CrystEngComm

Importance of Raw Material Features for Prediction of Flux Growth of Al_2O_3 Crystals Using Machine Learning

Journal:	*CrystEngComm*
Manuscript ID	CE-ART-01-2022-000010.R1
Article Type:	Paper
Date Submitted by the Author:	24-Feb-2022
Complete List of Authors:	Yamada, Tetsuya; Shinshu Daigaku, Center for Energy and Environmental Science Watanabe, Takanori ; Central Research Laboratories, DIC Corporation Hatsusaka, Kazuaki ; Central Research Laboratories, DIC Corporation Yuan, Jian-jun; DIC Corporation, Central Research Laboratories Koyama, Michihisa; Shinshu University Faculty of Engineering Teshima, Katsuya; Shinshu University, Department of Materials Chemistry; Shinshu University, Center for Energy and Environmental Science

Journal Name

ARTICLE

Importance of Raw Material Features for Prediction of Flux Growth of Al_2O_3 Crystals Using Machine Learning

Tetsuya Yamada,[a,b] Takanori Watanabe,[c] Kazuaki Hatsusaka,[c] Jianjun Yuan,[c] Michihisa Koyama,[a] Katsuya Teshima[a,b]

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

The flux method is an efficient liquid-phase crystal growth technique. It is expected to be one of the key technologies for the development of innovative inorganic materials in the future because it enables the production of high-quality crystals. However, owing to the complexity of the mechanism of crystal growth in fluxes, it is difficult to establish a guideline for the experimental recipe for growing crystals. Thus, flux crystal growth still needs a long process of trials and errors. Our goal is to develop "process informatics" (PI)-assisted flux method, supported by machine learning. To predict flux crystal growth by linking it to the process, essentially, the experimental parameters must be converted into explanatory variables. However, there is a limit to the explanatory power of describing crystal growth using only process conditions, such as raw materials and flux species, their preparation amounts, and heating conditions. In this study, we focused on using information on raw materials (raw material information) as explanatory variables and investigated their influence on the prediction of flux crystal growth. Aluminum oxide ($Al_2O_{3)}$, in which raw materials have abundant lot numbers, was selected as target material. After performing 185 growth experiments, we made regression models composed of process conditions and various raw material information as explanatory variables and Al_2O_3 particle size distribution as the objective variable. The obtained models clarify the effect of the raw material information on the accuracy of prediction of crystal growth. Our findings provide new insights into the PI-assisted flux method in terms of the importance of raw material information and effective descriptions. This would contribute to the development of highly accurate prediction models for data-driven experimental suggestion and clarification of important factors in flux crystal growth.

Introduction

In recent years, the development and rapid adoption of innovative materials have been required for the transition to a sustainable society. In this context, process informatics (PI) has attracted attention as an approach for material development by acquiring and analyzing the knowledge implicit in the process. It is expected that PI will realize, through effective learning from human experience and knowledge, an efficient synthesis of innovative crystalline materials that cannot be developed by conventional routines. As the flux method allows for liquid-phase crystal growth at high temperatures, high-quality inorganic crystal materials can be achieved with high crystallinity and surface development. Therefore, this method is expected to be a key technology in the future development of innovative inorganic materials. The flux method comprises major elementary processes, such as flux melting, transport of molten flux to solute, solute dissolution and re-deposition, nucleation, and crystal growth. More than 20 variables, including phase transitions, chemical reactions, heat transfer, diffusion, solvation, and supersaturation, as well as other related factors may contribute to each process. As such multidimensional factors must be considered, it is difficult to understand the mechanism of flux crystal growth and formulate a theory-based growth guideline. Consequently, actual development of crystal growth must rely on trial and error based on the experience and intuition of the experimenter, resulting in spending many years. Therefore, it is difficult to respond to demands for rapid developments in crystal optimization of various materials on an industrial scale. If PI were introduced into the flux method (PI-assisted flux method), it could lead to many innovative materials being developed in a short time.

There have been several reports of machine learning being used in crystal growth research.[1] Tsunooka et al. succeeded in rapidly predicting the convection in a melt using machine learning.[2] Boucetta et al. showed that the measurement of the location around the reaction vessel in the furnace, that is, the

[a] *Research Initiative for Supra-Materials, Shinshu University, 4-17-1 Wakasato, Nagano 380-8553, Japan*
[b] *Department of Materials Chemistry, Faculty of Engineering, Shinshu University, 4-17-1 Wakasato, Nagano 380-8553, Japan*
[c] *Central Research Laboratories, DIC Corporation, 631, Sakado, Sakura, Chiba 285-8668, Japan*
† Contact address for the corresponding author: teshima@shinshu-u.ac.jp

Electronic Supplementary Information (ESI) available: Definition of target and non-target peak areas in XRD profiles (Figure S1), Histograms for all descriptors in input data (Figure S2), Box plot for all descriptors in input data (Figure S3), Identification and extraction of XRD parameters for raw materials (Figure S4), Scatter plots and heat maps of XRD descriptors (Figure S5), Box plots and histograms of XRD descriptors (Figure S6), Correlation-coefficients plots for all data (Figure S7), Parity plots of OLS, LASSO, DT, RF, SVR, and NN for dataset without XRD, PD parameters, and without XRD parameters (Figure S8), Parity plots of RF and SVR for dataset with only raw materials, and raw material conditions descriptors (Figure S9), Model coefficients for four kinds of datasets (Figure S10), Model coefficients for dataset in table (Table S1). See DOI: 10.1039/x0xx00000x

region with a high-temperature thermal gradient, is important for predicting the thermal distribution in the furnace.[3] Dang et al. used a genetic algorithm to optimize the crystal solidification step by correlating two different heater cooling rates and a heat gate rate with crystal quality and growth time.[4] Yao et al. were able to discriminate the single crystalline nature of the target material by learning a case study of crystal growth using the flux method.[5] However, there are no examples of the PI-assisted flux method being applied to control crystallographic features quantitatively.

