The displaying manner is identical to Figure A4. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis," Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

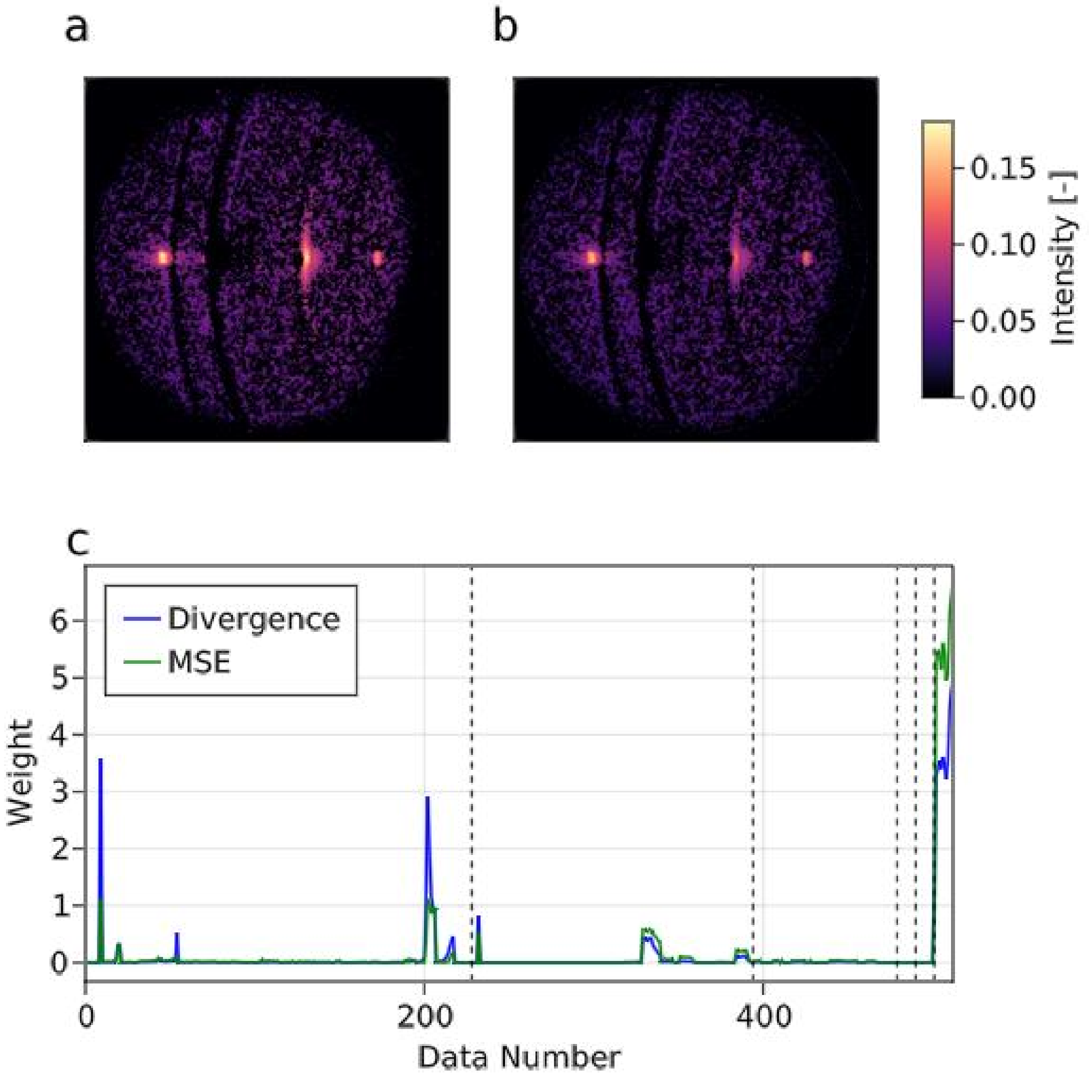

Figure A8. The result of the factor 7. The displaying manner is identical to Figure A4. The results of the factors 5 and 6 are explained in the main text. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis," Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

