Nanoscale Advances

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Nanoscale Adv.*, 2025, **7**, 4620

Data-scientific validation of prediction models for the controlled syntheses of exfoliated nanosheets†

Yuka Kitamura,[a] Yuki Namiuchi,[b] Hiroaki Imai,[a] Yasuhiko Igarashi *[b] and Yuya Oaki *[a]

Exfoliated nanosheets have attracted considerable interest as two-dimensional (2D) building blocks. In general, the yield, size, and size distribution of the exfoliated nanosheets cannot be easily controlled or predicted because of the complexity in the processes. Our group studied the prediction models of the yield, size, and size distribution based on the small experimental data available. Sparse modeling for small data (SpM-S) combining machine learning (ML) and chemical insight was used for the construction of predictors. In SpM-S, the weight diagram visualizing the significance of explanatory variables plays an important role in variable selection to construct the models. However, the processes of variable selection were not validated in a data-scientific manner. In the present work, the significance of data size, visualization method, and chemical insight for variable selection was studied to validate the processes of model construction. The data size had a lower limit to extract appropriate descriptors. The weight diagram had an appropriate visualizing range for variable selection. Chemical insight as domain knowledge supplemented the limitation caused by the data size. These studies indicated that SpM-S can be applied to construct predictors, straightforward linear regression models, for the controlled syntheses of other 2D materials, even based on small data.

Received 4th March 2025
Accepted 4th June 2025

DOI: 10.1039/d5na00215j

rsc.li/nanoscale-advances

1. Introduction

Liquid-phase exfoliation is a general method used to obtain 2D materials, including monolayers and few-layers.[1–9] Various precursor layered materials can be exfoliated into nanosheets. However, the exfoliation behavior cannot be easily controlled because of the complex and random downsizing processes in both the lateral and thickness directions. For example, the yield, size, and size distribution of the exfoliated nanosheets are not easily controlled by specific parameters based only on professional experience. In recent years, data-scientific approaches have been applied to the field of 2D materials for their design, synthesis, and characterization.[10–14] Our group has focused on the construction of prediction models to control the yield, size, and size distribution of surface-modified nanosheets exfoliated from precursor layered composites (Fig. 1a–c).[4,10,15–18] The precursor layered composites are typically synthesized by the intercalation of the guest organic molecules in the interlayer space of host layered transition-metal oxides (Fig. 1a). The surface-modified nanosheets are then obtained through

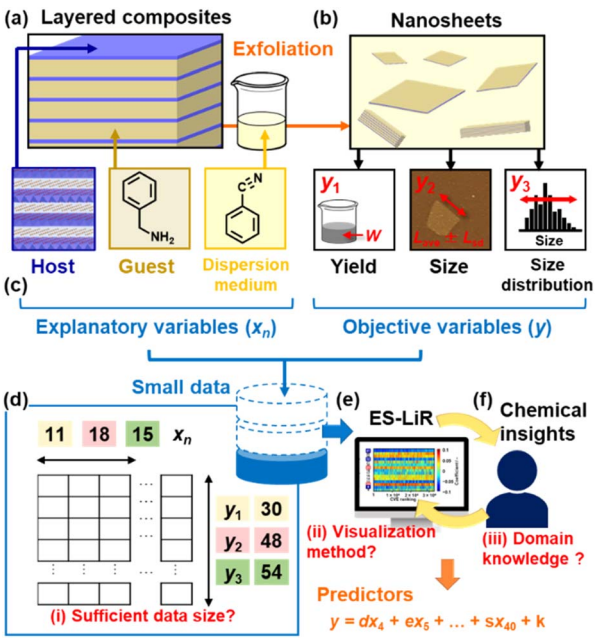

Fig. 1 Overview of the small-data-driven exfoliation experiments. (a) Precursor layered composites of inorganic hosts and organic guests and their exfoliation in organic dispersion media. (b) Yield (y_1), lateral size (y_2), and size distribution (y_3) of the surface-functionalized nanosheets. (c) Explanatory variables (x_n) and objective variables (y: y_1, y_2, y_3). (d) Small datasets and their contents. (e and f) Variable selection using ES-LiR (e) and our chemical insight (f) as ML and domain knowledge, respectively, for the construction of predictors.