To implement the PI-assisted flux method, a dataset that properly represents crystal growth must be prepared. For this purpose, the key to PI crystal growth prediction is specifying what kind of features to obtain and how to describe them in the dataset. The former involves time and equipment to collect, while the latter requires accuracy of machine-learning analyses. To properly prepare the dataset, we have to understand the importance of each factor or feature, and accurately represent them as explanatory variables. When considering experimental conditions as explanatory variables, parameters such as raw material information (material type and preparation volume), flux information (flux type, preparation volume), heating conditions (heating rate, temperature, time, cooling rate), reaction vessel information (material, volume), and atmosphere (gas type, pressure) are easy candidates. However, it is questionable whether crystal growth can be accurately predicted using such simple information alone. One of concerns is insufficient features of raw material information. It is well known that the states of raw materials significantly contribute to the crystallographic characteristics and performance of the final crystal materials.[6] Yang et al. reported that the crystal size and dispersibility of the reactant, $Na_2Ti_3O_7$, is affected by the particle size of its raw material, TiO_2.[7] Yanagisawa et al. reported that the amorphous nature of the precursor is involved in the photocatalytic performance of TiO_2.[8] Huang et al. reported that the addition of organic materials affected the grain size and stable phase in the preparation of ZrO_2 nanoparticles.[9] Zaitseva et al. reported that impurities affect the crystal outline and size in the crystal growth of potassium orthophosphate crystals in solution.[10] Rajesh et al. showed that the growth rate of single crystals of potassium dihydrogen phosphate is related to the purity of their raw material.[11] These findings suggest that description of raw material information in greater detail is essential to accurately predict crystal growth.

Although we recognize the importance of raw material information, there is no guideline yet on what information to use and how to represent them. Raw material information includes grain size, shape, uniformity, crystallinity, and purity. Of these, grain size and uniformity can be obtained numerically by particle size distribution measurements. Shape and crystallinity can be obtained by electron microscopy, but this method is not so reliable in terms of consistency and reproducibility of description of the population because it may be measured with subjectivity to some degree. Therefore, we focused on X-ray diffraction (XRD) spectra as an alternative method of microscopy. The ability of XRD to describe crystallographic information has already been reported. Inoue et al. experimentally demonstrated that the XRD intensity of $CaSO_4 \cdot 2H_2O$ is related to its crystal morphology.[12] Suzuki et al. have been working on discriminating crystalline systems by machine learning of feature information obtained by XRD.[13] Several research groups have succeeded in accelerating Rietveld analysis[14,15] and identifying crystal phases[16,17] after transforming XRD into a representation using a convolutional neural network (CNN). These studies suggest that XRD simultaneously obtains information on crystal size, shape, and impurities.

In this study, we aimed to understand the importance of raw material information in improving the prediction accuracy of flux crystal growth using machine learning and to propose an effective description method. α-Al_2O_3 is aluminum oxide in its most commonly occurring crystalline form, called corundum. We chose this material as the model material in this study. Al_2O_3 is known to grow into a variety of polyhedral shapes when grown with molybdenum trioxide (MoO_3)-containing fluxes[18-24] and into plate-like shapes when grown with sulfuric acid-based fluxes.[25,26] Oxide additives are also known to change the crystal outline of Al_2O_3.[27,28] In addition, Al_2O_3 is a well-known industrial material and there are many raw materials with similar compositions but of different purity and crystallinity. Thus, we regarded Al_2O_3 as an appropriate candidate to study the effects of raw materials on flux crystal growth.

Experiments

2.1 Crystal Growth and Identification of Al_2O_3

$Al(OH)_3$ or AlO(OH) were the raw materials used to grow Al_2O_3 plate crystals following the flux method.[24, 28] Here, 40 different lot numbers of $Al(OH)_3$ and 17 different lot numbers of AlO(OH) were used as raw materials. MoO_3 (Nippon Inorganic Colour & Chemical Co., Ltd., purity > 99%) was used as flux. Various oxides were used as additives. After mixing the raw materials, flux, and additives at predetermined ratios for approximately 5 min., the mixture was poured into in an alumina crucible and placed in an electric furnace (fast heating electric furnace SC-2045D-SP, Motoyama Corporation). Crystal samples were obtained by heating and sintering under the specified temperature conditions (heating rate, holding temperature and time, and cooling rate) in an air atmosphere. After calcination, the samples were washed by ammonia water to remove the residue flux, and finally dried up at 130°C. The chemical phase was identified using an X-ray diffractometer (XRD, Ultima IV, Rigaku Corporation). The Copper kα (Cu kα) radiation (λ = 0.154 nm) was used as the characteristic X-ray with a sweep speed of 0.05° min^{-1} and a step size of 0.02° in the region of 2θ = 10°-80°. The crystal morphology of the samples was observed using a scanning electron microscope (SEM; JEOL JCM7000), JEOL Corporation) under an accelerating voltage of 15 kV. The particle size distribution was measured using a laser diffraction-type measurement system (HELOS (H3355) & RODOS, Japan Laser Corporation). The elemental composition was measured using inductively coupled plasma emission spectrometry (Optima 8300, PerkinElmer).

2.2 Data Set Preparation

Table 1 lists the experimental variables used in the study. The total number of factors for the explanatory and objective variables were 47 and 3, respectively. The explanatory variables consisted of basic variables (experimentally controllable parameters) and advanced variables (observables). In Descriptor Category 1, the basic variables are of four types: raw material, flux, additives, and heating conditions, and the numbers of variables belonging to these categories are presented as Descriptor Category 2. For possible description of the ratio of reagents, the quantities of reagents other than the raw material are expressed as wt% (percentage by weight). The heating and cooling rates were identical in all experiments and were excluded from the explanatory variables. D_x (x = 10, 50, 90) refers to the particle size when the volume ratio of particles smaller than this is x%. The objective variables were the Al_2O_3 particle size distributions D_{10}, D_{50}, and D_{90}.