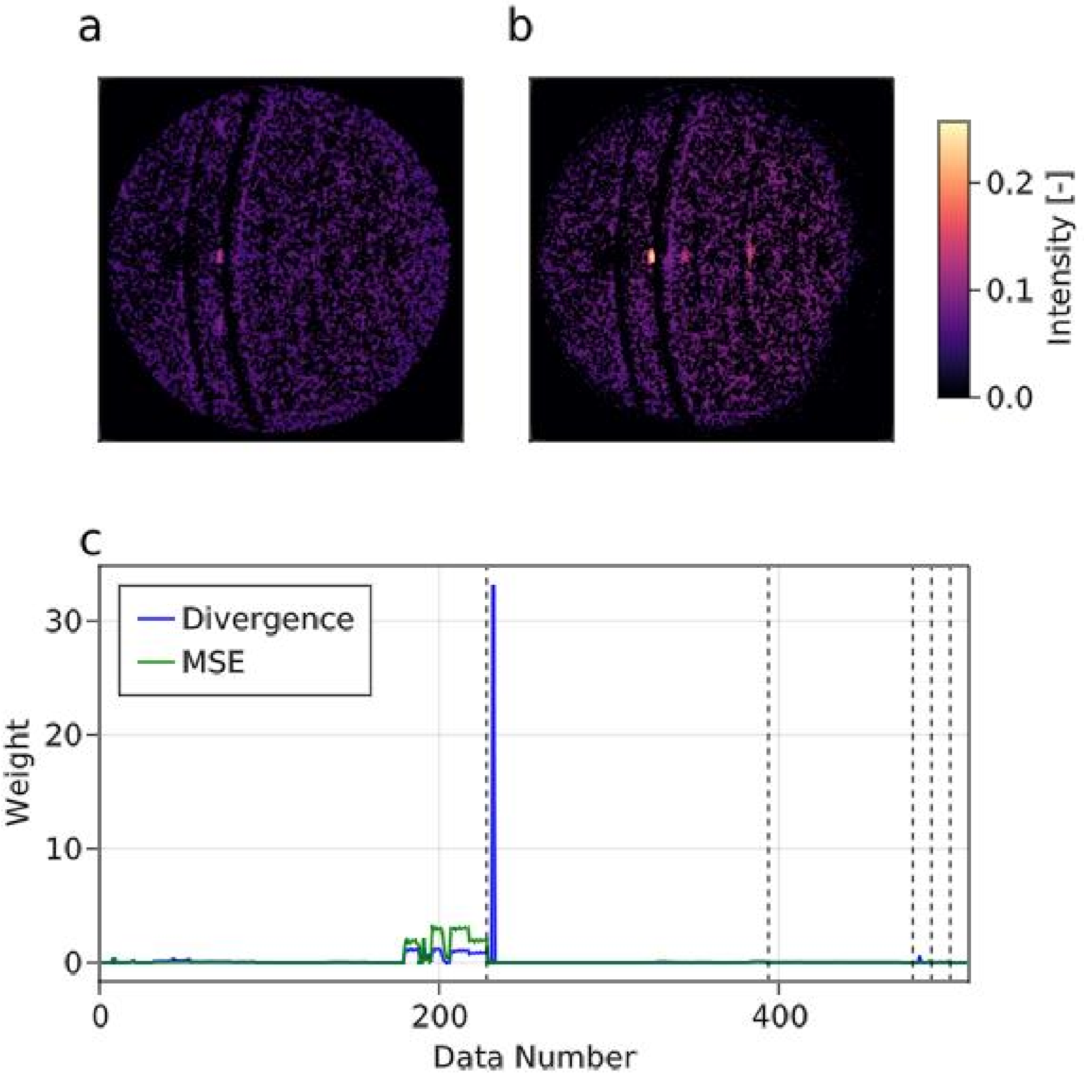

Figure A9. The result of the factor 8. The displaying manner is identical to Figure A4. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, "Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis," Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

