Skill-Agnostic analysis of reflection high-energy electron diffraction patterns for Si(111) surface superstructures using machine learning

Asako Yoshinari, Yuma Iwasaki, Masato Kotsugi, Shunsuke Sato, and Naoka Nagamura

ABSTRACT
Reflection high-energy electron diffraction (RHEED) data are important for the in-situ characterization of surface conditions during physical vapor deposition. Surface superstructures obtained by adsorbing exotic atoms onto a clean silicon surface, which exhibit various physical properties, were identified using RHEED. However, this information is too abundant for quantitative analysis; therefore, scientists rely on their expertise to interpret RHEED patterns to assess surface structures and evaluate film thickness, and a large amount of information remains unused. In this study, we adapted machine learning for a RHEED pattern dataset of a clean Si(111) surface during indium deposition in molecular-beam epitaxy growth to use the entire RHEED pattern image information and investigated appropriate machine learning analysis methods. First, we aimed to determine RHEED pattern similarities in the dataset. Then, five structural phases, 7 x 7 (clean surface), √3 x √3, √31 x √31, 4 x 1, and 4 x 1 (Room Temperature), were automatically detected by hierarchical clustering using Ward’s method. Next, we aimed to extract the information for each surface superstructure from the dataset. Using non-negative matrix factorization, we successfully estimated the optimal forming conditions for each surface superstructure separately more accurately than the conventional methods. Our proposed strategies could be widely applied to surface structural analysis.

Machine Learning

Skill-dependent analysis

| Substrate | RHEED images |
| --- | --- |
| Vapor Deposition | Judgment |
| * steps of structural transition | * positions of spots...etc. |
| knowledge and experience |

Skill-agnostic analysis

| Detection | Estimation |
| --- | --- |
| the structural surface transition | the optimal deposition time |
| Hierarchical Clustering | Non-negative Matrix Factorization |

CONTACT Naoka Nagamura

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

ARTICLE HISTORY
Received 21 March 2022
Revised 04 May 2022
Accepted 17 May 2022

KEYWORDS
RHEED; machine learning; surface superstructure
1. Introduction

Clean semiconductor surfaces with well-controlled atomic structures provide attractive stages for various low-dimensional systems such as atomic-layer materials [1–4], heteroepitaxial multi-layers [5–7], topological materials [8–10], and Rashba films [11–13]. These surfaces and thin films are fabricated by physical vapor deposition (PVD). Reflection high-energy electron diffraction (RHEED) is a widely used technique for structural analysis of surfaces obtained by PVD growth. RHEED enables researchers to observe surface structures in situ during film fabrication because the deposition rate can be evaluated by monitoring the dynamic changes in the diffraction patterns and intensity oscillations of the selected diffraction spots. Thus, RHEED has been used in the process optimization for sample preparation. However, the intensity of the diffraction spot is affected by the background, neighboring spots, incident electron beam intensity fluctuations, and sample temperature, among other factors. Moreover, expert knowledge is required to interpret diffraction patterns. RHEED images contain a lot of surface information, including surface atomic structural and film quality information, such as step density, growth mode, and strain [14–17]. Therefore, an appropriate data processing technique is required to extract the necessary information and perform a high-accuracy analysis.

Recently, informatics strategies have been utilized for big data analysis in materials [18–21] such as the promotion of novel material explorations [22–24], high-throughput measurements [25–27], and optimization of fabrication conditions [28–30]. Machine learning and data mining techniques are powerful tools for automatically analyzing image datasets with high accuracy and low cost, thus avoiding the subjectivity of experts. Several approaches have been presented for RHEED pattern analysis, such as convolutional neural networks (CNNs), principal component analysis (PCA), and k-means clustering [31–36].

In this study, we propose a skill-agnostic analysis method for RHEED images. We applied unsupervised machine learning, multi-dimensional scaling (MDS), hierarchical clustering, PCA, independent component analysis (ICA), and non-negative matrix factorization (NMF) to RHEED pattern images of a Si(111) surface during indium deposition in molecular-beam epitaxy (MBE) growth. By introducing machine learning, scientists can thoroughly analyze a large amount of information. In particular, unsupervised learning methods are useful in practice for categorizing datasets in which we do not know which element will appear, such as time-series RHEED pattern image datasets taken during parameter sweep. Aiming to perform a skill-agnostic and fast analysis, we applied unsupervised learning to one-dimensional (1D) RHEED image arrays and worked on a highly accurate process optimization for sample preparation, as compared to the conventional thickness monitoring method.