In addition, we used the XRD descriptors as the raw material states as shown in Table 2. This category includes the following five items (see Descriptor Category 3). The extraction target from XRD was defined to represent the crystal face development, regularity, and impurity of the crystal. The terms for each feature and its expected role are as follows:

- Euclid_average: Represents crystal orientation and shape of raw material
- Euclid_variance: Represents the magnitude of the bias in the crystal orientation and shape of the raw material.
- FWHM_average: Represents the crystallinity of the raw material. Here, FWHM is full width at half maximum.
- FWHM_variance: Is the axis dependence of crystallinity of raw material
- Euclid_impurity: Indicates impurity state and crystallinity of raw material

The method of extracting XRD features is described below; each peak in the literature value of the XRD pattern was normally distributed and then normalized for the maximum value of the strongest peak to be 1. This artificial XRD profile was used as the XRD pattern for reference. ICDD 00-007-0324 for $Al(OH)_3$, ICDD 01-074-1895 for AlO(OH), and ICDD 00-042-1468 for Al_2O_3 were used as the XRD references. The peaks of the objects were determined if the ratio of the maximum and minimum values in the 2θ region of ±1° from the center of each peak of the comparison XRD was more than 10. The experimental values for the maximum value at the strongest peak of the XRD patterns obtained from the experiment were normalized to be 1, which was used as the XRD pattern for the experiment. The Euclidean distance of the intensity difference between each peak of the obtained XRD was used for reference, and the XRD for the experiment was calculated. The Euclidean distance *d(ref,exp)* between the literature values and experimental values is expressed in Formula 1.

$$d(ref,\exp)=\sum\sqrt{(2\theta_{ref}-2\theta_{\exp})^2+(\mathrm{I}_{ref}-\mathrm{I}_{\exp})^2} \quad \text{Formula 1}$$

Here, $2\theta_{ref}$ and $2\theta_{exp}$ are the reference and experimental values of the 2θ region (target peak area) at ±1° of each peak in the comparison XRD. I_{ref} and I_{exp} are the reference and experimental values of the peak intensity in the same region. In this study, the same material was used for both the experimental and reference samples. In reality, there is an error within 0.02°, which is the measurement resolution, but we decided to ignore it on this occasion. If we consider $2\theta_{ref}$ and $2\theta_{exp}$ to be the same, *d(ref,exp)* becomes Formula 2.

$$d(ref,\exp)=\sum\sqrt{(\mathrm{I}_{ref}-\mathrm{I}_{\exp})^2} \quad \text{Formula 2}$$

Table 1 Descriptor's information about input & output data.

	Descriptor category 1	Descriptor category 2	Descriptions	Number of factors in each category	Units
Input					
Basic variables	Raw material	raw material amount	$Al(OH)_3$, AlO(OH)	2	gram
	Flux	flux amount	MoO_3	1	wt%
	Additives	additive amount	add1-add19	19	mol%
	Heating conditions	Heating temperature	1st Heating temp., 2nd Heating temp.	2	ºC
		Heating time	1st Heating time., 2nd Heating time	2	ºC· h^{-1}
Advanced variables	Raw material conditions	Particle distribution of raw materials (PD)	D_{10}, D_{50}, D_{90}	3	µm
		Impurity elements(IE)	Imp1-imp18	18	ppm
output					
	Particle distribution of Al_2O_3		Al_2O_3_D_{10}, Al_2O_3_D_{50}, Al_2O_3_D_{90}	3	µm

The mean and variance were calculated for the Euclidean distance obtained from each peak and used as features (Euclid peak_average, Euclid_variance). The full width at half maximum (FWHM) of each peak in the experimental XRD was obtained, and its mean and variance were used as features (FWHM_average, FWHM_variance). The Euclidean distance (Euclid_impurity) between the non-target peak area of the reference XRD and experimental XRD was also calculated. The non-target peak area here is defined as the 2θ region (no XRD peak region) that does not correspond to the area of ±1° of the peak centered on each peak position shown in the XRD for reference. In Euclid_impurity, I_{ref} in Formula 2 is almost zero because it is the intensity of the base region of the normal distribution created by the peak. Figure S1 shows the definitions of the target and non-target peak areas of $Al(OH)_3$ and AlO(OH) in the XRD pattern. The total number of explanatory variables after addition of XRD features was 52.

2.3 Statistical Analysis

Through executing the extraction algorithm of XRD features, we also automatically identified the chemical phase for the obtained 185 samples. As a result, single-phase Al_2O_3 was achieved at 170 samples, and these were used for further analyses. The machine-learning analysis was performed using the Python programming language (version 3.8), programmed using Spyder 4.2 in the Anaconda software package. To prevent multicollinearity, one of the explanatory variables with a correlation coefficient of 0.95 or higher was eliminated. Next, regression analysis was conducted using the cleaned data. Six algorithms were used: ordinary least squares (OLS), least absolute shrinkage and selection operator (LASSO), decision tree (DT), random forest (RF), support vector regression (SVR), and neural network (NN). For regression analyses, we focused on the clarification of the relation between prediction accuracy and raw-material information along with the identification of the important descriptors. We chose these prevalently employed algorithms because of the following considerations: OLS and LASSO are linear regression algorithms, and we used them to first verify the linearity with sparse modelling. DT and RF were adopted to classify the importance of the descriptors in linear or non-linear manners. SVR and NN algorithms were adopted to verify the accuracy depending on descriptors, in contrast to the visualization of important descriptors. The selection for descriptors in OLS and the hyperparameters for the other regression methods were tuned in the following ways:

- OLS: Stepwise method was used for feature selection. Variables with a significance level greater than 5% obtained by model fitting were removed. This process was repeated five times, and the remaining features were used to create the final model.
- LASSO: penalty term λ (2^{-n} (n = 0 ~ 15))
- DT: The max depth = (2, 4, 6, 8, 10), and the maximum leaf nodes = (2, 4, 6, 8, 10)
- RF: The number of trees = (5, 10, 30, 50), the maximum depth = (2, 5, 15), the max feature number is the number of explanatory variables, and the minimum number of samples available to split = (2, 4, 8)
- SVR: The kernel = (linear, poly, rbf), the cost parameter = (1,10,15), and epsilon = 0.1
- NN: The number of hidden layers = (1, 2, 3), and the number of neural net nodes was set to 100 at each layer. Further, batch size = (20, 40, 60), and the max number of iterations was 10000.

