Taylor & Francis
Taylor & Francis Group

Science and Technology of Advanced Materials: Methods

ISSN: 2766-0400 (Online) Journal homepage: www.tandfonline.com/journals/tstm20

Bayesian inference for peak feature extraction and prediction of material property in X-ray diffraction data

Ryo Murakami, Taisuke T. Sasaki, Hideki Yoshikawa, Yoshitaka Matsushita, Keitaro Sodeyama, Tadakatsu Ohkubo, Hiroshi Shinotsuka & Kenji Nagata

To cite this article: Ryo Murakami, Taisuke T. Sasaki, Hideki Yoshikawa, Yoshitaka Matsushita, Keitaro Sodeyama, Tadakatsu Ohkubo, Hiroshi Shinotsuka & Kenji Nagata (2024) Bayesian inference for peak feature extraction and prediction of material property in X-ray diffraction data, Science and Technology of Advanced Materials: Methods, 4:1, 2384352, DOI: 10.1080/27660400.2024.2384352

To link to this article: https://doi.org/10.1080/27660400.2024.2384352

Published online: 15 Aug 2024.

SCIENCE AND TECHNOLOGY OF ADVANCED MATERIALS: METHODS
2024, VOL. 4, NO. 1, 2384352
https://doi.org/10.1080/27660400.2024.2384352

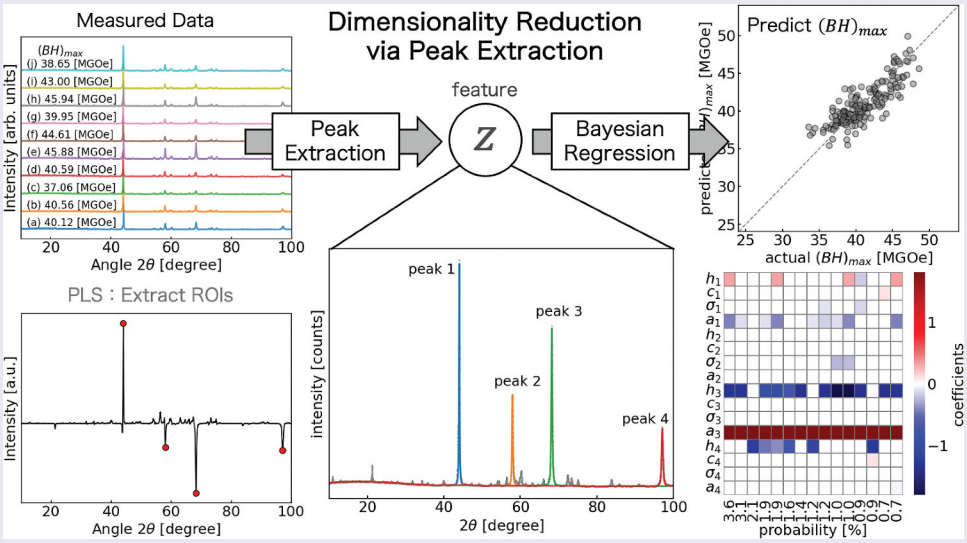
Bayesian inference for peak feature extraction and prediction of material property in X-ray diffraction data

Ryo Murakami[a], Taisuke T. Sasaki[b], Hideki Yoshikawa[a,c], Yoshitaka Matsushita[a], Keitaro Sodeyama[c], Tadakatsu Ohkubo[b], Hiroshi Shinotsuka[a] and Kenji Nagata[c]

[a]Research Network and Facility Services Division, National Institute for Materials Science, Tsukuba, Japan; [b]Research Center for Magnetic and Spintronic Materials, National Institute for Materials Science, Tsukuba, Japan; [c]Center for Basic Research on Materials, National Institute for Materials Science, Tsukuba, Japan

ABSTRACT

To advance the development of materials through data-driven scientific methods, appropriate methods for building machine learning (ML)-ready feature tables from measured and computed data must be established. In materials development, X-ray diffraction (XRD) is an effective technique for analysing crystal structures and other microstructural features that have information that can explain material properties. Therefore, the fully automated extraction of peak features from XRD data without the bias of an analyst is a significant challenge. This study aimed to establish an efficient and robust approach for constructing peak feature tables that follow ML standards (ML-ready) from XRD data. We challenge peak feature extraction in the situation where only the peak function profile is known a priori, without knowledge of the measurement material or crystal structure factor. We utilized Bayesian estimation to extract peak features from XRD data and subsequently performed Bayesian regression analysis with feature selection to predict the material property. The proposed method focused only on the tops of peaks within localized regions of interest (ROIs) and extracted peak features quickly and accurately. This process facilitated the rapid extracting of major peak features from the XRD data and the construction of an ML-ready feature table. We then applied Bayesian linear regression to the maximum energy product $(BH)_{max}$, using the extracted peak features as the explanatory variable. The outcomes yielded reasonable and robust regression results. Thus, the findings of this study indicated that *004* peak height and area were important features for predicting $(BH)_{max}$.



ARTICLE HISTORY

Received 13 February 2024
Revised 16 July 2024
Accepted 18 July 2024

KEYWORDS

Materials informatics; spectral decomposition; Bayesian estimation; feature selection; al-ready

IMPACT STATEMENT

Our method performs peak feature extraction from XRD data and consistently predicts physical property from peak features within the framework of Bayesian estimation, without requiring any crystal phase information.

CONTACT Kenji Nagata NAGATA.Kenji@nims.go.jp Center for Basic Research on Materials, National Institute for Materials Science, Tsukuba 305-0044, Japan

1. Introduction

In an effort to accelerate material development via data-driven science, the construction of a machine learning (ML)-ready feature table from measurement or computation data is essential. This necessitates appropriate analysis tools that extract the interpretable low-dimensional features from high-dimensional data such as spectra and images. In materials development, X-ray diffraction (XRD) is an effective technique for analysing crystal structures and other microstructural features that have information that can explain material properties. The crystal structure of a material can be understood by analyzing the diffraction peaks in the XRD data. However, XRD data often has numerous peaks and observation points, making it difficult to fully automatically extract the peak features from XRD data without the bias of an analyst.

The peak search method employing differential operation is often used in XRD analysis [1–3]. Peak search can find the primary peaks; however, it is unable to find sub-peaks, i.e. the peak of a hem. Consequently, the analyst has to manually assign sub-peaks to ensure that an ML-ready peak feature table is not constructed rather rapidly. The measurement data can be used as a feature table; however, it requires strong constraints, such as the need to used identical measurement equipment and measurement conditions. To share the feature table across a variety of material development processes, we believe that the features of the peak that are robust with respect to the measurement equipment and conditions are required. Thus, there is a need for a fully automated peak separation method in case of XRD data.

In recent years, XRD data analysis methods have been proposed for the automatic extraction of material information [4–7]. These methods are remarkably effective for XRD analysis of well-known materials and experimental systems. However, they require an organized database or a large dataset. In addition, they strongly reduce the XRD data to the lattice parameters. This reduction is stronger than the peak extraction. As a result, the peak parameters containing rich material information cannot be stored in the material database or utilized appropriately. Studies are actively attempting to extract information from large datasets in XRD data using multivariate analysis based on large-scale or non-parametric models [8–12]. It is desirable to have a large dataset with the same measurement conditions: slit size, scan step, X-ray resolution, etc. However, such a methodology may not be suitable for efficient sharing of data and features from multiple locations/institutions.

This study aimed to develop a method for automatically extracting high-intensity peak features. We challenge peak feature extraction in the situation where only the peak function profile is known a priori, without knowledge of the measurement material or crystal structure factor. This study demonstrated the effectiveness of the proposed framework in predicting the maximum energy product, $(BH)_{max}$, an important magnetic property in permanent magnets such as neodymium. Our framework first extracted the region of interest (ROI) for predicting the material property using partial least squares (PLS) regression [13,14]. The PLS method can perform multivariate analysis for predicting the objective variables and hence can visualize the important regions of the spectrum. PLS regression generates latent variables based basis components by linear projection of the input variables $\boldsymbol{y}_m$ and seeks to minimize the error between the objective variable t_m and the output of a linear model with the latent variable as input; PLS regression can provide us with basis components and its coefficients in a linear model. Then the proposed method extracted the peak features in the ROIs. The proposed peak separation method extracted peak features using information from the peak tops only. This facilitated the removal of the bias in peaks with a small base. The framework rapidly provided the peak feature from the XRD data by focusing on the peak top in an ROI. Furthermore, this method was used to calculate the accuracy of peak feature extraction using Bayesian estimation [15–18].