To determine the number of surface superstructural phases included in the dataset, we attempted MDS and hierarchical clustering (3.1). Using Ward’s method of hierarchical clustering, we automatically found that five structural phases were present. Subsequently, to extract information on each surface superstructure – namely 7 × 7 (clean surface), √3×√3, √31×√31, 4 × 1, and 4 × 1 (Room Temperature, RT)—PCA, ICA, and NMF were tried (3.2) [37–39]. Using NMF, the optimal forming conditions for each surface superstructure were individually estimated with high accuracy. Si surfaces are the most basic and widely studied semiconductor surfaces. Over the past few decades, various superstructures on Si(111) surfaces induced by several adsorbed atoms have been reported [1,2]. Each superstructure shows an amazingly different characteristic; therefore, it is important to identify them and optimize the sample preparation conditions.

The remainder of this paper is structured as follows. Section 2 presents the experimental procedures. Section 3 discusses the automatic detection of surface superstructural phase transitions and optimization of the formation of each surface superstructure along with the several machine learning methods. Finally, section 4 provides a brief summary and outlook.

2. Experimental procedures

2.1. Sample preparation

The substrate was an n-type (resistivity of 1–10 Ω-cm at 300 K) Si(111) wafer, the size was 3 mm × 20 mm. Si(111) is three-fold symmetric; however, to form single-domain 4 × 1-In chains, vicinal substrate with a miscut of 1.8° along the [112] direction was used. The oxide layer was removed by direct current (DC) heating to 1200 °C for a few seconds in an ultra-high vacuum (UHV) chamber (1 × 10⁻⁸ Pa). Then the substrate was annealed at 1050 °C to produce a Si(111)-7 × 7 clean surface terrace with one atom high steps. Indium was deposited onto this 7 × 7 surface from an alumina-coated tungsten wire basket at 450 °C until approximately a 1.0 monolayer (ML) indium chains were formed along the steps. Here, 1.0 ML corresponds to 7.84 × 10¹⁴ atoms/cm² for Si(111).
2.2. RHEED image dataset

We started the DC sample heating and RHEED pattern recording and then opened the shutter of the deposition source to start the atomic layer growth after the sample temperature stabilized. The acceleration voltage for RHEED was 15 keV. The angle of incidence and beam current of RHEED were 3.0° and 36 µA, respectively. The shutter was closed to finish indium deposition after 392 s and DC sample heating was stopped after 421 s from the start of deposition. Finally, the RHEED pattern recording was stopped after the sample had cooled to room temperature. A total of 375 RHEED pattern images ([112] incident) were continuously taken for 480 s using a charge-coupled device (CCD) camera to monitor the structural changes caused by the surface phase transition due to the change in the indium deposition amount. The average interval between shots was 1.28 s, and the size of RHEED pattern images was 480 × 640 (h×w) pixels.

We used 208,251 pixels within a fluorescent screen, as shown in Figure 1 (red circle), from the 307,200 (480 × 640) pixels in a rectangular image from the CCD camera. The luminance of each pixel was obtained and the output was expressed using a 1D array.

2.3. Machine learning analysis

Machine learning analysis was attempted for 374 images after excluding one image with a shifted view; therefore, 374 1D luminance arrays were used as the dataset. We adopted MDS, hierarchical clustering, PCA, ICA, and NMF for RHEED image analysis. Machine learning was performed using the Scikit-learn packages for MDS, PCA, ICA, and NMF, and the SciPy package for hierarchical clustering. Unless specified, the default parameters were used. Data analysis was performed using the following machine learning software and libraries: Python (Version: 3.7.6), Scikit-learn (Version: 0.22.2), and SciPy (Version: 1.4.1). The computer used for the calculations was equipped with an Intel(R) Core(TM) i5 CPU with four cores at 1.60 GHz and 16 GB memory.