All the regression models were created and validated as follows: The training data and test data were randomly selected at a ratio of 80:20. The training data were further divided into training and validation data for regression model building, 5-fold cross-validation, and hyperparameter tuning. Using the hyperparameters tuned here, the model was re-built with all the training data, and the prediction accuracy was determined using the coefficient of determination (COD), root mean square error (RMSE), and mean absolute error (MAE) for the test data. Because the data points handled in this study were fewer than 200, bias in the training data and the resulting decrease in prediction accuracy were concerns. To solve this problem, we randomly divided the training and test data into 20 combinations, created models and hyperparameter tunings for each of them. Then, we adopted a combination of training and test data that had the highest COD. To enable valid comparisons, it was necessary to use common combinations of training and test data in creating each regression model and selecting features for analysis. In addition, all the hyperparameters' random variables had to be identical. These problems were solved by executing a "random

Table 2 Extracted XRD descriptors from XRD profiles for raw materials.

Descriptor category 2 in raw material conditions	Descriptor category 3	Descriptions	Units
XRD descriptors	Average Euclidean distance of XRD peaks between experimental and reference data	Euclid_average	-
	Euclidean distance variance of XRD peaks between experimental and reference data	Euclid_variance	-
	FWHM average of XRD peaks	FWHM_average	º
	FWHM variance of XRD peaks	FWHM_variance	(º)^2
	Euclidean distance of XRD signal between experimental and reference data at no XRD peak region	Euclid_impurity	-

seed ()" command in the Python programming. Consequently, the 20 generated training and test data sets were always the same.

Results and Discussion

3.1 Feature Engineering

3.1.1 Descriptive Statistics

In our study about crystal growth, we used the conventional hypothesis-testing approach. The number of explanatory variables was 47, while the number of experimental groups employed in this study was 170. Initially, all the explanatory variables were checked for bias using histograms and box plots. The experimental parameters were the raw material species and their amounts, amounts of additives, and heating conditions. Note that particle distribution (PD), impurity species, and their amounts were uniquely determined by choosing different lot numbers of raw materials. As shown in Figure S2, the histograms of the explanatory variables show that all variables have continuous values, but these are highly biased. For example, the heating conditions are identical in many cases, and thus the bias is severe. The number of impurities in raw material is usually zero; therefore, the frequency of zero is high. The additive variables were often two discrete values, one for no use and one for use at a fixed amount. The box plots in Figure S3 show that in many cases, the data points of the explanatory variables were outside the range of the first quartile Q1 to the third quartile Q3. In conventional experiments, many variables are fixed, and the experimental conditions are tuned, resulting in this type of data.

Figure 1 shows the crystallographic evaluations of the resultant Al_2O_3 crystals. Figure 1(a) shows the SEM images of the samples obtained in different experiments. They show various distributions of crystal shapes and sizes, including crystals with a diameter of approximately 10 μm (upper left and lower left) developed in a plate-like form, and a mixture of small particles of less than 1 μm (upper right and lower right) and plate-like particles. Figure 1(b) shows the scatter plot of PD of Al_2O_3 for D_{10}, D_{50}, and D_{90}. All the images are autoscaled to see the distributions more easily. The diagonal plot shows the histogram of each PD. For all PDs, the data frequencies were sparse, but there were no major outliers. The off-diagonal plots show scatter plots of the two variables. Figure 1(c) is the PD correlation heat map that shows a correlation coefficient of more than 0.7 in the non-diagonal term, indicating that there is a strong correlation among the three variables of PD. This suggests that there is a commonality in the functional model explaining the PDs, and we decided to focus on only D_{10} in the subsequent crystal data analysis.

3.1.2 Extraction of Descriptors from XRD Features

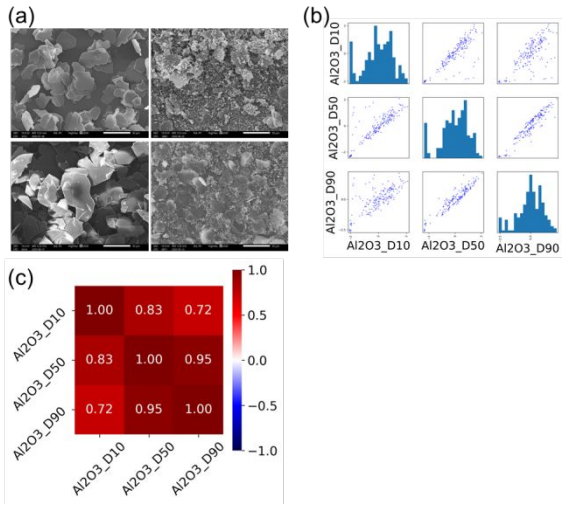

Figure 1 (a) Scanning electron microscope (SEM) images of representative flux growth of Al_2O_3 crystals, (b) scatter plots and (c) heat maps for the output data.

The XRD patterns of each raw material were analyzed to extract the feature values as shown in Figure S4. 40 types of $Al(OH)_3$ and 17 types of AlO(OH) were examined in these analyses. The peaks identified with red circles in each figure were the peaks assigned as the raw materials. As a result, we confirmed that all samples were single-phase compositions as the raw materials. Figure S5 shows the scatter plots and correlation heat maps of XRD feature values; ((a) and (c) correspond to $Al(OH)_3$, and (b) and (d) correspond to AlO(OH)). The diagonal of the scatter plot shows that the XRD feature values of each raw material are well distributed. This can also be confirmed from the box plots and histograms of the $Al(OH)_3$ and AlO(OH), as shown in Figure S6. The non-diagonal plots show that some of the variables were correlated. The heat map showed that the correlation between certain variables e.g., Euclid_average, Euclid_variance, and Euclid_impurity was greater than 0.9. Euclid_average is related to crystal orientation and crystallinity, while Euclid_variance can be interpreted as the magnitude of the bias in crystal orientation and crystallinity. In addition, Euclid_impurity increases as the peak signal-to-noise ratio of the target decreases, which is also related to the crystallinity. Therefore, the strong correlation between these three variables is easily expected. On the other hand, there was no significant correlation between the FWHM_average and FWHM_variance. This is explained by the FWHM of any peaks in a single composition being considered always similar.