In this study, $(BH)_{max}$ was regressed using the peak features based on the Bayesian linear regression model with the feature selection [19]. This regression model can estimate the probability of a feature being used in the regression by pseudo-exploring all combinations using the replica exchange Monte Carlo method [15,20, 21–24]. Consequently, the proposed framework indicated that the peak area and peak height of the *004* plane peak were important to predict the $(BH)_{max}$. Therefore, our framework can provide highly interpretable analysis results by using peak features.

2. Data

2.1. X-ray diffraction data of neodymium magnets by hot extrusion

We measured the XRD and magnetic properties, such as the maximum energy product $(BH)_{max}$, of Neodymium magnets fabricated by hot extrusion. In the dataset denoted as $\{(\mathcal{D}_m, t_m)\}_{m=1}^{M}$, M is the number of data samples. In this paper, an XRD datapoint is denoted as $\mathcal{D} = \{(\boldsymbol{x}, \boldsymbol{y})\} = \{(x_n, y_n)\}_{n=1}^{N}$ where N is the number of data points, and a data point (x_n, y_n) indicates the observation angle 2θ [degree] and intensity [counts], respectively. Figure B2 shows a part of the measured XRD data. The symbol $\boldsymbol{t} = (t_1, t_2, \ldots, t_M)^\top$ denotes the magnetic properties, for example, the maximum energy product $(BH)_{max}$.

In this measurement data, the number of XRD data was 176, and one XRD dataset had 9000 data points. The range of the observation angle 2θ was 10.0–100.0 [degree], and the observation step was 0.01 [degree]. The intensity unit of XRD datapoints was the counting unit [counts].

3. Concept

Concept of our method is that the peak feature extraction from the measurement XRD data and the prediction of properties from the peak feature extraction are consistently performed by Bayesian estimation framework without any crystal phase information at all. Figure B1 shows a schematic diagram of our analysis procedure. As shown in the Figure B1, it is important to perform the step of projecting the measurement data into the peak feature space. It allows for a highly interpretable and robust analysis. In particular, we focused on the peak features to achieve high interpretability. To realize the construction of feature tables, the proposed method extracted peak features for ROIs with large peak intensities. This method estimated the posterior distribution of the peak features by extracting the peaks through Bayesian estimation using the replica exchange Monte Carlo method. This facilitated discussions on the estimation accuracy of the extracted peak features. Furthermore, the method provided analyst-independent global solutions so that the extracted peak features were less analyst-dependent. In addition, we regressed a material property on the peak features using Bayesian linear regression with feature selection. Thus, the proposed method provides peak extraction and regression analysis based on a full Bayesian framework.

4. Result and discussion

In this study, we constructed an ML-ready peak feature table from XRD data following three steps. Figure B3 shows the schematic of the three steps: (Step 1) Peak regions of interest (ROI) extraction, (Step 2) XRD peak decomposition using Bayesian estimation, (Step 3) Bayesian linear regression with feature selection. This section discusses the results of Step 2 and 3. We described the results of Step 1 in Appendix B.1 since Step 1 is a subordinate claim of this paper.

4.1. Peak feature extraction using peak-top fitting

Extracting peak features is crucial for the understanding of humans. This sub-section presents the results of XRD peak fitting. The peak decomposition model is described in Appendix A.1. describes the setting of the prior distributions in the Bayesian estimation of the peak features. In the defined model, we set the intensity spread σ_h, the position spread σ_c, width mode μ_σ, and width spread σ_σ as the hyper-parameters of the prior distribution. We set the hyper-parameters of the prior distribution to $\sigma_h = 2.0$, $\sigma_c = 1.0$, $\mu_\sigma = 0.1$, $\sigma_\sigma = 0.2$. Our method extracted peak features focusing on the four ROIs. We performed peak fitting by assuming the existence of one dominant main peak in the ROI for improved interpretability and high calculation throughput. Thus, the fine structure of the peak hem was not considered in this study.

However, if the minor peaks cannot be assigned, a bias is observed in the estimation of the major peaks. Therefore, in this study, a method was developed to automatically extract the major peak features. In this case, the bias of the minor peaks was removed by using information related only to the peak tops.

Figure B4 presents an example of the peak fitting result of our method. In this study, we performed peak fitting using 30 points around the peak top. In this figure, the gray scatter plot indicates all data points, and the blue, orange, green, and red peaks are the fitted peaks. We indexed the peaks in order from the low-angle side. As shown in Figure B4, the extracted four peaks well represented the characteristic of high-intensity diffraction peaks. We describe the posterior distribution of this fitting in the Appendix B.3. Our method transformed 9000 observation data points into 12 parameters $\Theta = \{h_k, c_k, \sigma_k | k = 1, \ldots, \mathrm{K} = 4\}$ by peak fitting. We extracted the peak features exhibiting a Bayesian posterior distribution from 176 XRD datapoints in approximately 180 minutes, despite the sequential analysis of XRD data. Therefore, our method facilitated obtaining peak features with confidence intervals in approximately 1 minute. We shows the heatmap of the peak feature table in the Appendix B.4.

Peak features are robust to measurement perturbations and statistical noise. Multivariate analysis, which is often used, provides us with a mixture of spectral changes originating from the measurement environment and those linked to the material properties. Therefore, it is not always the case that all of the basis spectral shapes are linked to the properties. Our proposed framework projects spectra into semantic space (position, height and width) by introducing a simple Bayesian peak separation. We then use Bayesian linear regression to perform feature selection in the property prediction task exactly. Our method allows us to isolate each peak and its shape parameter and discuss its correlation with the properties. In addition, when we know the resolution of the instrument, the peak features are easy to correct. In contrast, features in spectral space are more difficult to correct than peak feature. Moreover, our method allows us to normalize each peak feature separately by converting

it to a table of peak features. This allows us to treat large and small changes in the peaks equally. Normalizing the measured data by 2θ would overemphasize the observed noise in the lower intensity areas.

If we consider 2θ space (measurement data) as is the feature without considering z, its dimension is 9000. We can reduce the dimension of features from 9000 to dozens (16 dimensions in this case) by Bayesian peak separation in a simple model. By converting to a table of peak features, the characteristics of each peak can be normalized separately. This allows for treating changes in large and small peaks equally. If the measurement data were normalized by 2θ, observation noise would be overemphasized in areas of low intensity.

We also considered the peak area a_k as a peak feature. In this study, we denoted the feature vector for material properties as $\boldsymbol{z} = \{h_k, c_k, \sigma_k, a_k | k = 1, \ldots, \mathrm{K} = 4\}$. In the next sub-section, we present the results of Bayesian linear regression for $(BH)_{max}$ using peak features.

4.2. Bayesian linear regression with feature selection

This sub-section shows the prediction of the magnetic property $(BH)_{max}$ using the peak features. In particular, our method automatically selected the peak feature needed to perform predictions using Bayesian estimation. We described the Bayesian linear regression model in Appendix A.2. In this study, we set the variance λ^{-1} to 0.05. Further, we used the replica exchange Monte Carlo (REMC) method in subsection A.3 to sample from the poster distribution $p(\boldsymbol{g}, \boldsymbol{w} | \boldsymbol{t}, \boldsymbol{z})$. As a result, we estimated the optimal noise level (the optimal inverse temperature) using Bayesian free energy.

Figure B5 shows the prediction results of the magnetic property $(BH)_{max}$ using the peak features. As shown in Figure B5, the regression results were reasonably good. This good regression result was achieved with only four features $\{h_1, a_1, h_3, a_3\}$ estimated by feature selection using our method. The feature subset $\{h_1, a_1, h_3, a_3\}$ indicates the peak area and height of peaks 1 and peak 2, respectively.