3. Results and discussion

3.1. Surface superstructure transition

The Si(111) clean surface shows a structural phase transition according to the amount of indium adsorption [1]; therefore, the RHEED pattern image dataset taken during the deposition process includes information on several surface superstructural phases. To detect the number of surface superstructural phases, we applied MDS and clustering, which are general methods for confirming sample similarity in datasets. First, we performed MDS, which can express the overall relationships among all samples in two-dimensional mapping, to visualize whether clusters corresponding to each surface structure appear in the mapping. Figure 2 shows the MDS results using a six-point distance scale: ‘Euclidean distance,’ ‘Manhattan distance,’ ‘Chebyshev distance,’ ‘Squared Euclidean distance,’ ‘Cosine distance,’ and ‘Hamming distance.’ In Figure 2, the color of the plot changes from blue to light green over time. From Figure 2, the number of phases and phase transitions during indium deposition do not appear clearly. However, in Figure 2(a–e), the light green markers form a cluster, and the other markers are arranged in a continuous manner. The latter distribution occurs because the spot size of the electron gun in the RHEED measurement system was large (100–200 µm) compared to the crystal nuclei and domains, and the structural phase transitions in the RHEED patterns under the deposition process seems to occur continuously. The former distribution (i.e. the light green cluster) includes the RHEED images taken after the DC sample heating was stopped. At that moment, the intensity of the RHEED spots increased discontinuously with sample cooling. On the other hand, in Figure 2(f), with Hamming distance, the markers show a spiral-like shape, and the time-dependent change tendency in the mapping plot cannot be interpreted. This means that the Hamming distance is not suitable for our dataset.

Next, we performed a clustering analysis, specifically, hierarchical clustering. Non-hierarchical clustering, such as the k-means clustering, has often been used as a typical clustering analysis method in previous studies [34–36]; however, the number of clusters must be determined prior to the analysis by
knowledgeable and experienced scientists. Furthermore, the k-means clustering is vulnerable to unbalanced cluster sizes. Therefore, we applied hierarchical clustering to the RHEED image dataset. In hierarchical clustering analysis, the relationship between samples is shown in a hierarchical structure called dendrogram, which helps estimate the number of clusters visually. To perform hierarchical clustering, it is necessary to select the routine for combining clusters and the distance scale among samples.

**Figure 2.** MDS results for (a) Euclidean distance, (b) Manhattan distance, (c) Chebyshev distance, (d) Squared Euclidean distance, (e) Cosine distance, and (f) Hamming distance.

**Figure 3.** Hierarchical clustering results with Euclidean distance using the Single linkage method, Group average method, and Ward’s method. (a) Dendrograms for each method; (b) Temporal cluster change. The dynamical change in diffraction patterns is detected by all linkage methods, but the number of clusters are different for each.
Considering the MDS principle, in which similar samples are plotted closer and dissimilar samples are plotted farther in the selected distance scale, it is expected that the distance scales that capture the features of RHEED pattern changes in MDS mapping are appropriate for performing the hierarchical clustering. From Figure 2(a–f), the behavior using the Euclidean and Squared Euclidean distances show a particularly reasonable behavior. We adopted Euclidean distance, which is highly versatile for cluster-combining methods, and used seven routines for combining clusters, namely the single linkage, complete linkage, group average, weighted average, centroid, and median methods, as well as Ward’s method. Figure 3(a) shows the dendrogram results and Figure 3(b) shows the temporal cluster change at the farthest distance in the hierarchy for the single linkage method, group average method, and Ward’s method. The results using the complete linkage, weighted average, centroid, and median methods are shown in Figure S1 in the Supplementary Material. The number of clusters for each routine is summarized in Table 1. Although we succeeded in detecting the dynamical change in diffraction patterns using all linkage methods, as shown in Figure 3 and S1, the number of

| Linkage method      | Single linkage | Complete linkage | Group average | Weighted average | Centroid | Median | Ward’s |
|---------------------|----------------|------------------|---------------|------------------|----------|--------|--------|
| Number of clusters  | 2              | 3                | 3             | 3                | 3        | 3      | 5      |

**Table 1. Number of clusters comparison via hierarchical clustering using seven linkage methods.**

Figure 4. Average images of clusters obtained using Ward’s method. Each cluster corresponds to surface superstructures 7×7, √3×√3, √31×√31, 4×1, and 4×1 (RT) after the DC sample heating was stopped.