Figure 2 shows the results of principal component analysis (PCA) of the XRD feature values of the raw materials. Figure 2(a) shows the contribution ratio and cumulative contribution ratio plots. The 1st and 2nd principal components accounted for more than 60% and less than 20%, indicating more than 80% of the total description information. Figure 2(b) shows a scatter plot of the first and second principal components, where the blue and red circles represent $Al(OH)_3$ and AlO(OH), respectively. There is a notable difference in the plot positions of the two raw materials and indicates that we were able to

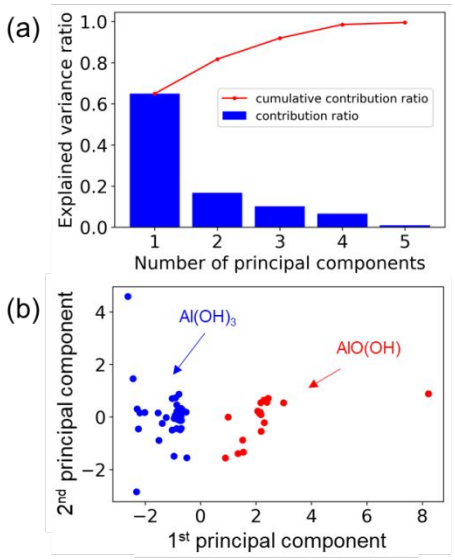

Figure 2 PCA results of XRD feature values for $Al(OH)_3$ and AlO(OH). (a) Explained variance ratio, and (b) scatter plot for 1st and 2nd principal component axes.

express the differences of raw materials using the XRD feature values. In addition, we also represented the differences among same raw materials with different lot numbers. This result suggests that feature values extracted from XRD profile has explanatory power to represent the differences in raw material samples.

3.2 Machine-Learning Analysis

The efficacy of raw material information for the prediction of flux crystal growth was investigated. The original explanatory variables (47 types) and XRD feature variables (5 types) were used to create a regression model to explain $Al_2O_3\_D_{10}$. Figure 3 shows the machine learning flow. The dataset contains all explanatory variables (dataset with all data), explanatory variables other than XRD feature values (dataset without XRD), explanatory variables other than XRD and PD feature values (dataset without XRD, PD), and explanatory variables other than XRD, PD, and impurity elements (IE) feature values (dataset without XRD, PD, IE). These variables were standardized and cleansed using correlation analysis. Then, regression models of OLS, LASSO, DT, RF, SVR, and NN for the training data for each dataset were created through a 5-fold cross validation to tune the hyperparameters. After the modeling, the obtained functions were validated using the test data. The obtained results were output as 1) accuracy and 2) factor importance.

3.2.1 Data Cleansing

During data cleansing, two of the IEs (imp8 and imp18) were deleted because their variables were all zero in the dataset. Actually, imp8 and imp18 had non-zero variables, but these variables were deleted at Al_2O_3 identification & deletion process for 185 data. Figure S7(a) shows a heat map of the correlation coefficients of the dataset with all the data.

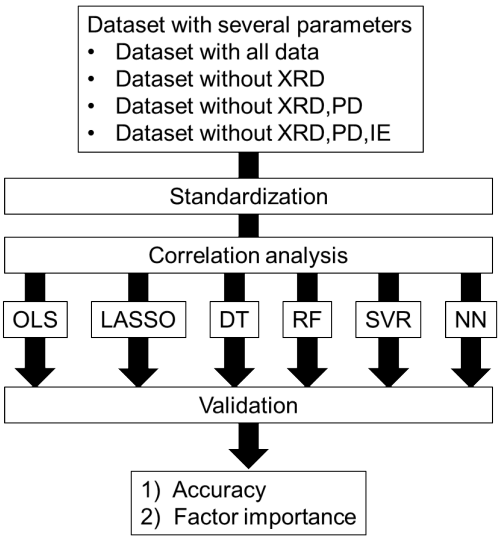

Figure 3 Machine-learning workflow in this study.

Variables with strong correlations were found in the non-diagonal lines. To eliminate multicollinearity, one of the pairs of explanatory variables with an off-diagonal component greater than 0.95 was deleted. The results are presented in Fig. S7(b). Here, Euclid_variance, Euclid_impurity, D_{50}, D_{90}, and 2nd heating time, were reduced, and the total number of explanatory variables was reduced to 45. The same cleansing operation was performed for the other datasets, and the number of explanatory variables was reduced to 42 (dataset without XRD), 41 (dataset without XRD, PD), and 25 (dataset without XRD, PD, IE).

3.2.2 Regression Model Creation and Validation

Using the cleansed data, we applied six different regression models to four different datasets. The model accuracy of each regression result for the test data is summarized in Table 3. Figure 4 shows the results with the largest COD in the 20 different CODs obtained for the randomly partitioned datasets using the regression algorithms. Figure 4(a) shows the parity plot of the predicted COD for each regression model using the dataset without XRD, PD, and IE, which is the dataset with the least number of explanatory variables. The blue triangles and the red circles represent the fitting results using training and test data, respectively. The linear regression fittings such as OLS, LASSO, and DT show relatively discrete estimations for both the training and test data and the COD was less than 0.5. The SVR was treated as a nonlinear regression model because a radial basis function (RBF) was adopted for the kernel as a result of hyperparameter tuning. For the RF and SVR nonlinear fittings, the estimated values became continuous, but the COD was still less than 0.55. Figure 4(b) shows the parity plots for each regression model created using the dataset with all data, which has the largest number of explanatory variables. The CODs of models except of OLS improved to approximately 0.650. Fitting using RF exhibits a COD of 0.741, the largest among all models and datasets except of NN. In the case of using the NN, the COD improved up to 0.767 for the dataset without XRD, while a slightly lower COD was attained for the entire dataset. This may be due to overfitting arising from a large number of explanatory