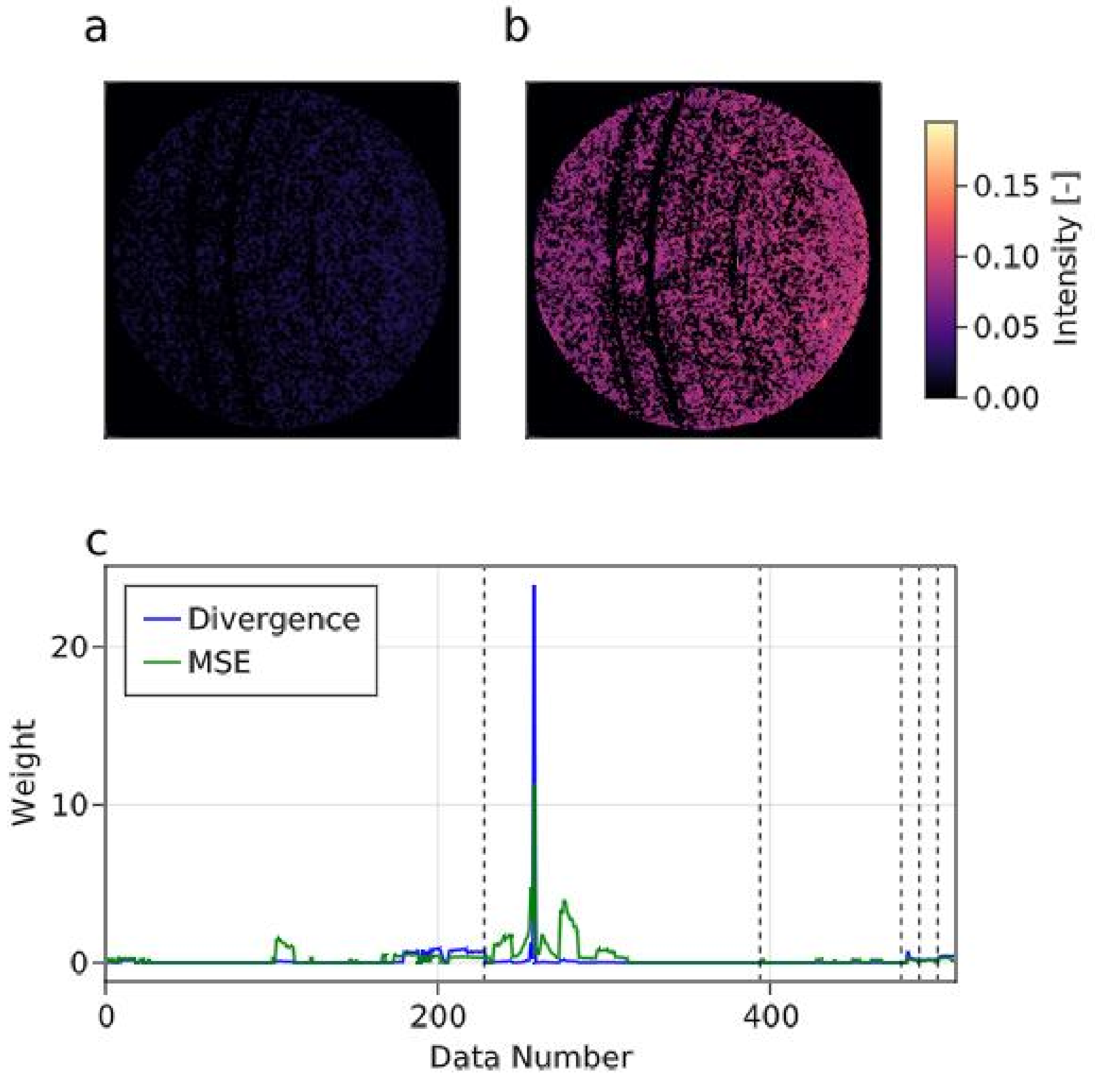

Figure A10. The result of the factor 9. The displaying manner is identical to Figure A4. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis,” Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