[a]*Department of Applied Chemistry, Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan. E-mail: oakiyuya@applc.keio.ac.jp*

[b]*Institute of Engineering, Information and Systems, University of Tsukuba, 1-1-1 Tennodai, Tsukuba 305-8573, Japan. E-mail: igayasu1219@cs.tsukuba.ac.jp*

† Electronic supplementary information (ESI) available: Datasets, weight diagrams. See DOI: https://doi.org/10.1039/d5na00215j

exfoliation by the dispersion of the layered composites in organic dispersion media (Fig. 1b). Although the exfoliation behavior could be changed by varying the combinations of the host, guest, and medium, their effects on the yield, size, and size distribution of the nanosheets were unclear. In recent years, prediction models for the control of these parameters have been constructed by combining machine learning and our chemical insight on small experimental data.[4,10,15–18] Moreover, controlled syntheses have been achieved using the predictors in a limited number of experiments. However, the model construction processes have not yet been fully studied in a data-scientific manner. If the model construction processes could be validated, then similar predictors could be constructed by SpM-S for various other 2D materials.

Data-driven approaches have been used in a broad range of chemistry and materials science areas.[19–27] For instance, the combination of big data, machine learning (ML), and a robotic system has been studied to develop fully automated AI chemists.[28–32] These AI-oriented methods are supported by the availability of a sufficient size of data for the ML. However, a sufficient size of data is not always available for all experimental systems. For example, big data is not efficiently collected from conventional experimental works including batch processes. Specific methods are thus required to apply ML to small data.

ML for small data has been increasingly studied in recent years.[33–45] Specific approaches, such as transfer learning, have been developed for the use of small data.[33–45] However, the interpretability and generalizability are lower for modeling based on complex modeling algorithms. Recent reports have indicated the significance of domain knowledge and the use of simple regression models.[43–45] Our group has focused on sparse modeling (SpM), a method for describing whole high-dimensional data by a small number of significant descriptors.[46–48] SpM has already been applied in a variety of fields, such as image compression and materials science.[15–18,49–54] We have studied SpM for small data (SpM-S) combining ML and domain knowledge.[10,45] The method was applied to controlled the synthesis of nanosheets and the exploration of electrode active materials based on small data.[15–18,51–54] In SpM-S, the descriptors are extracted from a small training dataset using a ML algorithm, and an exhaustive search with linear regression (ES-LiR), as mentioned later (Fig. 1c–e). Then, the descriptors are further selected based on our domain knowledge as chemists (Fig. 1f). A straightforward linear regression model is then constructed using the selected descriptors. In our previous work,[45] the prediction results of SpM-S combining linear regression and our chemical insight were compared with those obtained from other linear and nonlinear algorithms, such as least absolute shrinkage and selection operate (LASSO) and neural network regression (NN-R), in terms of the accuracy, interpretability, and generalizability, especially for small data. Although nonlinear algorithms, such as NN-R, generally exhibit high expressive power, they tend to overfit small chemical experimental datasets because of the insufficient generalizability.[45] However, the processes remain unclear with problems persisting regarding the variable selection, one of the significant steps for modeling (problems (i)–(iii) in Fig. 1d–f) regarding the required data size (problem (i)), visualization method of the weight diagram (problem (ii)), and the significance of domain knowledge (problem (iii)). The present study aimed to solve these problems to improve the understanding of SpM-S. The results indicated that similar predictors could be constructed by SpM-S based on small experimental data for various other 2D materials.

2. Results and discussion

2.1. Prediction models for exfoliated nanosheets

The prediction models of the yield, size, and size distribution were constructed using the small training datasets I–III in our previous works, respectively (Fig. 1 and Tables S1–S3 in the ESI†).[16–18,45] The objective variables (y) were the yield (y_1), lateral size (y_2), and lateral-size distribution (y_3) of the nanosheets (Fig. 1a and b). The yield ($y_1 = 100 \times W/W_0$) was calculated from the weight of the collected nanosheets with the filtration (W) and that of the precursor layered materials (W_0).[16] The average lateral size (L_{ave}) and its standard deviation (L_{sd}) were measured by dynamic light scattering as for a high-throughput method. The lateral size ($y_2 = R_L$) is defined as its reduction rate $R_L = L_{ave}/L_0$,[17] where L_0 is the lateral size of the precursor layered materials. The size distribution ($y_3 = L_{CV}$), polydispersity, is represented by the coefficient of variation about the lateral size ($L_{CV} = L_{sd}/L_{ave}$).[18]