Table 3 Calculated accuracies for all regression models

COD	without XRD,PD,IE	without XRD,PD	without XRD	all data
OLS	0.462	0.489	0.477	0.487
LASSO	0.500	0.569	0.567	0.666
DT	0.416	0.457	0.486	0.636
RF	0.535	0.715	0.721	0.741
SVR	0.515	0.648	0.645	0.664
NN	0.519	0.751	0.767	0.731

RMSE	without XRD,PD,IE	without XRD,PD	without XRD	all data
OLS	0.803	0.733	0.650	0.734
LASSO	0.775	0.590	0.591	0.58
DT	0.837	0.753	0.736	0.606
RF	0.747	0.546	0.474	0.457
SVR	0.699	0.533	0.535	0.516
NN	0.76	0.445	0.529	0.568

MAE	without XRD,PD,IE	without XRD,PD	without XRD	all data
OLS	0.671	0.587	0.511	0.572
LASSO	0.648	0.445	0.458	0.433
DT	0.651	0.570	0.507	0.471
RF	0.583	0.400	0.345	0.323
SVR	0.525	0.395	0.399	0.371
NN	0.546	0.316	0.41	0.443

descriptors, and optimizing the descriptors would give higher accuracy. The improvement in accuracy was also confirmed by looking at the common decrease in the RMSE and MAE. It is notable that the estimated values of the OLS model became continuous. This is explained by the fact that the increase in the number of explanatory variables affords the function expression more flexibility. The discrete nature of the estimation in the DT model is not completely eliminated, even using the dataset with all data. Here, the maximum values of the hyperparameters, max depth and maximum leaf nodes were 10 and 10, respectively, and the parameter settings were sufficient to output 10^{10} discrete values. As a result of tuning in DT fitting, a small max depth and leaf node maximum was selected, which may indicate that the degree of discreteness was increased to suppress overtraining. Thus, we concluded precise prediction using DT is difficult with the current dataset.

Figures S8(a) and S8(b) show the results of modeling with the dataset without XRD and the dataset without XRD and PD as explanatory variables, respectively. Although there was no significant improvement in the prediction accuracy or the trend of prediction value discreteness in either case compared to Fig. 4, it was found that a reasonable level of accuracy was obtained, overall. The same regression analysis was performed using only raw material information (raw materials, PD, IE, and XRD descriptors), and COD values of approximately 0.650 were obtained by RF and SVR (Fig. S9). This result suggests that the raw material information solely had explanatory power in predicting crystal growth. However, the prediction results are discrete, indicating that a highly accurate prediction is not possible using the raw material information alone. There was a commonality in the data partitioning methods used in many cases (Nos. 3, 9, 12, and 14 were often used). This indicates that the data partitioning methods has a significant impact on prediction accuracy. In fact, among the 20 data splits, some fitting results were significantly less accurate, indicating that it is essential to consider the data splitting method when applying machine learning analysis to small data.

The test data prediction accuracy of each regression model on the four datasets is summarized in Figure 5. COD is least accurate in OLS, with no superiority or inferiority on any of the datasets. In the same linear plot, the accuracy of LASSO and DT improved by increasing the number of explanatory variables; therefore, it cannot be said that the linear representation is not suitable. OLS uses a stepwise variable reduction operation judging from the significance level, while LASSO and DT tune hyperparameters with test data for modeling. Thus, they would show large difference in variable

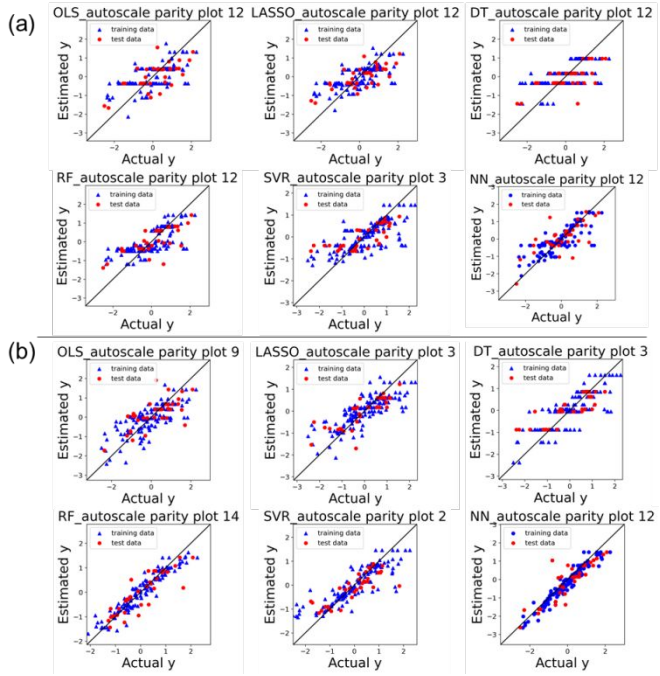

Figure 4 Parity plots of OLS, LASSO, DT, RF, SVR, and NN for (a) dataset without XRD, PD, IE parameters, and (b) dataset with all parameters.

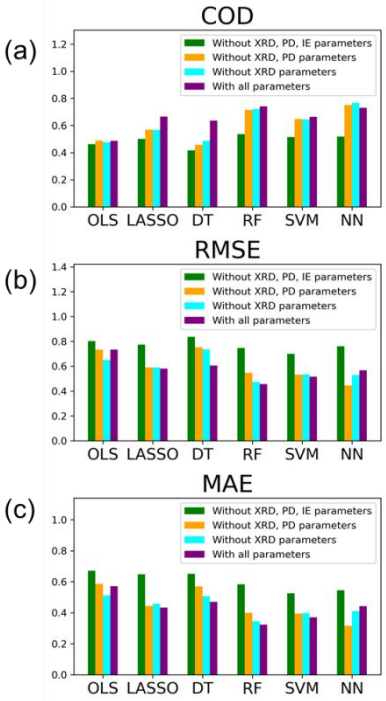

Figure 5 Accuracies at (a) COD, (b) RMSE, (c) MAE for regression models of OLS, LASSO, DT, RF, SVR, and NN for test data.