Figure 5. Average images of clusters obtained using Ward’s method. Each cluster corresponds to the phase diagram of indium deposited on silicon surface. At 450 °C, where indium deposition was performed in this study, five surface superstructures (7×7, √3×√3, √31×√31, 4×1, and 4×1 (RT) after stopping the DC sample heating) are expected to appear. This figure is reproduced from Ref [1].
clusters was different, as shown in Table 1. However, each cluster does not necessarily correspond to a specific surface superstructure. Ward’s method shows the largest number of clusters, five is the optimum, as shown in Table 1. The average images of each cluster are shown in Figure 4. In Figure 4, Clusters 2–5 represent four completely different patterns, while Cluster 1 represents the pattern in which the intensity discontinuously increased by stopping the DC sample heating. Compared with the phase diagram (Figure 5 [1]), clusters correspond to the expected five indium surface superstructures on Si(111): 7 × 7 (clean surface), √3×√3, √31×√31, 4 × 1, and 4 × 1 (RT) after stopping the DC sample heating. Although the 4 × 1 pattern represented by Cluster3 and the 4 × 1 (RT) pattern represented by Cluster1 are essentially the same, they differ in the intensity and number of visible spots with change in substrate temperature. Due to these difference of diffraction patterns depending on temperature, the 4 × 1 pattern and the 4 × 1(RT) pattern were detected as different clusters.

However, the other routines did not work well; for example, four structures were recognized during heating as a single cluster when a single linkage was used. Ward’s method combines clusters together in such a manner that the increase in the square of the distance from the center of gravity due to the combining clusters is minimized. Although computationally expensive, Ward’s method is highly sensitive, and therefore, in our case, it is considered to be the most successful in categorization. From these results, Ward’s method was determined as the most appropriate for hierarchical clustering for RHEED pattern change detection. As shown in Figure 2, the dynamic change in the diffraction patterns due to the structural phase transition driven by the amount of adatom deposition occurs continuously. Therefore, cluster boundary detection is inherently difficult. Although Ward’s method is based on hard clustering, our results suggest that Ward’s method is a useful means to easily understand surface superstructural phase transitions.

### 3.2. Optimal deposition time estimation

In the previous section, automatic RHEED pattern detection was achieved via machine learning without prior knowledge or experience. Figure 6 presents the information on the phase transition obtained from the Ward’s method in section 3.1, together with the experimental procedure. Although the hierarchical clustering is hard clustering, the actual phase transitions occurred in a continuous manner; therefore, a conceptual diagram of the phase transitions is shown in Figure 6. Next, we searched the image when the pattern of each structure most clearly appeared, that is, the condition in which each structure occupied the largest area of the surface. This enabled us to determine the optimal deposition time for fabricating each surface superstructure separately. Searching for the best conditions for synthesizing target materials using machine learning is one of the goals of materials informatics.

Scientists have generally identified surface superstructures based on their expertise. In addition, they have optimized the deposition time to fabricate

![Figure 6](image-url)  
*Figure 6.* Experimental procedure and phase transitions described in section 3.1. The phase transition is shown in the conceptual diagram.
a certain surface superstructure by observing time-dependent intensity profile of RHEED spots from the aimed surface superstructure relying on their technique. This method is also used for thickness estimation using the intensity oscillation of RHEED spots. Typically, the intensity profile is detected at only one arbitrary RHEED spot. To obtain the intensity profile, a rectangular area around one arbitrary RHEED spot is defined; then, the luminance sum in that range is calculated. However, the accuracy of this method is poor because spots derived from unintended superstructures could exist in the selected rectangular area, and the effect of the background intensity, such as that from Kikuchi patterns, is inevitable. For example, in Figure 7, we attempted to estimate the optimal deposition time required to obtain a Si(111)4×1-In structure using the above mentioned conventional method. The RHEED spots in the red and blue rectangular areas are both derived from the 4×1 structure (Figure 7(a)). As shown in Figure 7(b), in the case of the red rectangular area, the intensity peak clearly appeared at 290 s, which can be determined as the optimal deposition time. However, in the case of the blue rectangular area, we cannot determine the optimal deposition time. This means that the position and size of the rectangular area significantly affect the results. Moreover, the results of this conventional method depend on the analyst’s expertise. Accordingly, by using machine learning, we expect a more accurate estimation of the best condition using information from the entire RHEED pattern image dataset.

Therefore, we aimed to extract the components that reflect the information of each surface superstructure via dimensional reduction. First, we applied PCA, which is a typical dimensional reduction method and has been used in previously [34–36]. Figure 8 shows the cumulative contribution ratio of the PCA results. It was found that the number of principal components could be reduced to four when more than 90% of the information in the dataset was retained. Figure 9 shows the basis pattern images of the four components and their score change over time. A higher component score indicates that the RHEED pattern contains the principal component more strongly at that time. In Figure 9(a), the basis patterns differ from those in Figure 4, which correspond to the diffraction patterns of specific surface superstructures obtained by hierarchical clustering using Ward’s method. The behavior of the component score change over time in Figure 9(b) is too complicated to interpret easily. Therefore, PCA is considered unsuitable for estimating the optimal deposition time for fabricating a specific surface superstructure. In PCA, the components are orthogonal to each other, but the RHEED patterns do not have orthogonal relationships, so PCA does not work well. Therefore, it is necessary to use a dimensional reduction method that extracts information under non-orthogonal constraints.