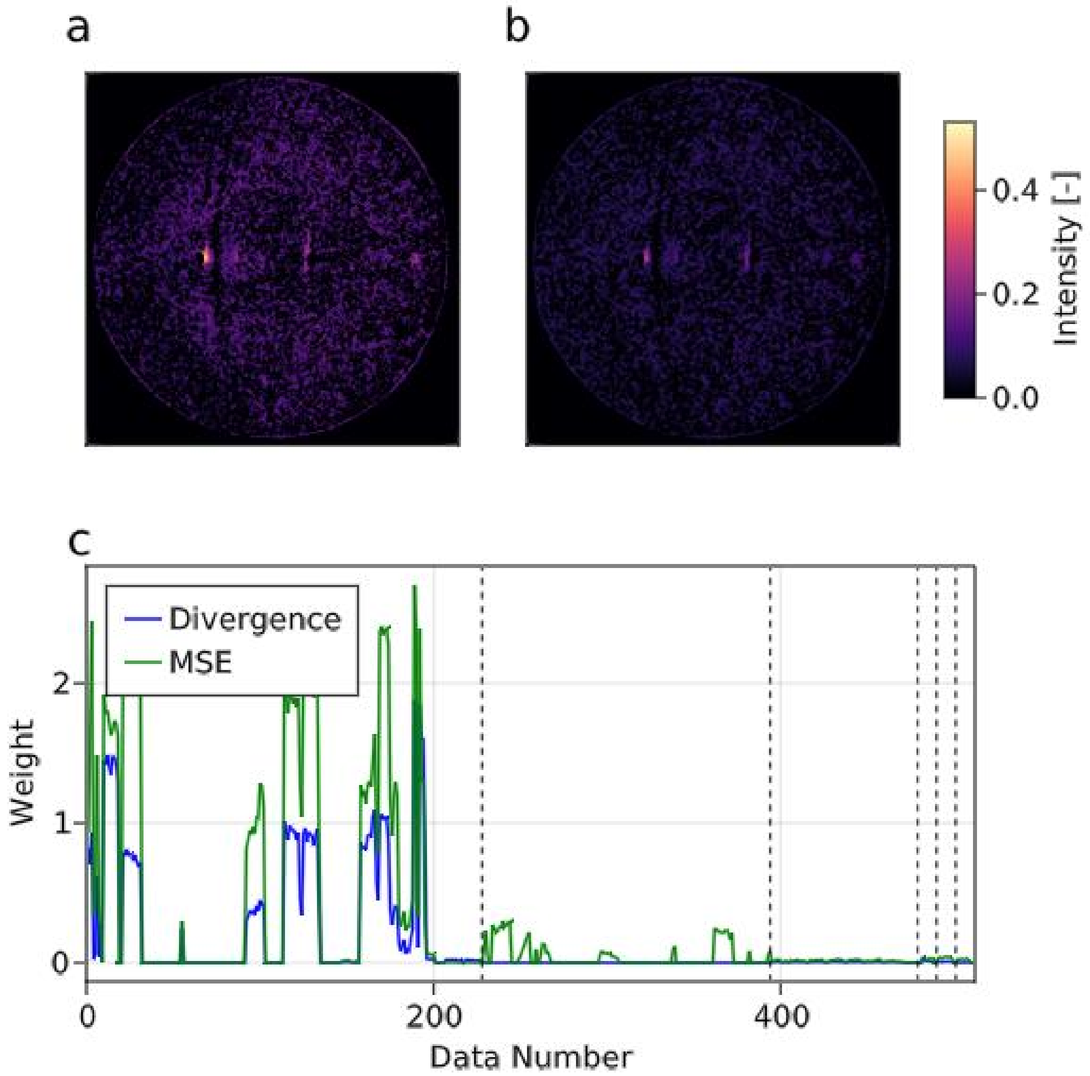

Figure A11. The result of the factor 10. The displaying manner is identical to Figure A4. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis,” Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