The explanatory variables (x_n: n = 1–41), such as the physicochemical parameters of the guests and media, were the related physicochemical parameters selected by our chemical insights (Table 1). In the total 41x_n, the selected x_n were used as the potential descriptors for y_1–y_3. The datasets contained the following numbers of y and x_n (Table 1): 30y_1 and 11x_n (n = 2, 4, 5, 8, 10, 14, 16–18, 36, 40) (dataset I), 48y_2 and 18x_n (n = 1, 3–5, 14–21, 30–32, 34, 36, 40) (dataset II), 54y_3 and 15x_n (n = 4, 8, 10, 13, 14, 16–18, 21, 30–32, 36, 40, 41) (dataset III) (Fig. 1c and d). In our previous works, the descriptors were extracted from x_n by SpM using ES-LiR (Fig. 1d and e). Then, the descriptors were further selected with the assistance of our chemical insights (Fig. 1e and f). The linear regression models eqn (1)–(3) were constructed using the selected two to eight x_n.[16–18]

$$y_1 = 35.00x_3 - 32.33x_5 + 34.07 \quad (1)$$

$$y_2 = -0.159x_3 - 0.096x_4 + 0.257x_7 - 0.017x_8 - 0.018x_{10} + 0.028x_{13} - 0.050x_{14} + 0.061x_{18} + 0.267 \quad (2)$$

$$y_3 = -0.0599x_7 + 0.0802x_9 + 0.0699x_{20} - 0.0681x_{28} - 0.0623x_{37} + 0.266 \quad (3)$$

As the coefficients are converted to the normalized frequency distribution with mean 0 and standard deviation 1 for the variables in each model, the weight of the contribution is represented by the coefficients. In the present work, the processes of the variable selection were studied to validate the models themselves and their construction processes.

2.2. Effects of the data size on the extraction of the descriptors

The extractability of the descriptors generally depends on the data size. In the present work, the descriptors were extracted

Table 1 List of x_n (n = 1–41) for y_1, y_2, and y_3 (ref. 45)

n	Parameters	x_n for
Dispersion media		
1	Molecular weight	y_1, y_2, y_3
2	Molecular length[b]	y_1
3	Melting point[a]	y_1, y_2, y_3
4	Boiling point[a]	y_1, y_2, y_3
5	Density[a]	y_1, y_2, y_3
6	Relative permittivity[a]	y_1, y_2, y_3
7	Vapor pressure[a]	y_1, y_2, y_3
8	Viscosity[a]	y_1, y_2, y_3
9	Refractive index[a]	y_1, y_2, y_3
10	Surface tension[a]	y_1, y_2, y_3
11	Heat capacity[b]	y_1, y_2, y_3
12	Entropy[b]	y_1, y_2, y_3
13	Enthalpy[b]	y_1, y_2, y_3
14	Dipole moment[b]	y_1, y_2, y_3
15	Polarizability[b]	y_1, y_2, y_3
16	HSP-dispersion[b]	y_1, y_2, y_3
17	HSP-polarity[b]	y_1, y_2, y_3
18	HSP-hydrogen bonding[b]	y_1, y_2, y_3
Guest molecules		
19	Molecular weight	y_1, y_2, y_3
20	Polarizability[b]	y_1, y_2, y_3
21	Dipole moment[b]	y_1, y_2, y_3
22	Heat capacity[b]	y_1, y_2, y_3
23	Entropy[b]	y_1, y_2, y_3
24	Enthalpy[b]	y_1, y_2, y_3
25	Molecular length[b]	y_1
26	Layer distance[c]	y_1, y_2, y_3
27	Layer distance expansion[c]	y_3
28	Composition (x)[c]	y_1, y_2
29	Interlayer density[c]	y_1, y_2
30	HSP-dispersion terms[b]	y_1, y_2, y_3
31	HSP-polarity terms[b]	y_1, y_2, y_3
32	HSP-hydrogen bonding terms[b]	y_1, y_2, y_3
Guest-medium combinations		
33	Δ polarizability ($=x_{15} - x_{20}$)[b]	y_3
34	Δ polarizability ($=\lvert x_{33}\rvert$)[b]	y_1, y_2, y_3
35	Δ dipole moment ($=x_{14} - x_{21}$)[b]	y_3
36	Δ dipole moment ($=\lvert x_{35}\rvert$)[b]	y_1, y_2, y_3
37	Product of dipole moment ($=x_{14} \times x_{21}$)[b]	y_3
38	Δ heat capacity ($=x_{11} - x_{22}$)[b]	y_3
39	Δ heat capacity ($=\lvert x_{38}\rvert$)[b]	y_1, y_2, y_3
40	HSP distance[b]	y_1, y_2, y_3
Host		
41	Bulk size[c]	y_3