Figure B5 shows the ranking of peak feature combinations. The x- and y-axes denote the feature labels and probability, respectively. The color indicates the estimated weight coefficients for each combination. This figure presents the rankings from 1–15. The combination on the left had a higher ranking. The regression results for the leftmost combination on are same as those shown in Figure B5. As shown in Figure B6, the feature h_3, a_3 is an important feature that is essential for prediction because the features h_3 and a_3 were frequently used in higher rankings. Similarly, the features h_1 and a_1 are also considered as comparatively important features.

Comparison with the reference peak list of the $Nd_2Fe_{14}B$ phase, the main phase of the Neodymium magnet, shows that peaks 1 and 3 correspond to the c-planes of *006* and *004*, that is the axis of easy magnetization. Then we considered the weight coefficients of the *004* peak area and height, which were the most important features to predict $(BH)_{max}$. The sign coefficients of the area and height were positive (red) and negative (blue), respectively, indicating that the low peak height and large peak area are expected for high $(BH)_{max}$. This suggests that decreasing crystallite size improves $(BH)_{max}$. Thus, our method provided highly interpretative results because it extracted the peak features and selected a few peak features for predicting $(BH)_{max}$.

5. Limitation and scope

Our framework calculated the features by focusing on ROIs with large peak intensities to accelerate the process and eliminate the intentional genre of an analyst. Therefore, our framework did not consider small peaks of intensity. Thus, the method cannot extract important features if small amounts of impurities contribute significantly to the material properties.

The proposed method required a certain number of ROIs to be the focus of attention. The larger the number of ROIs considered, the more likely it is to obtain important features, although this increases the computational complexity of feature extraction. It is recommended to set the number of ROIs at an acceptable computational cost.

6. Conclusion

Our framework was developed to perform peak feature extraction from XRD data and regression analysis with feature selection on material properties using the peak features, all using Bayesian estimation. By using Bayesian estimation, the reliability (Bayesian posterior distribution) of the extracted features and regression results can be discussed. In addition, in this method, the peak features were extracted robustly by focusing only on the peak tops of the peak regions of interest. As a result, an MI-corresponding feature table of major peak features can be constructed from the XRD measurement data.

We applied our framework to XRD data of Neodymium magnets. We extracted the peak features with Bayesian posterior distribution from 176 XRD datapoints in about 180 min even though XRD data was analyzed in sequence. Therefore, our method enabled us to obtain peak features with confidence intervals in approximately 1 min. The magnetic properties were

subjected to Bayesian linear regression using the extracted peak features and reasonably good regression results were obtained. The peak height and peak area of the *004* and *006* peaks were important as the necessary features, which were reasonable and interpretable.

Disclosure statement

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

Funding

This work was supported by MEXT Program: Data Creation and Utilization-Type Material Research and Development Project Grant Number JPMXP1122715503, and a Grant-in-Aid for Scientific Research [KAKENHI Grants No. 19H05819].

ORCID

Ryo Murakami http://orcid.org/0000-0001-8585-9268
Keitaro Sodeyama http://orcid.org/0000-0002-9228-0729

References

[1] Taupin D. Automatic peak determination in x-ray powder patterns. J Appl Crystallogr. 1973;6 (4):266–273. doi: 10.1107/S0021889873008666
[2] Huang TC. Precision peak determination in x-ray powder diffraction. Australian J Phys. 1988;41 (2):201–212. doi: 10.1071/PH880201
[3] Rodríguez-Carvajal J. Recent advances in magnetic structure determination by neutron powder diffraction. Physica B: Condens Matter. 1993;192(1–2):55–69. doi: 10.1016/0921-4526(93)90108-I
[4] Suzuki Y, Hino H, Hawai T, et al. Symmetry prediction and knowledge discovery from x-ray diffraction patterns using an interpretable machine learning approach. Sci Rep. 2020;10(1):12. doi: 10.1038/s41598-020-77474-4
[5] Suzuki Y. Automated data analysis for powder x-ray diffraction using machine learning. Synchrotron Radiat News. 2022;35(4):9–15. doi:10.1080/08940886.2022.2112496
[6] Surdu V-A, Győrgy R. X-ray diffraction data analysis by machine learning methods—a review. Appl Sci. 2023;13(17). doi: 10.3390/app13179992
[7] Greasley J, Hosein P. Exploring supervised machine learning for multi-phase identification and quantification from powder x-ray diffraction spectra. J Mater Sci. 2023 03;58(12):5334–5348. doi: 10.1007/s10853-023-08343-4
[8] Suzuki Y, Taniai T, Saito K, et al. Self-supervised learning of materials concepts from crystal structures via deep neural networks. Mach Learn: Sci Technol. 2022 dec;3(4):045034. doi: 10.1088/2632-2153/aca23d
[9] Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett. 2018;120(14):145301. doi:10.1103/PhysRevLett.120.145301
[10] Ziletti A, Kumar D, Scheffler M, et al. Insightful classification of crystal structures using deep learning. Nat Commun. 2018;9(1):2775. doi: 10.1038/s41467-018-05169-6
[11] Choudhary K, DeCost B. Atomistic line graph neural network for improved materials property predictions. Npj Comput Mater. 2021;7(1):185. doi: 10.1038/s41524-021-00650-1
[12] Kovacs A, Fischbacher J, Oezelt H, et al. Physics-informed machine learning combining experiment and simulation for the design of neodymium-iron-boron permanent magnets with reduced critical-elements content. Front Mater. 2023;9 (1094055):1094055. doi: 10.3389/fmats.2022.1094055
[13] Gerlach RW, Kowalski BR, Wold HOA. Partial least-squares path modelling with latent variables. Anal Chim Acta. 1979;112(4):417–421. doi:10.1016/S0003-2670(01)85039-X
[14] Wold S, Sjöström M, Eriksson L. Pls-regression: a basic tool of chemometrics. Chemom And Intell Lab Syst. 2001; 58(2):109–130. doi: 10.1016/S0169-7439(01)00155-1
[15] Nagata K, Sugita S, Okada M. Bayesian spectral deconvolution with the exchange monte carlo method. Neural Networks. 2012;28(28):82–89. doi:10.1016/j.neunet.2011.12.001
[16] Fancher CM, Han Z, Levin I, et al. Use of bayesian inference in crystallographic structure refinement via full diffraction profile analysis. Sci Rep. 2016;6 (1):2016. doi: 10.1038/srep31625
[17] Mikhalychev A, Ulyanenkov A. Bayesian approach to powder phase identification. J Appl Crystallogr. 2017 Jun;50(3):776–786. doi: 10.1107/S1600576717004393
[18] Murakami R, Matsushita Y, Nagata K, et al. Bayesian estimation to identify crystalline phase structures for x-ray diffraction pattern analysis. Sci Technol Of Adv Mater: Methods. 2024;4(1):2300698. doi: 10.1080/27660400.2023.2300698
[19] Obinata K, Nakayama T, Ishikawa A, et al. Data integration for multiple alkali metals in predicting coordination energies based on bayesian inference. Sci Technol of Adv Mater: Methods. 2022;2 (1):355–364. doi: 10.1080/27660400.2022.2108353
[20] Hukushima K, Nemoto K. Exchange monte carlo method and application to spin glass simulations. J Phys Soc of Jpn. 1996;65(6):1604–1608. doi: 10.1143/JPSJ.65.1604
[21] Thompson P, Cox DE, Hastings JB. Rietveld refinement of Debye–Scherrer synchrotron X-ray data from Al_2O_3. J Appl Crystallogr. 1987 Apr;20(2):79–83. doi: 10.1107/S0021889887087090
[22] Alan Young R. *The Rietveld method*, volume 5. Oxford, England: International union of crystallography; 1993.
[23] Erb D. pybaselines: a python library of algorithms for the baseline correction of experimental data; 2023. https://pybaselines.readthedocs.io/en/latest/citing.html.
[24] Okajima K, Nagata K, Okada M. Fast bayesian deconvolution using simple reversible jump moves. J Phys Soc Jpn. 2021;90(3):034001. doi: 10.7566/JPSJ.90.034001

Appendix A. Method

A.1. Model — Fitting model for peak features

We consider the observation process of the XRD data. The XRD data $y_n \in \mathbb{N}$ were observed from a Poisson distribution $P(y_n|f_n)$ with the profile function $f_n = f(x_n;\Theta)$ as expected value:

$$y_n \sim P(y_n|f_n) = \frac{f_n^{y_n} e^{-f_n}}{y_n!}. \tag{A1}$$

The profile function $f(x_n;\Theta)$ can be represented by linear sum of the peak signal $S(x_n;\Theta)$ and the background $B(x_n)$:

$$f(x_n;\Theta) = S(x_n;\Theta) + B(x_n), \tag{A2}$$

where (x_n, y_n) denotes the measured data points, the function $S(x_n;\Theta)$ denotes the peak signal, the function $B(x_n)$ denotes the background, and Θ is the profile parameter set.