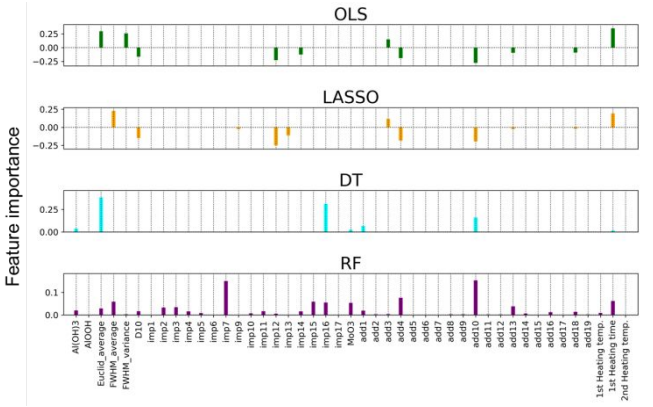

Figure 6 Summary of regression coefficients or importance (Feature importance) of OLS, LASSO, DT, RF for the dataset with all parameters.

selection. The significance level is derived from the t-test assuming that the target variable is normally distributed, but as shown in Fig. 1(b), $Al_2O_3\_D_{10}$ has slightly different characteristics from the normal distribution. As variable reduction in stepwise method in OLS rely on p-value, this modeling was possibly not effective in this study. In models except of OLS, there was a clear improvement in COD when IE and XRD parameters were added as explanatory variables. In general, when explanatory variables that are not important factors are added, overfitting occurs, and the COD for test data decreases. In this study, the COD of the test data was improved by adding raw material information, which means improvement of generalization performance. Therefore, the importance of raw material information, IE, and XRD, in explaining $Al_2O_3\_D_{10}$ is significant. In Figure 5(b) and 5(c), the RMSE and MAE were almost uniformly decreased by enriching the explanatory variables except for OLS. This negative correlation with COD also suggests adding raw material information during modeling is an efficient approach.

Figure 6 shows the factor importance when using the regression model excluding the uninterpretable SVR and NN for the dataset with all data. In linear regression models (OLS, LASSO, DT), XRD feature values, D_{10}, imp12, add 3, 4, 10, and 1st heating time have a high contribution in common. More variables were chosen in the nonlinear regression model RF. Notably, XRD feature values $D_{10,}$ imp16, add 3, 4, 10, and 1st heating times were commonly chosen in both the linear and non-linear regression models. Figure S10 shows the importance of each regression model using each partitioned dataset employed in Figures 4 and S8. Table S1 also summarizes their numerical information. It can be seen that in all regression models, when IE, PD, and XRD feature values were added to the explanatory variables, they gained a certain level of importance, and the balance of the overall importance changed significantly. Looking at the importance factor in the linear regression, it can be seen that the larger the D_{10}, the more negative the effect on $Al_2O_3\_D_{10}$. The importance factor also points out that larger XRD feature values, including FWHM, lead to increase of $Al_2O_3\_D_{10}$. They are consistent with a general knowledge that crystallographic characteristics of raw materials, including size, crystallinity, and shape are related to the reaction and dissolution rate of solutes, as well as the crystal size of the target material. The contribution of the "imp" and "add" variables was also observed, which is consistent with the finding that crystal growth is affected by impurities and additives.[29,30] From the RF analysis results in Figure 6, imp7 has a relatively high importance among the raw material features. Thus, the impurity

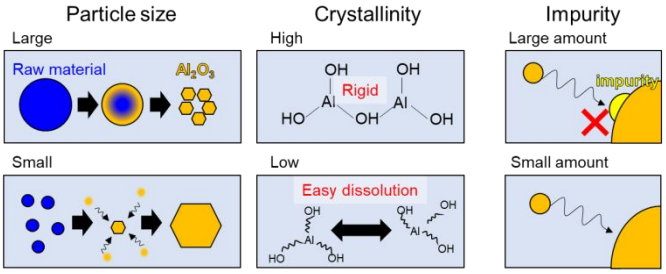

Figure 7 Suggested effects of features of raw material on crystal growth of Al_2O_3 in flux.

species and concentration in the raw material have a dominant effect on the crystal growth of Al_2O_3. Linking the present AI results to existing crystallization knowledge, we suggest that the impurity, particle size, and crystallinity of the raw material should be regarded as important parameters for controlling the crystal growth of Al_2O_3 and probably also of other inorganics.

Finally, a quantitative perspective was also considered for these values. In Figures S10(d)-1 and S10(d)-2, regression coefficients of FWHM are positive, while D10 and "imp" are negative. This means that low crystallinity, small particle size, and low impurity facilitate the growth of Al_2O_3. This trend is schematically explained in Figure 7. The raw material with a large particle size requires substantial time to dissolve in the flux, resulting in inhibition of crystal growth of Al_2O_3 due to the lower supply of solute from the solution. Decreasing the particle size may facilitate the dissolution of raw material to grow large target crystals. High crystallinity may result in rigid atomic bonding, whereas low crystallinity weakens the bonding, allowing for easy dissolution of raw material and thus enhancing the crystal growth of the target. Impurities may inhibit the adsorption of the solute from solution. Of course, it is necessary to verify these assumptions experimentally, and these effects are not common, depending on the raw material species and flux species. More studies using various materials are desirable to classify the role of these raw material parameters in the future.

Conclusions

In this study, to develop a methodology for PI-assisted flux method, we attempted to clarify the effect of raw material information on the prediction accuracy of crystal growth. Al_2O_3 crystals were grown using MoO_3 flux under 185 experimental conditions with different raw material species, raw material lot numbers, additives, and heating conditions. The crystal phases were identified by XRD, and the formation of the target material was confirmed at 170 experiments. Next, a regression model was created using raw material information and experimental conditions as explanatory variables, and the grain size distribution of the Al_2O_3 crystals as the objective variable. Six kinds of regression algorithms (OLS, LASSO, DT, RF, SVR, and NN) were executed to evaluate the prediction accuracy of the regression model depending on the raw material information. The results showed that the prediction accuracy was improved by incorporating XRD features values, raw material particle size distribution, and raw material impurity elements into the explanatory variables.