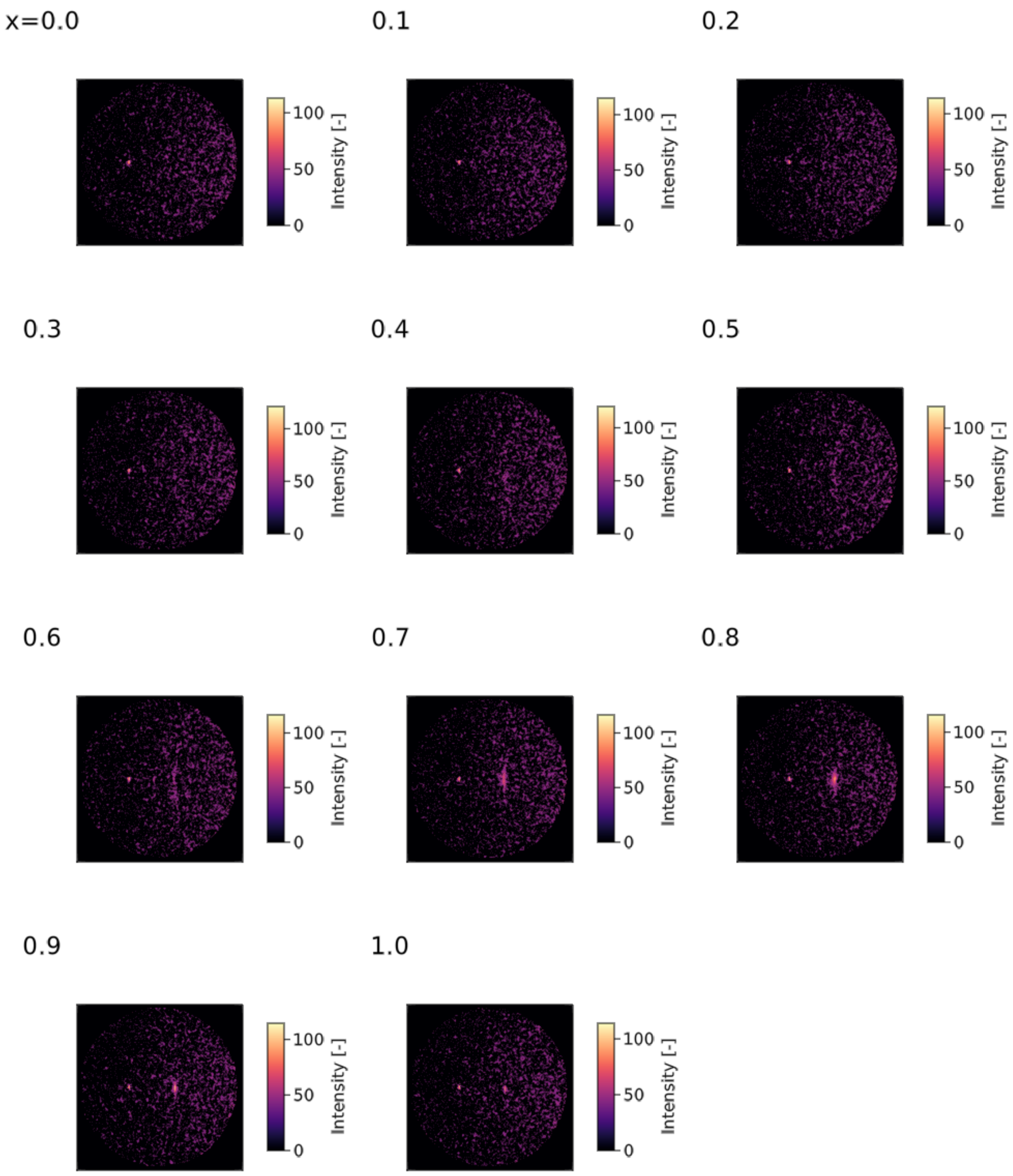

Figure A12. 2D-XRD images of the $(Ga_{1-x}In_x)_2O_3$ composition spread discussed in Section 2.3.3 and Figure 2.8 in the main text. Note that the intensities are different from other 2D-XRD images in this thesis because this figure shows original 2D-XRD images, not normalized ones.