The peak signal $S(x_n;\Theta)$ can be represented by a linear sum of the peak functions, which is the Voigt function $V(x_n)$ [21]:

$$S(x_n;\Theta) = \sum_{k=1}^{\mathrm{K}} h_k V(x_n; c_k, \sigma_k), \tag{A3}$$

$$\text{where } V(x_n; c_k, \sigma_k) = \int_{-\infty}^{\infty} G(x'; c_k, \sigma_k) L(x_n - x'; c_k, \sigma_k) dx'. \tag{A4}$$

The most common profile functions are the Voigt function $V(x_n)$, which is a convolution of the Lorenz and Gauss functions, and the Pearson VII function, which is a generalization of the exponential part of the Lorenz function [22]. We employed the forked function, which is the most commonly used.

The integer value K is the number of peaks, and the parameter set Θ is $\Theta = \{h_k, c_k, \sigma_k\}_{k=1}^{\mathrm{K}}$. The functions $G(x_n)$ and $L(x_n)$ are the Gaussian and Lorentzian functions, respectively. Although Gaussian and Lorentzian widths are often defined with different parameters, they are defined with the same parameters for simplicity in this study. This method focuses only on peak tops in order to easily separate peaks in XRD data. Therefore, since the hem of the XRD peak cannot be discussed, a model focused on extracting only the width of the peak profile was applied. The estimated parameter set $\hat{\Theta}$ corresponded to the peak features for predicting material properties.

We present a set of prior distributions as follows:

$$h_k \sim \mathcal{G}\big(h_k|\eta = \eta(\mu_h, \sigma_h), \theta = \theta(\mu_h, \sigma_h)\big),$$

$$c_k \sim \mathcal{N}(c_k|\mu = \mu_c, \sigma = \sigma_c),$$

$$\sigma_k \sim \mathcal{G}\big(\sigma_k|\eta = \eta(\mu_\sigma, \sigma_\sigma), \theta = \theta(\mu_\sigma, \sigma_\sigma)\big),$$

where the probability distributuin $\mathcal{G}(\eta, \theta)$ and $\mathcal{N}(\mu, \sigma)$ are the Gamma and the Gaussian distributions, respectively. The parameters η and θ are shape and scale parameter of the Gamma distribution, respectively. Here, the functions $\theta(\mu, \sigma)$ and $\eta(\mu, \sigma)$ calculated the shape and scale parameter of the Gamma distribution from the mode μ and the variance σ^2, respectively. The functions $\theta(\mu, \sigma)$, $\eta(\mu, \sigma)$ can be expressed as follows:

$$\theta(\mu, \sigma) = \frac{1}{2}\left(-\mu + \sqrt{\mu^2 + (2\sigma)^2}\right),$$

$$\eta(\mu, \sigma) = 1 + \frac{\mu}{\theta}.$$

Note that this variance σ^2 and peak width σ_k have different meaning.

As the peak height and width have positive values: $h_k \in \mathbb{R}^+$ and $\sigma_k \in \mathbb{R}^+$, we set the Gamma distribution to a prior distribution of h_k and σ_k. We set the position (angle) and intensity of maximum value in the ROI to μ_c and μ_h. We estimated the background $B(x_n)$ using pybaselines [23] before peak extraction. The pybaselines can perform backgroud estimation with model-free by the iterative least squares method mostly.

A.2. Model — Bayesian linear regression with feature selection

The linear regression model with feature selection is represented by an output t_m, an input (features) vector $z_m \in \mathbb{R}^D$, indicator $g \in \{0, 1\}^D$ [24], the weight coefficients $w \in \mathbb{R}^D$, and statistical noise ε_m as follows:

$$t_m = (g \circ w)^{\mathrm{T}} z_m + \varepsilon_m, \tag{A5}$$

where the operation $\circ$ denotes the Hadamard product. In this study, statistical noise ε_m followed a Gaussian distribution with variance λ^{-1}: $\varepsilon_m \sim \mathcal{N}(\varepsilon_m|0, \lambda^{-1})$.

Therefore, the probability distribution of the output t_m is represented as follows:

$$p(t_m|g, w, z_m) = \mathcal{N}(t_m|(g \circ w)^{\mathrm{T}} z_m, \lambda^{-1}). \tag{A6}$$

The prior distribution of the weight coefficients $p(w)$ was set to an uninformative broad distribution. The prior distribution of the indicator $p(g)$ is the Bernoulli distribution with the Bernoulli distribution with the expected value of 0.5.

The poster distribution of parameters $\boldsymbol{w}$, $\boldsymbol{g}$ is represented as follows:

$$p(w, g|z, t) \propto p(w)p(g) \prod_{m=1}^{M} p(t_m|g, w, z_m), \tag{A7}$$

$$\propto \prod_{m=1}^{M} \mathcal{N}(t_m|(g \circ w)^{\mathrm{T}} z_m, \lambda^{-1}). \tag{A8}$$

This study approximated an exhaustive state search by sampling the indicators $\boldsymbol{g}$ using the replica exchange mcmc method.

A.3. Algorithm — Replica exchange Monte-carlo method

We performed posterior visualization and the maximum a posteriori (MAP) estimation through sampling from the posterior distribution. A popular sampling method is the Monte Carlo (MC) method, which may be bounded by local solutions for cases when the initial value is affected or the cost function landscape is complex.

Therefore, the replica exchange Monte Carlo (REMC) method [15, 20] was used to estimate the global solution. For sampling using the REMC method, a replica was prepared with the inverse temperature β introduced as follows:

$$p(\theta|y;\beta=\beta_\tau)=\exp(-\beta_\tau E(\theta))p(\theta), \tag{A9}$$

where the function $E(\theta)$ denote a loss function, this is, a negative log likelihood with a parameter set θ. The inverse temperature β is $0=\beta_1<\beta_2<\cdots<\beta_\tau<\beta_T=1$. For each replica, the parameters were sampled using the Monte Carlo method.

Appendix B. Supplement

B.1. Extraction of peak regions of interest (ROI)

The proposed framework first extracted the ROI. In this study, we used PLS regression. We provided a detailed explanation of PLS regression in Appendix B.2. PLS regression reduces the dimensionality of the input data to latent variables based from which the objective variable (material properties) can be predicted. We decided that the ROIs should be those with information that could predict the objective variable. We removed the background from the XRD data as a pre-processing step for the PLS regression since there was no significant change in the background in the present XRD dataset. We removed the background using pybaselines [23] before performing PLS regression. The pybaselines can perform backgroud estimation with model-free by the iterative least squares method mostly.

Figure B7 shows the basis components obtained by PLS regression. A correlation with the objective variable t_m was observed. We present the results of the PLS regression. Figure B7 (a) show the 1st PLS component. We show the peak search results in the 1st PLS component with red dots. We show some XRD data for reference in Figure B7 (b). We performed peak detection using an absolute intensity threshold of 20% of the maximum absolute intensity of the PLS basis. As shown in Figure B7, there were four diffraction angle regions (ROI) of high component intensity. We analyzed the four regions of interest (ROI) obtained by peak search.