We also applied ICA, which is a dimensional reduction method under the constraint that the bases are independent. We assumed the number of independent components to be five because the number of RHEED patterns detected by hierarchical clustering was five. Figure 10 shows the images of the independent components and their score change over time. As with PCA, a higher component score
indicates that the RHEED pattern contains the independent component more strongly at that time. Similar to PCA, the basis pattern images of each component in Figure 10(a) cannot be assigned to the typical diffraction patterns of specific surface superstructures, and the score change over time in Figure 10(b) is also complicated.

The reason for the incompatibility of PCA and ICA for dimensional reduction is that positive and negative values are included in the components; therefore, diffraction patterns from different structures appear depending on the values (positive or negative). To estimate the optimal vapor deposition time for fabricating a specific surface superstructure, it is desirable to extract non-negative components. Therefore, we applied NMF, which is a dimensional reduction method with the constraint that only non-negative matrices can be decomposed. We performed NMF with number of basis $k = 5–10$. Figure 11 shows the difference between the images reconstructed using the basis patterns and the original images, that is, the reconstruction error. From the basis patterns with number of basis $k = 5–10$ shown in Figures S2–S6 and Figure 12(a), it is evident that when the number of basis is less than 7, each structure cannot be separated, whereas when the basis number is greater than 7, the structures are too divided to be accurately interpreted. Therefore, basis number 7 appears to be valid. It should be noted that the number of basis patterns in NMF and clusters obtained via hierarchical clustering using Ward's method are different. The appropriate number of components for performing NMF should be determined by confirming each component, in addition to the reconstruction error.

Figure 12 shows the NMF results with number of basis $k = 7$. Figure 12(a) shows the seven basis patterns obtained via dimensional reduction using NMF. Compared to the hierarchical clustering results using Ward’s method, each pattern in NMF seems to correspond to the background (Basis Pattern 1), $7 \times 7$ structure divided into two basis(Basis Pattern 2,3), $\sqrt{3} \times \sqrt{3}$ structure(Basis Pattern 4), $\sqrt{31} \times \sqrt{31}$ structure(Basis Pattern 5), $4 \times 1$ structure(Basis Pattern 6), and $4 \times 1$ (RT) structure after stopping the DC sample heating.

Figure 9. PCA results: (a) Patterns of the principal components; (b) Component score change over time. A larger component score means that the RHEED image contains the principal component more strongly at that time.

Figure 8. Cumulative contribution ratio of PCA: 90% of the information is retained using four principal components after dimensional reduction.
Basis Pattern 7. Of the two $7 \times 7$ patterns (Basis Pattern 2, 3), Basis Pattern 3 is suggested to be $\delta 7 \times 7$ [40]. This means that the RHEED pattern images are decomposed into the background and five structures using NMF. As in section 3.1, the $4 \times 1$ structure (Basis Pattern 6) and $4 \times 1$ (RT) after the DC sample heating was stopped (Basis Pattern 7) are also separated in NMF. In Basis Pattern 7, a rapid coefficient change over time occurred when the heating of the sample was stopped. The intensity of the RHEED pattern decreases with increasing substrate temperature due to the effect of lattice thermal vibrations. In other words, the coefficient change in Basis Pattern 7 may reflect the effect of lattice vibration as the substrate temperature changes.

Let us focus on the time-dependent change in the coefficients of each basis pattern (Figure 12(b)). Similar to PCA and ICA, the higher the coefficient value, the more strongly the RHEED pattern contained the basis pattern in the NMF analysis. Even though the $7 \times 7$ structure is divided into two basis patterns, Basis Patterns 2 and 3, Basis Pattern 3 oscillates when Basis Pattern 2 decreases fast, resulting in an overall decrease.