Although the raw material information was already considered important in experiments, an achievement of this study was to numerically estimate the importance by proposing a method of expressing of them. This would contribute towards making more accurate prediction models for data-driven experimental suggestion and clarification of important factors in flux crystal growth. Conversely, because the prediction accuracy is at most less than 0.8, the poor expressiveness of explanatory variables remains an issue currently. A reason may be that insufficient information is extracted by XRD. The correspondence between the XRD spectra and crystallographic features remains uncertain, and further clarification of this relationship is essential. In future, we plan to improve the accuracy of predicting flux crystal growth by enhancing the extraction of information from XRD spectra. We also plan to add new variables, such as crystal structure information, mixture state of raw materials, and their interaction terms.

Authors' Contributions

T.Y and K.T conceived the idea and designed the research. T.W and J.Y performed the crystal growth and general characterization. T.Y and K.H analyzed the data using machine learning. T.Y wrote the first draft of the paper. M.K and K.T. discussed the results and modified the paper. All of the authors discussed and commented on the paper. Therefore, the manuscript was written with the contributions of all the authors. All the authors have approved the final version of the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This research was partially supported by MEXT Program for Building Regional Innovation Ecosystems, JST Center of Innovation, Young Collaborative Research Fund, Digital Field R1WD12, and Wakasato association in Shinshu university.

Bibliographic references & notes

1 N. Dropka, M. Holena, *Crystals,* 2020, **10**, 663.
2 Y. Tsunooka, N. Kokubo, G. Hatasa, S. Harada, M. Tagawa, T. Ujihara, *CrystEngComm,* 2018, **20**, 6546-6550.
3 A. Boucetta, K. Kutsukake, T. Kojima, H. Kudo, T. Matsumoto, N. Usami, *Appl. Phys. Express,* 2019, **12**, 125503.
4 Y. Dang, L. Liu, Z. Li, *J. Cryst. Growth,* 2019, **522**, 195-203.
5 T.-S. Yao, C.-Y. Tang, M. Yang, K.-J. Zhu, D.-Y. Yan, C.-J. Yi, Z.-L. Feng, H.-C. Lei, C.-H. Li, L. Wang, L. Wang, Y.-G. Shi, Y.-J. Sun, H. Ding, *Chin. Phys. Lett*., 2019, **36**, 068101.
6 Y. Oaki, *Bull. Chem. Soc. Jpn*., 2017, **90**, 776-788.
7 J. Yang, D. Li, H. Wang, X. Wang, X. Yang, L. Lu, *Mater. Lett.,* 2001, **50**, 230-234.

8 K. Yanagisawa, J. Ovenstone, *J. Phys. Chem. B,* 1999, **103**, 7781-7787.
9 W. Huang, J. Yang, X. Meng, Y. Cheng, C. Wang, B. Zou, Z. Khan, Z. Wang, X. Cao, *Chem. Eng. J.,* 2011, **168**, 1360-1368.
10 N. Zaitseva, L. Carman, I. Smolsky, R. Torres, M. Yan, *J. Cryst. Growth,* 1999, **204**, 512-524.
11 P. Rajesh, U. C. In, P. Manyum, P. Ramasamy, *Materials Res. Bull*., 2014, **59**, 431-434.
12 M. Inoue, I. Hirasawa, *J. Cryst. Growth,* 2013, **380**, 169-175.
13 Y. Suzuki, H. Hino, T. Hawai, K. Saito, M. Kotsugi, K. Ono, *Sci. rep*., 2020, **10**, 21790.
14 W. B. Park, J. Chung, J. Jung, K. Sohn, S. P. Singh, M. Pyo, N. Shin, K.-S. Sohn, *IUCrJ,* 2017, **4**, 486-494.
15 F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. T. P. Hartono, S. Ramasamy, B. L. DeCost, S. I. P. Tian, G. Romano, A. G. Kusne, T. Buonassisi, *NPJ Comput. Mater.,* 2019, **5**, 60.
16 J.-W. Lee, W. B. Park, J. H. Lee, S. P. Singh, K.-S. Sohn, *Nat. Commun.,* 2020, **11**, 86.
17 K. Kaufmann, C. Zhu, A. S. Rosengarten, D. Maryanovsky, T. J. Harrington, E. Marin, K. S. Vecchio, *Science* 2020, **367**, 564-568.
18 A. B. Chase**,** J. A. Osmer, *J. Am. Ceram. Soc*., 1970, **53**, 343-345.
19 S. Oishi, K. Mochizuki, S. Hirano, *J. Ceram. Soc. Japan,* 1994, **102**, 502-504.
20 K. Teshima, A. Takano, T. Suzuki, S. Oishi, *Chem. Lett*., 2005, **34**, 1620-1621.
21 K. Teshima, K. Matsumoto, H. Kondo, T. Suzuki, S. Oishi, *J. Ceram. Soc. Japan,* 2007, **115**, 379-382.
22 S. Oishi, K. Teshima, H. Kondo, *J. Am. Chem. Soc*., 2004, **126**, 4768–4769.
23 K. Teshima, H. Kondo, S. Oishi, *Bull. Chem. Soc. Japan,* 2005, **78**, 1259-1262.
24 J. Yuan, H. Kinoshita, *U.S. Patent,* 9604852B2, 2013.
25 S. Hashimoto, A. Yamaguchi, *J. Eur. Ceram. Soc*., 1999, **19**, 335-339.
26 S. Hashimoto, S. Horita, Y. Ito, H. Hirano, S. Honda, Y. Iwamoto, *J. Eur. Ceram. Soc*., 2010, **30**, 635-639.
27 C. W. Park, D. Y. Yoon, *J. Am. Ceram. Soc*., 2000, **83**, 2605-2609.
28 J. Yuan, H. Kinoshita, *U.S. Patent*, 10808131B2, 2017.
29 T. Rabizadeh, T. M. Stawski, D. J. Morgan, C. L. Peacock, L. G. Benning, Crystals. *Cryst. Growth Des*. 2017, **17**, 582-589.
30 S. Dobberschütz, M. R. Nielsen, K. K. Sand, R. Civioc, N. Bovet, S. L. S. Stipp, M. P. Andersson, *Nat. Commun.,* 2018, **9**, 1578.