We could consider using all the peaks that can be detected. However, it would take an enormous amount of time to compute all peak features with confidence intervals (Bayesian posterior probability) using Bayesian inference. Moreover, if the presence or absence of fine peaks is estimated, it is difficult to make a feature table with organized peak assignments. Using all peaks would result in a very high-dimensional space, and there is a large risk of falling into a local solution. Our framework aims to extract and utilize local information but robust features rather than sensitive features that can lead to various local solutions depending on initial values. We have incorporated the extraction of peak features into our framework by narrowing down the region of interest and utilizing a exact algorithm.

In this step 1, it is doesn't matter to just specify K high intensity peaks. For example, the number of peaks K can be determined by the required computation time. Our framework is not limited to using PLS regression as step 1. On the other hand, this study uses PLS regression, which uses information on property properties, to select regions of interest. Therefore, it is possible to detect areas of sensitivity to the property. For example, a peak with a large intensity will not appear in the basis component of a PLS regression if it is completely insensitive to changes in properties.

This study focused on the four ROIs and performed the XRD peak decomposition. The result of peak decomposition is presented in sub-section 4.1.

In this study, we used the four peak regions of interest. If the regression is not successful with only the peak features in these regions, it is possible to lower the threshold for peak detection and utilize more fine peaks. Our method can evaluate the effectiveness of the focused peak features by performing Bayesian estimation of the peak features and properties. This method is also capable of searching for a better peak regions of interest by repeating the three steps.

B.2. Partial least squares regression

The partial least squares regression, which is known as PLS regression, is one of the regression analysis methods for linear models [13,14]. The PLS regression is a method that is very similar to the principal component analysis (PCA), which is a well-known dimensionality reduction method.

The main difference with PCA is that the PLS conversion is supervised. PCA extracts principal components so that the variance of the latent variable is maximized, while PLS extracts principal components so that the variance of the latent variable and the objective variable is maximized. In PCA, fluctuations in data that are not related to material properties, such as fluctuations caused by measuring equipment, also affect the extracted principal components. PLS has the advantage of being able to extract the principal components that are correlated with material properties.

The aim of PLS regression is to extract latent variables $U\in\mathbb{R}^{M\times L}$ from target variables $t\in\mathbb{R}^M$ and explanatory variables $Y\in\mathbb{R}^{M\times N}$ which is spectrum in this case. The integer value L denote the dimension of the latent variable. We give the basic formula for PLS regression below:

$$\begin{aligned} Y&=\sum_{l=1}^{L}u_l p_l^{\mathrm{T}}+d=UP^{\mathrm{T}}+d,\\ t&=\sum_{l=1}^{L}u_l q_l+e=Uq+e, \end{aligned} \tag{B1}$$

where $p_l\in\mathbb{R}^N$ denote l-th basis spectrum described with latent variable $u_l\in\mathbb{R}^M$. The vector p_l is generally referred to the l-th loading. The vector $q\in\mathbb{R}^L$ is the coefficients to convert from u_l to target variables $\boldsymbol{t}$. The vectors $\boldsymbol{d}$ and $\boldsymbol{e}$ denote the reconstruction error of the spectrum $\boldsymbol{Y}$ and the regression error of the target variable $\boldsymbol{t}$. Note that PLS regression assumes noise $\boldsymbol{d}$ which follows a Gaussian distribution, which is not an appropriate modeling from the perspective of the observation process. It is important to develop PLS regression that assumes noise that follows a Poisson distribution, and this is a topic for the future.

Assuming that u_l can be expressed as a linear combination of Y, we express u_l as follows:

$$u_l=Yw_l. \tag{B2}$$

First, we calculate the first ($l=1$) principal component u_1. We determine c to maximize the covariance ($t^{\mathrm{T}}u_1$) between $\boldsymbol{t}$ and u_1 subject to $\parallel w_l \parallel=1$. The vector w_1 can be calculated by $w_1=Y^{\mathrm{T}}t/\parallel Y^{\mathrm{T}}t\parallel$ from the method of Lagrange multiplier. After obtaining w_1, we calculate u_1 using the equation 11. Then, we calculate p_1 and q_1 to minimize the residual sum of squares $\boldsymbol{d}$ and $\boldsymbol{e}$ by $p_1=Y^{\mathrm{T}}u_1/\parallel u_1^{\mathrm{T}}u_1\parallel$ and $q_1=t^{\mathrm{T}}u_1/\parallel u_1^{\mathrm{T}}u_1\parallel$ from the least squares method. Through this process, we can calculate the first principal component.

Second, we calculate the partial data $Y_{\setminus 1}$, $t_{\setminus 1}$ that cannot be expressed by the first principal component:

$$Y_{\backslash 1} = Y - u_1 p_1^{\mathrm{T}}, \quad t_{\backslash 1} = t - u_1 q_1. \tag{12}$$

By applying a similar process to the partial data $Y_{\backslash 1}$, $t_{\backslash 1}$, we calculate the second ($l = 2$) principal component. Subsequently, we can calculate the third principal component and beyond ($l \geq 3$) in the same way.

B.3. Posterior distribution of peak parameters

Figure B8 shows the history of the cost function (the negative log-likelihood) and posterior distribution of each peak parameter: (b) peak height, (c) peak position, and (d) peak width. Figure B8 corresponds to the fitting in Figure B4. We performed burn-in on 3000 samples, and then performed production sampling on 2000 samples. A total of 5000 MC sampling steps were performed for fitting of one spectral datapoint. To reduce the correlation between samples, we visualised the posterior distribution using the history of 1000 MC samples at one sample interval. As shown in Figure Figure B4 (b)–(d), we believed that the posterior distributions had sufficient sharpness to use material description feature. In particular, peak positions were estimated with high precision. If the posterior distribution is very broad, we can eliminate this feature from the explanatory variables. Our method can provide a posterior distribution for discussing the reliability of the features.

B.4. Peak feature table

Figure B9 shows the heatmap of the feature vector standardized at each feature. The x- and y-axes show the index of XRD data and label of a feature, respectively. The suffix of the feature label denotes the peak index. The heatmap was sorted by $(BH)_{max}$ values. In this figure, warm colors exhibited a larger value and cold colors had a smaller value. This figure shows the ML-ready table data. This figure shows a low-dimensional feature representation of high-dimensional spatial data projected through a peak model to a lower dimension by Bayesian estimation. The machine learning (ML)-ready feature table is the table data that is able to input to the ML library such as Scikit-learn. This corresponds to the mid-level (features) z of the three-step data-driven analysis strategy in Figure B1. It can also be understood as a projection of high-dimensional XRD measurement data into an interpretable variable space.

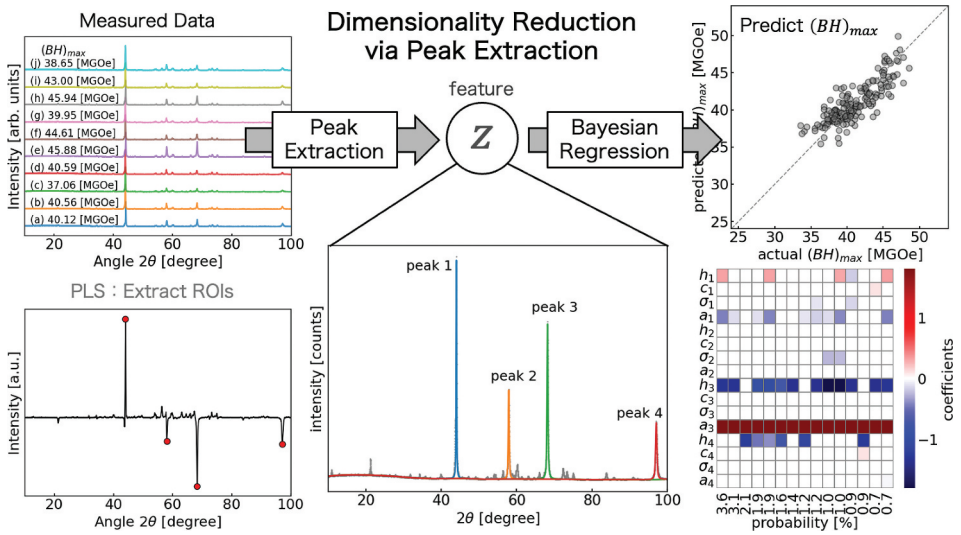


Figure B1. Schematic diagram of our analysis procedure.