Figure 10. ICA result: (a) Patterns of the independent components; (b) Component score change over time. A larger component score means that the RHEED image contains the independent component more strongly at the time.
This is consistent with the expected surface superstructural transition, in which the Si(111)-7×7 clean surface changes into the next surface superstructure due to an increase in indium adsorption over time. Basis Pattern 4 in Figure 12 corresponds to the $\sqrt{3}\times\sqrt{3}$ structure basis pattern and coefficients. Its maximum value was obtained at 129 s, so the optimal deposition time of indium for the $\sqrt{3}\times\sqrt{3}$ structure was found to be 129 s. Accordingly, the optimal deposition times of indium for the $\sqrt{31}\times\sqrt{31}$ and $4\times1$ structures are found to be 184 and 292 s via Basis Patterns 5 and 6, respectively, in Figure 12. Lastly, Basis Pattern 7 in Figure 12 shows the $4\times1$ (RT) structure after stopping the DC sample heating and the coefficients start to increase sharply at 421 s, when the DC sample heating was stopped. This behavior is consistent with the experimental procedure. From these NMF results, it is considered that NMF is a suitable method for decomposing the RHEED pattern into the corresponding main factors and estimating the optimal deposition time for fabricating each surface superstructure.

Table 2 shows an estimated deposition time comparison between the NMF and conventional methods, the latter is shown in Figure 7. The NMF method has the potential to estimate the optimal deposition time automatically without prior knowledge regarding diffraction patterns and is more accurate than the conventional method. Figure 13 shows the optimal deposition time estimated by NMF method for each structural phase. The structures $\sqrt{3}\times\sqrt{3}$ and $4\times1$ are composed of 1/3 ML and 1 ML of indium surface density, respectively [37]. Nevertheless, the estimated time for $4\times1$ was not 3 times greater than that for $\sqrt{3}\times\sqrt{3}$. This may be related not only to the stability of deposition source, but also to the desorption and the diffusion of indium atoms. Film thickness gauges do not necessarily reflect the accurate adsorption amount of deposited atoms. Therefore, in-situ structural monitoring by RHEED and machine learning analysis of diffraction patterns are a powerful combination to estimate the optimal deposition conditions with high accuracy.

4. Conclusions

Herein, we propose an efficient method to analyze the entire RHEED pattern image dataset using machine learning for indium deposition on Si(111) surface superstructures. MDS, hierarchical clustering, PCA, ICA, and NMF were applied.

Hierarchical clustering analysis by Ward’s method easily clarified the five steps of the surface superstructure transition associated with indium deposition: $7\times7$ structure, $\sqrt{3}\times\sqrt{3}$ structure, $\sqrt{31}\times\sqrt{31}$ structure, $4\times1$ structure, and $4\times1$ (RT) structure after stopping the DC sample heating.
Figure 12. NMF results: (a) Basis patterns, each pattern shows the background, two 7×7 structure patterns, √3×√3 structure pattern, √31×√31 structure pattern, 4×1 structure pattern, and 4×1 structure pattern (after stopping the DC sample heating); (b) Coefficient change over time for each basis pattern. A high coefficient value means that the RHEED pattern image contains the basis pattern strongly at the time. ▼ indicates the optimal deposition time for fabrication of the √3×√3 (Basis Pattern 4), √31×√31 (Basis Pattern 5), and 4×1 (Basis Pattern 6) surface superstructures.

Figure 13. Optimal deposition time in NMF results with conceptual figure of phase transition. The optimal deposition time estimated by NMF are plotted as ▼.
Compared to non-hierarchical clustering, represented by the k-means clustering used in previous studies, the proposed method has an advantage in that the number of phases does not have to be determined in advance. Hence, Ward’s method can be applied to the discovery of diffraction patterns of new structural phases that were not expected in conventional observations.

In contrast, the NMF analysis revealed the behavior of each surface superstructure. NMF analysis is more interpretable than methods such as PCA or ICA because the latter treat both positive and negative values and extract components of orthogonal or independent relationships. The proposed method enables an automatic estimation of the optimal deposition time for fabricating $\sqrt{3}\times\sqrt{3}$, $\sqrt{31}\times\sqrt{31}$, and $4\times1$ structures, each without prior knowledge or experience. In addition, it has the potential to obtain the optimal deposition time more accurately than conventional methods.

We used post-acquisition image dataset in this report. However, further work is needed to improve the method for in-situ analysis, taking advantage of the in-situ observation capability of RHEED. In addition, the 1D arrangement arrays of lumiance of the images are easily affected by external brightness and camera movement. Therefore, an improvement in feature engineering is required.

In future work, we intend to adopt these machine learning RHEED analysis methods for other systems, such as silicon surface superstructures with various adsorption materials and other phase transitions such as those driven by temperature. We believe that our method can be widely applied to surface structural analysis, such as metal oxide and III–V compound semiconductor film growth.