Appendix B

Additional Information for Chapter 3

B.1 Supplemental Figures

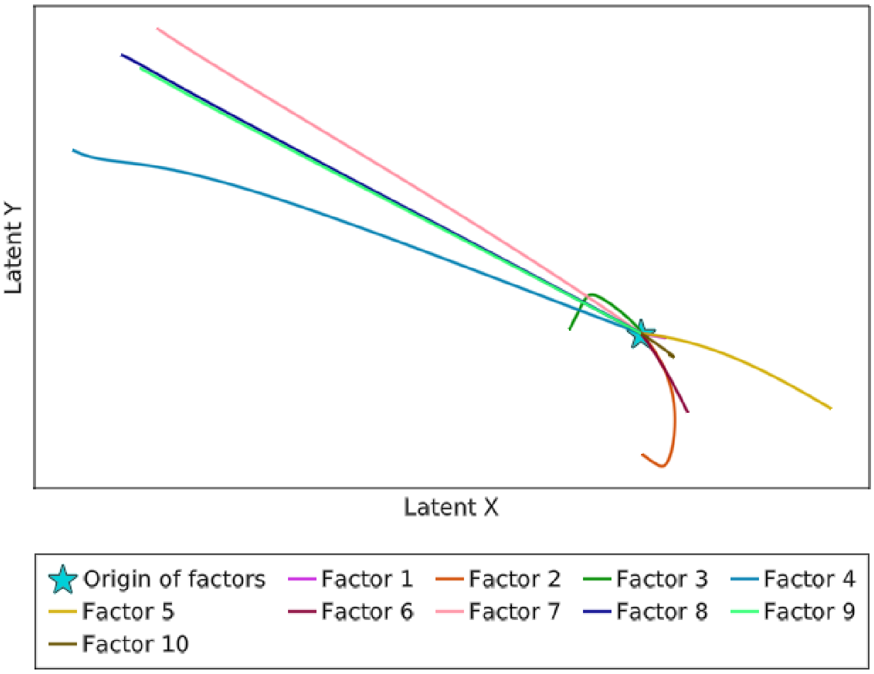

Figure B1 Origin and coordinates of the factors in the latent space. This figure is a reproduction under Creative Commons License 4.0 (CC BY) https://creativecommons.org/licenses/by/4.0/ from A. Yamashita, T. Nagata, S. Yagyu, T. Asahi, and T. Chikyow, “Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis,” Science and Technology of Advanced Materials: Methods, in press, https://doi.org/10.1080/27660400.2022.2029222, published by Taylor & Francis.

Acknowledgements

First of all, I would like to express my appreciation to my principal referee, Professor Toru Asahi for his kind support and guidance. I would also like to extend my deepest gratitude to Dr. Toyohiro Chikyow and Dr. Takahiro Nagata at National Institute for Materials Science for providing me invaluable advice and discussions throughout this research.

I would like to acknowledge the insightful discussions and feedbacks from my referees, Professor Takuro Katsufuji at Waseda University, Professor Kenji Miyatake at University of Yamanashi, Dr. Masahiko Ishida at NEC Corporation and Associate Professor Takuya Taniguchi at Waseda University.

Thanks should also go to Dr. Shinjiro Yagyu for his insightful discussion throughout my work at National Institute for Materials Science.

I also thank members of Nano Electronics Device Materials Group at National Institute for Materials Science and members of Asahi lab for kind support during my doctor course.

Finally I would like to thank my parents for their understanding and warm supports.