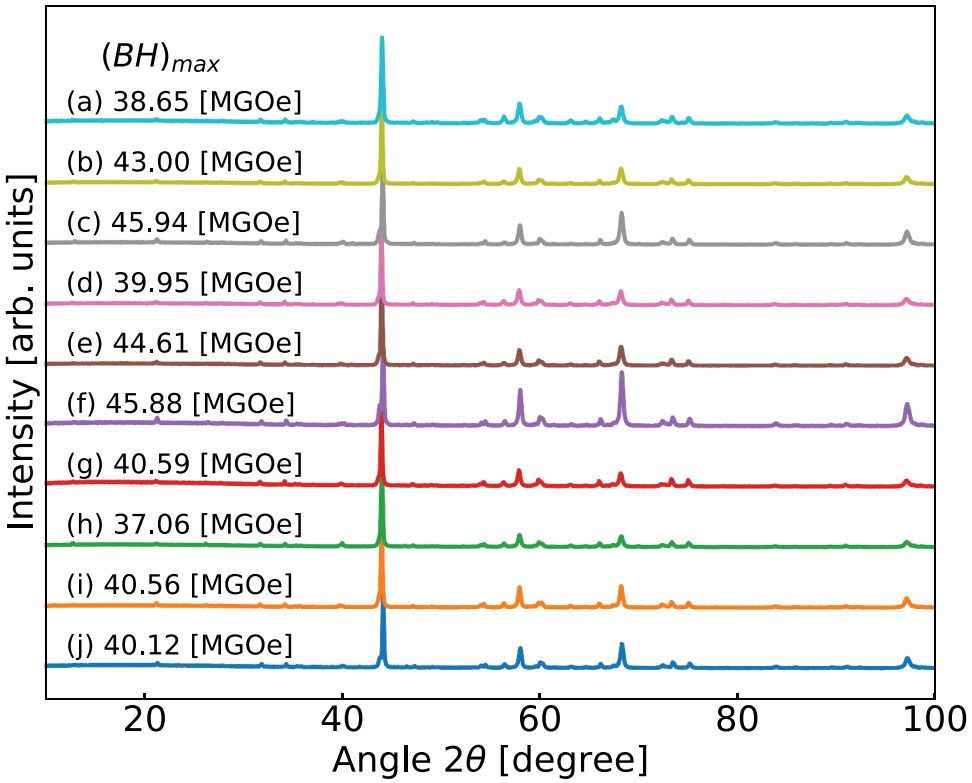


Figure B2. Part of measured X-ray diffraction (XRD) datasets of $Nd_2Fe_{14}B$ magnets under various hot extrusion conditions. The hot extrusion temperature T_{ext} [°C] and load limit F_{ext} [kN] for each data are shown as follows: [(a) $T_{ext} = 750$, $F_{ext} = 70$, (b) $T_{ext} = 750$, $F_{ext} = 100$, (c) $T_{ext} = 750$, $F_{ext} = 60$, (d) $T_{ext} = 750$, $F_{ext} = 60$, (e) $T_{ext} = 750$, $F_{ext} = 50$, (f) $T_{ext} = 775$, $F_{ext} = 50$, (g) $T_{ext} = 775$, $F_{ext} = 50$, (h) $T_{ext} = 750$, $F_{ext} = 50$, (i) $T_{ext} = 775$, $F_{ext} = 50$, (j) $T_{ext} = 750$, $F_{ext} = 35$].

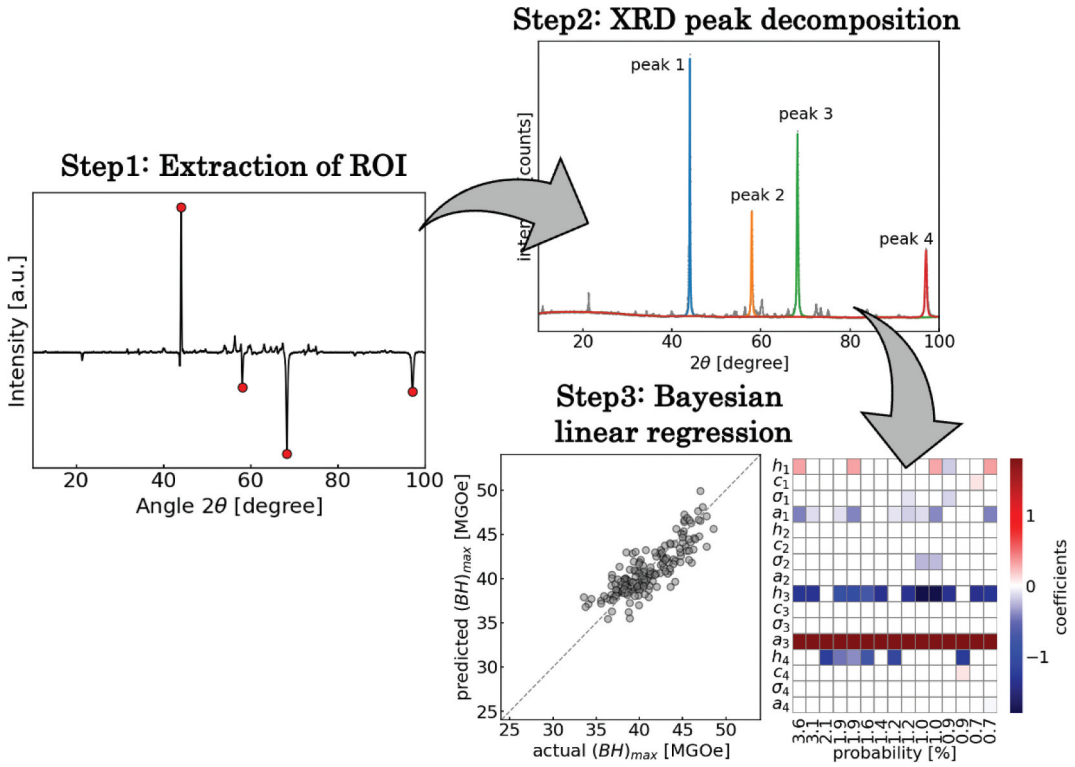


Figure B3. Analysis workflow in this study.

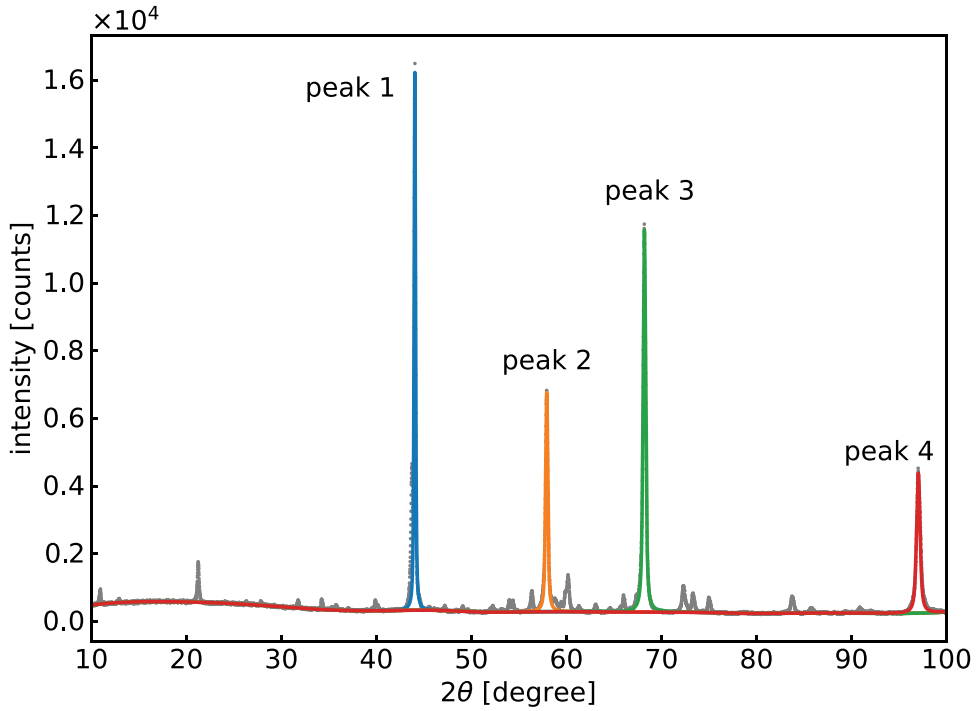


Figure B4. One example of XRD peak fitting focused on peak top in ROIs.