Acknowledgement

The RHEED spectral dataset was obtained using Shuji Hasegawa’s laboratory equipment at the University of Tokyo. We thank Rei Hobara for his cooperation in acquiring the RHEED images and for informative discussion. This study was supported by JSPS KAKENHI Grant No. 19H02561; JST-CREST Grant No. JPMJCR2101; and JST-PRESTO Grant Nos. JPMJPR20T7 and JPMJPR17NB.

Disclosure statement

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

ORCID

Asako Yoshinari http://orcid.org/0000-0002-8049-5028
Yuma Iwasaki http://orcid.org/0000-0002-7117-277X
Masato Kotsugi http://orcid.org/0000-0002-4841-1808
Shunsuke Sato http://orcid.org/0000-0001-5971-9368
Naoka Nagamura http://orcid.org/0000-0002-7697-8983

References

[1] V.G.Lifshits AAS and AVZ. Surface phases on silicon. Chichester (UK): J.WILEY; 1994.
[2] Hasegawa S, Tong X, Takeda S, et al. Structures and electronic transport on silicon surfaces. Prog Surf Sci. 1999;60:89–257.
[3] Kara A, Enriquez H, Seitsonen AP, et al. A review on silicene - New candidate for electronics. Surf Sci Rep. [Internet]. 2012;67:1–18. doi:10.1016/j.surfrep.2011.10.001
[4] Barth J, Costantini G, Kern K. Engineering atomic and molecular nanostructures at surfaces. Nature. 2005;671–679. doi:10.1038/nature04166
[5] Yu Z, Liang Y, Overgaard C, et al Advances in heteroepitaxy of oxides on silicon. Thin Solid Films. 2004;462–463:51–56.
[6] McKee RA, Walker FJ, Chisholm MF. Crystalline oxides on silicon: the first five monolayers. Phys Rev Lett. 1998;81:3014–3017.
[7] Nakada Y, Aksenov I, Okumura H. GaN heteroepitaxial growth on silicon nitride buffer layers formed on Si (111) surfaces by plasma-assisted molecular beam epitaxy. Appl Phys Lett. 1998;73:827–829.
[8] Li YY, Wang G, Zhu XG, et al Intrinsic topological insulator Bi 2 Te 3 thin films on Si and their thickness limit. Adv Mater. 2010;22:4002–4007.
[9] Sakamoto Y, Hirahara T, Miyazaki H, et al Spectroscopic evidence of a topological quantum phase transition in ultrathin Bi2Se3 films. Phys Rev B. 2010;81:2–5.
[10] Hilmi I, Lotnyk A, Gerlach JW, et al. Influence of substrate dimensionality on the growth mode of epitaxial 3D-bonded GeTe thin films: from 3D to 2D growth. Mater Des. [Internet]. 2019;168:107657. doi:10.1016/j.matdes.2019.107657.
[11] Gierz I, Suzuki T, Frantzeskakis E, et al. Silicon surface with giant spin splitting. Phys Rev Lett. 2009;103:1–4.
[12] Yaji K, Ohtsubo Y, Hatta S, et al. Large Rashba spin splitting of a metallic surface-state band on a semiconductor surface. Nat Commun. 2010;1:1–5.
[13] Sakamoto K, Kim TH, Kuzumaki T, et al. Valley spin polarization by using the extrardinary Rashba effect on silicon. Nat Commun. 2013;4:1–6.
[14] Braun W. Applied rhed. Springer; 1999.
[15] Shuji H. Reflection high-energy electron diffraction. In: Kaufmann E, editor. Charact Mater. [Internet]. John Wiley & Sons; 2012. p. 1925.
[16] Kawamura T. Relation between surface step density and RHEED intensity. Surf Sci. 1993;298:331–335.
[17] Lewis BF, Lee TC, Grunthaner FJ, et al. RHEED oscillation studies of mbe growth kinetics and lattice mismatch strain-induced effects during InGaAs growth on GaAs(100). J Vac Sci Technol B Microelectron Nanom Struct. 1984;2:419–424.
[18] Butler KT, Davies DW, Cartwright H, et al. Machine learning for molecular and materials science. Nature. [Internet]. 2018;559:547–555. doi:10.1038/s41586-018-0337-2.
[19] Agrawal A, Choudhary A. Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. APL Mater. [Internet]. 2016;4:053208.
[20] Schmidt J, Marques MRG, Botti S, et al. Recent advances and applications of machine learning in solid-state materials science. NPJ Comput Mater. [Internet]. 2019;5. doi:10.1038/s41524-019-0221-0

[21] Ramprasad R, Batra R, Pilania G, et al. Machine learning in materials informatics: recent applications and prospects. NPJ Comput Mater. [Internet] 2017;3. doi:10.1038/s41524-017-0056-5

[22] Oganov AR, Glass CW. Crystal structure prediction using ab initio evolutionary techniques: principles and applications. J Chem Phys. 2006;124:244704.