January, 2022

Akihiro Yamashita

List of research achievements for application of Doctor of Engineering, Waseda University

Full Name：山下　晶洸　　　　seal or signature

Date Submitted(yyyy/mm/dd): 2022/2/10

種類別 (By Type)	題名、　発表・発行掲載誌名、　　発表・発行年月、　　連名者（申請者含む） (theme, journal name, date & year of publication, name of authors inc. yourself)
論文	
○	1 Akihiro Yamashita, Takahiro Nagata, Shinjiro Yagyu, Toru Asahi, and Toyohiro Chikyow, "Direct feature extraction from two-dimensional X-ray diffraction images of semiconductor thin films for fabrication analysis" Science and Technology of Advanced Materials: Methods, in press
○	2 Akihiro Yamashita, Takahiro Nagata, Shinjiro Yagyu, Toru Asahi, and Toyohiro Chikyow, "Accelerating two-dimensional X-ray diffraction measurement and analysis with density-based clustering for thin films," Japanese Journal of Applied Physics, 60, SCCG04 (2020)
○	3 Takahiro Nagata, Takeshi Hoga, Akihiro Yamashita, Toru Asahi, Shinjiro Yagyu, and Toyohiro Chikyow "Valence Band Modification of a (GaxIn1–x)2O3 Solid Solution System Fabricated by Combinatorial Synthesis" ACS Combinatorial Science 2020, 22, 9, 433–439
講演	
	1 Akihiro Yamashita, Takahiro Nagata, Shinjiro Yagyu, Toru Asahi, and Toyohiro Chikyow, "Quasi-Continuous Representation of Crystal Structure of Thin Films with Two-Dimensional X-Ray Diffraction and Non-Negative Matrix Factorization" 2021 MRS Fall Meeting & Exhibit, Boston & Online, Nov. 29 - Dec. 2 & Dec. 6-8, 2021
	2 山下 晶洸、長田 貴弘、柳生 進二郎、朝日 透、知京 豊裕 「密度ベースクラスタリングを用いた2次元X線回折測定の高速化」 第68回応用物理学会春季学術講演会、オンライン、2021年3月
	3 Akihiro Yamashita, Takahiro Nagata, Shinjiro Yagyu, Toru Asahi, and Toyohiro Chikyow, "Automated Cluster Analysis of 2-Dimensional X-Ray Diffraction for Composition Spread Oxide Thin Film Fabricated by Combinatorial Synthesis, Aiming to Visual Information-Guided Material

List of research achievements for application of Doctor of Engineering, Waseda University

Full Name： 山下　晶洸 seal or signature

Date Submitted(yyyy/mm/dd): 2022/2/10

種類別 (By Type)	題名、　発表・発行掲載誌名、　　発表・発行年月、　　連名者（申請者含む） (theme, journal name, date & year of publication, name of authors inc. yourself)
	Discovery" 2020 Virtual MRS Fall Meeting & Exhbit, Online, Dec., 2020 4 Akihiro Yamashita, Takahiro Nagata, Shinjiro Yagyu, Toru Asahi, and Toyohiro Chikyow, "Accelerating 2-Dimensional X-Ray Diffraction Measurement and Analysis with Density-Based Clustering for Thin Films" 33rd International Microprocesses and Nonotechnology Conference (MNC2020), Online, 11, 2020 5 山下 晶洸、長田 貴弘、柳生 進二郎、朝日 透、知京 豊裕 「視覚情報による薄膜材料探索効率化に向けた、2次元X線回折手法の自動解析技術の開発」 第81回応用物理学会秋季学術講演会、オンライン、2021年9月 6 山下 晶洸、長田 貴弘、柳生 進二郎、朝日 透、知京 豊裕 「計測インフォマティクスに向けた2次元X線回折解析プログラムの開発」 第67回応用物理学会春季学術講演会、東京、2020年3月 7 宝賀剛、山下 晶洸、朝日 透、知京豊裕、長田貴弘 「(GaxIn1-x)2O3固溶体薄膜における結晶構造および電気的特性の検討」 第67回応用物理学会春季学術講演会、東京、2020年3月 8 Akihiro Yamashita, Takahiro Nagata, Shinjiro Yagyu, Toru Asahi, and Toyohiro Chikyow, "2D-X-Ray Diffraction Data Handling Flow for Time Efficient Measurement" 13th MANA International Symposium 2020, Tsukuba, 2, 2020