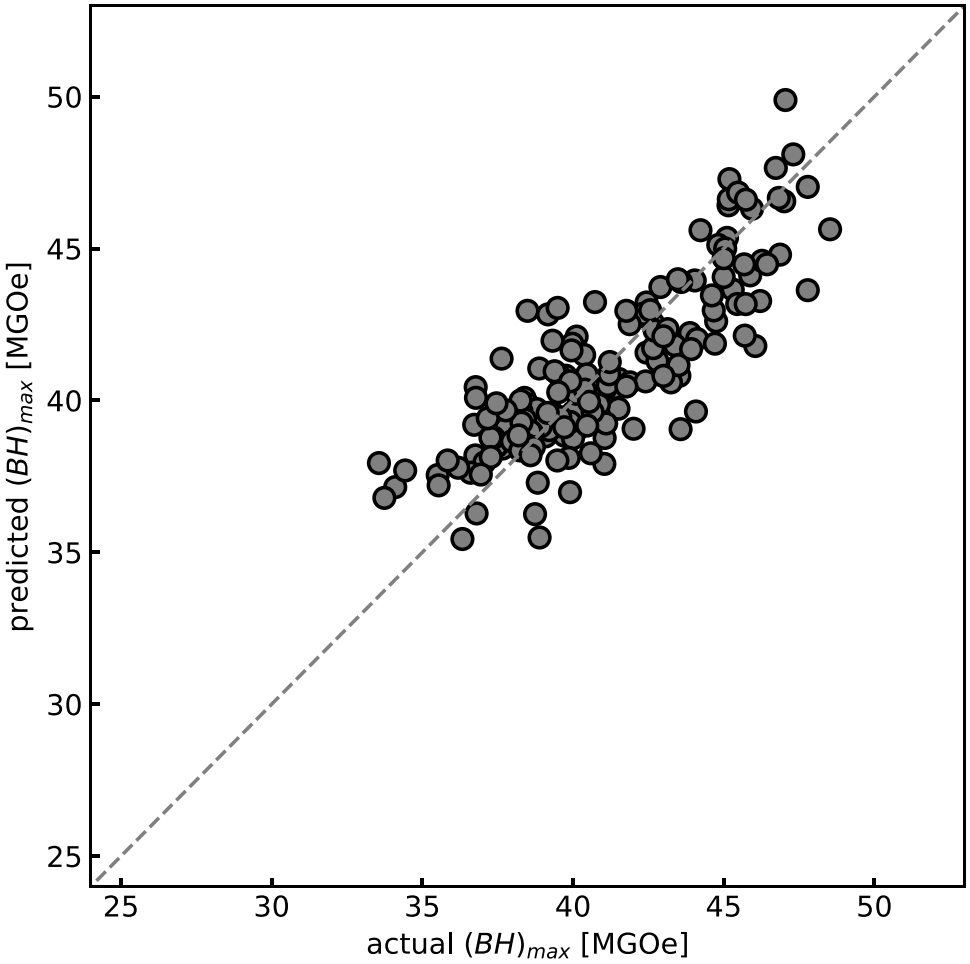


Figure B5. Results of Bayesian linear regression with feature selection. XRD peak features were used as explanatory variables.

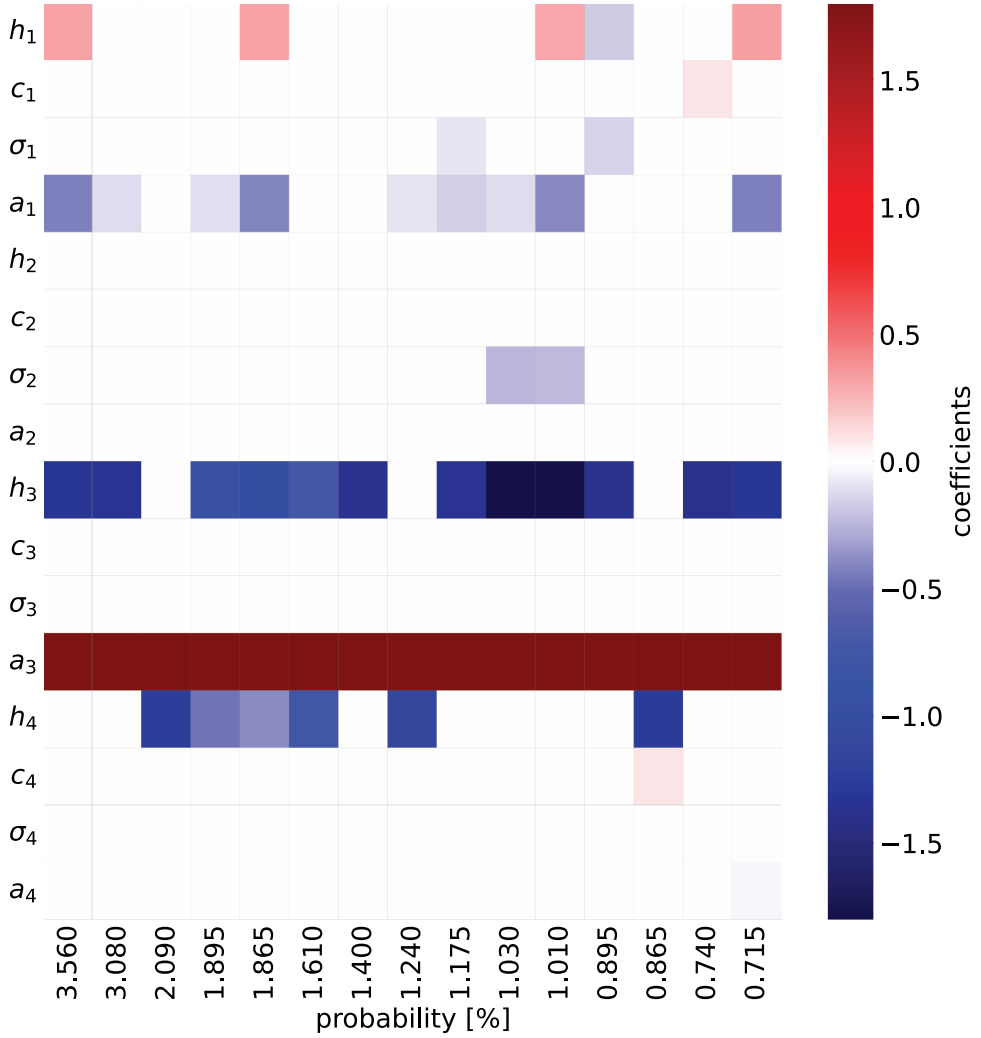


Figure B6. Results of feature selection by Bayesian linear regression. The parameter a_3 are suggested as the most possible from the regression.