[23] Ward I, Agrawal A, Choudhary A, et al. A general-purpose machine learning framework for predicting properties of inorganic materials. NPJ Comput Mater. 2016;2:1–7.

[24] Iwasaki Y, Sawada R, Saitoh E, et al. Machine learning autonomous identification of magnetic alloys beyond the Slater-Pauling limit. Commun Mater. [Internet]. 2021;2:1–7. doi:10.1038/s43246-021-00135-0

[25] Ueno T, Ishibashi H, Hino H, et al. Automated stopping criterion for spectral measurements with active learning. NPJ Comput Mater. [Internet]. 2021;7:1–9. doi:10.1038/s41524-021-00606-5

[26] Iwasawa H, Ueno T, Masui T, et al. Unsupervised clustering for identifying spatial inhomogeneity on local electronic structures. NPJ Quantum Mater. 2022;7. doi:10.1038/s41535-021-00407-5

[27] Matsumura T, Nagamura N, Akaho S, et al. Spectrum adapted the expectation-maximization algorithm for high-throughput peak shift analysis. Sci Technol Adv Mater. [Internet]. 2019;20:733–745. doi:10.1080/14686996.2019.1620123

[28] Nikolaev P, Hooper D, Webber F, et al. Autonomy in materials research: a case study in carbon nanotube growth. NPJ Comput Mater. 2016;2. doi:10.1038/npjcompumats.2016.31

[29] Burger B, Maffettone PM, Gusev V, et al. A mobile robotic chemist. Nature. [Internet]. 2020;583:237–241. doi:10.1038/s41586-020-2442-2

[30] Shimizu R, Kobayashi S, Watanabe Y, et al. Autonomous materials synthesis by machine learning and robotics. APL Mater. [Internet]. 2020;8:2–8. doi:10.1063/5.0020370

[31] Bensaoula A, Malki HA, Kwari AM. The use of multi-layer neural networks in material synthesis. IEEE Trans Semicond Manuf. 1998;11:421–431.

[32] Lee KK, Brown T, Dagnall G, et al Using neural networks to construct models of the molecular beam epitaxy process. IEEE Trans Semicond Manuf. 2000;13:34–45.

[33] Kwoen J, Arakawa Y. Classification of reflection high-energy electron diffraction pattern using machine learning. Cryst Growth Des. 2020;20:5289–5293.

[34] Vasudevan RK, Tselev A, Baddorf AP, et al Big-Data reflection high energy electron diffraction analysis for understanding epitaxial film growth processes. ACS Nano. 2014;8:10989–10908.

[35] Provence SR, Thapa S, Paudel R, et al. Machine learning analysis of perovskite oxides grown by molecular beam epitaxy. Phys Rev Mater. [Internet]. 2020;4:83807. doi:10.1103/PhysRevMaterials.4.083807.

[36] Gliebe K, Sehirlioglu A. Distinct thin film growth characteristics determined through comparative dimension reduction techniques. J Appl Phys. 2021;130:125301.

[37] Vlachos D, Kamaratos M, Foulias SD, et al. Indium growth on reconstructed Si(111)√3 × √3 and 4 × 1 in surfaces. J Phys Chem C. 2010;114:17963–17972.

[38] Saranin AA, Zotov VA, Tovpik AN, et al. Composition and atomic structure of the Si(111)√31 × √31-In surface. Surf Sci. 2000;450:34–43.

[39] Chukurov EN, Alekseev AA, Kotlyar VG, et al. First-Principles study of Si(111)√31 × √31 -In reconstruction. Surf Sci. [Internet]. 2012;606:1914–1917. doi:10.1016/j.susc.2012.07.038

[40] Horio Y, Ichimiya A, Kohmoto S, et al. Kinematical analysis of RHEED patterns from a hydrogen adsorbed Si(111) δ7 × 7 surface. Surf Sci. 1991;257:167–174. doi:10.1016/0039-6028(91)90789-U