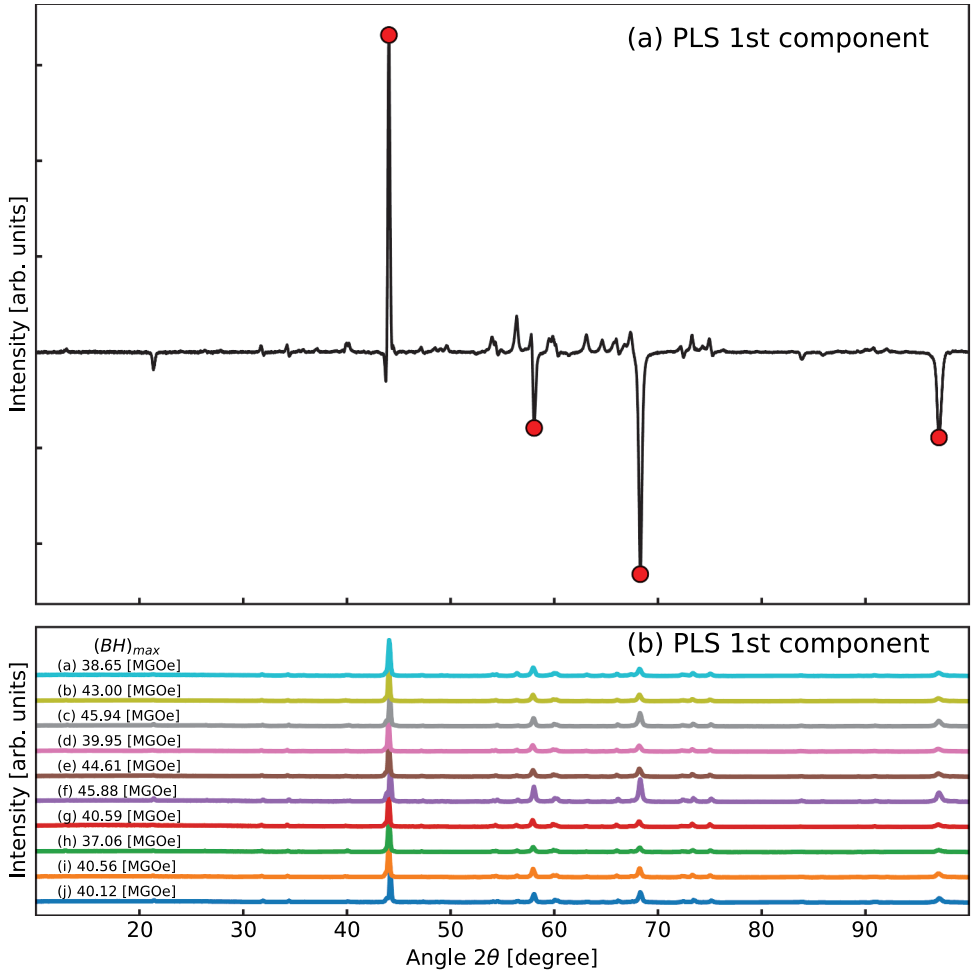


Figure B7. Results of PLS regression to extract peak regions of interest (ROI). (a) Basis component obtained by PLS regression. Red dots denote the peak search results in the 1st PLS component. (b) Some XRD data for reference.

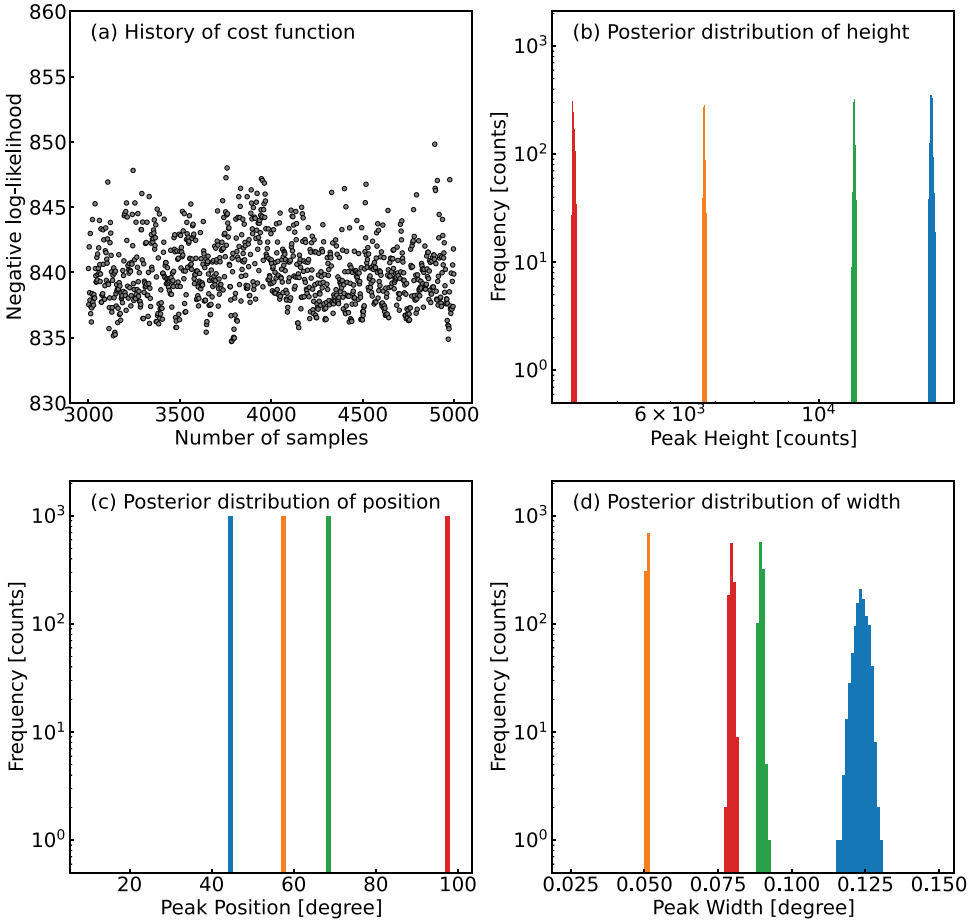


Figure B8. (a) History of cost function (the negative log-likelihood) and (b)–(d) Posterior distribution of each peak parameter: (b) peak height, (c) peak position, and (d) peak width. This figure corresponds to the Figure B4. In this figure (b)–(d), the colors denote peak ID.

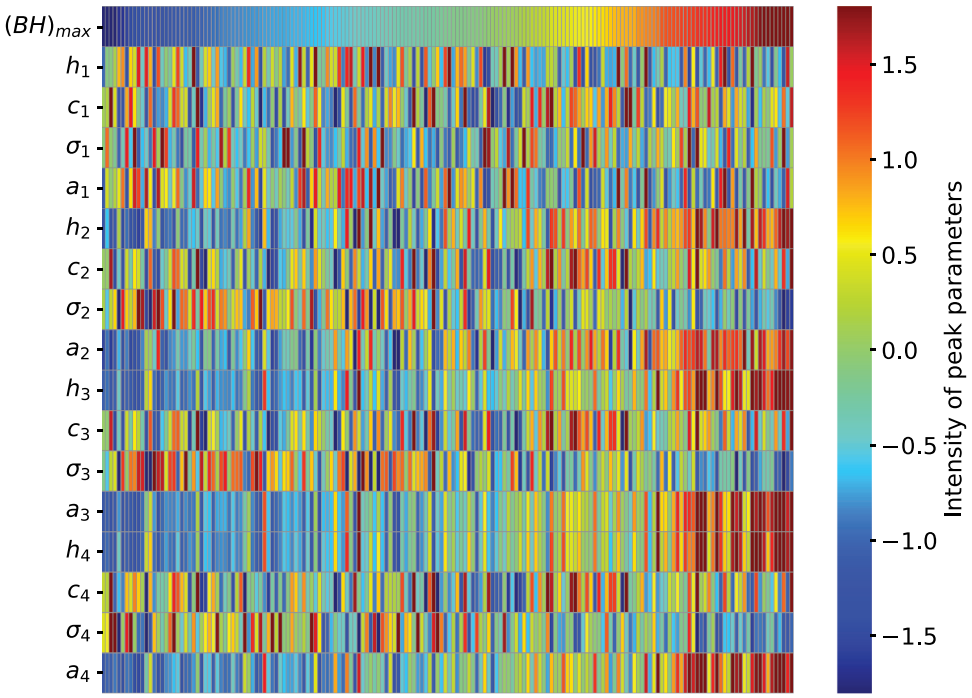


Figure B9. ML-ready XRD peak feature table built on our framework. The x-axis denotes the index of data. The suffix of the feature label denotes the index of peaks. The heatmap was sorted by $(BH)_{max}$ values.