[a] Literature data. [b] Calculation data. [c] Experimental data.

with reducing the data size in datasets I–III to study whether the data size was sufficient for the extraction of the descriptors (Fig. 2). The detailed procedure is described in the ESI.† The number of y (N) was decreased step-by-step (Fig. 2a and Tables S1–S3 in the ESI†). For example, the original $30y_1$ was reduced to $25y_1$ with the random subtraction of five y_1. Then, reduced datasets with $20y_1$ were prepared with a further subtraction of $5y_1$. The five subtracting data items were randomly selected and six different datasets were prepared at each N. In this manner, reduced datasets were prepared for each of y_1, y_2, and y_3.

The extractability of the descriptors was then studied using the reduced datasets. Weight diagrams were prepared by ES-LiR. Linear regression models were prepared by all the possible combinations of x_n (n = 1, 2, …, j), *i.e.*, total $2^j - 1$ combinations, on each dataset with five-fold cross-validation (Fig. 2b). As pointed out, Hastie *et al.* suggested "ten-fold cross-validation achieves an acceptable trade-off between bias and variance,"[55] and this has since become standard practice.[56] However, both 5- and 10-fold cross-validation are generally recognized as appropriate choices due to their superior stability compared to leave-one-out cross-validation (LOOCV). Despite its theoretical appeal, LOOCV exhibits high variance in performance estimates and is thus less reliable for model selection.[57] In our study, five-fold cross-validation was used in terms of its computational efficiency and as standard practice. After the models were sorted in ascending order of cross-validation error (CVE), *i.e.,* CVE ranking, the values of the coefficients for each regression model were visualized in the weight diagram (Fig. 2c). Conventional ML algorithms require tuning the hyperparameters to optimize the models.[58,59] Whereas, as ES-LiR just prepares all the possible linear regression models, the modeling method has no hyperparameters to be tuned compared with other ML algorithms. The contribution of each x_n was color-coded by the magnitude of the coefficients with their positive and negative values. The more deeply colored x_n with warmer and cooler colors have potential as more significantly contributed descriptors with the positive and negative correlations, respectively. The more densely colored x_n correspond to the more frequently used descriptors, implying a significant contribution to y. Weight diagrams were prepared for all the reduced datasets (Fig. S1–S3 in the ESI†). Based on the weight diagrams, we extracted x_n as the descriptors with reference to the deepness and density of the color (Fig. 2c and d). Here x_n in the already constructed models eqn (1)–(3) were assumed to be the true ones (Fig. 2e). If the visually extracted x_n from the weight diagram was found in the already constructed models, the extracted x_n here could be regarded as the correct ones. In contrast, the extracted x_n that were not found in the constructed models were regarded as incorrect ones (Fig. 2d and e). After the weight diagrams were prepared for the six different reduced datasets at each data size (N) (Fig. S1 in the ESI†), the numbers of correctly and incorrectly extracted x_n (n_c and n_i, respectively) were counted and are summarized in Fig. 2f–h. The mean and standard deviation of n_c and n_i were calculated for the six different datasets.

When N was reduced, n_c decreased and n_i increased (Fig. 2f–h). Here the threshold N to extract the correct descriptors (N_{min}) is defined as follows: the average n_i is less than two and n_c is more than 80% of the true n_c before the data reduction. The threshold data size to extract the correct x_n, N_{min}, was 20 for y_1 (the original data size: $N_0 = 30$), 45 for y_2 ($N_0 = 48$), and 45 for y_3 ($N_0 = 54$) (red-colored areas in Fig. 2f–h). These results indicate that the correct x_n can be extracted from the weight diagrams based on datasets with $N_{min} < N$. N_{min} can be regarded as the minimum required data size to extract the correct descriptors

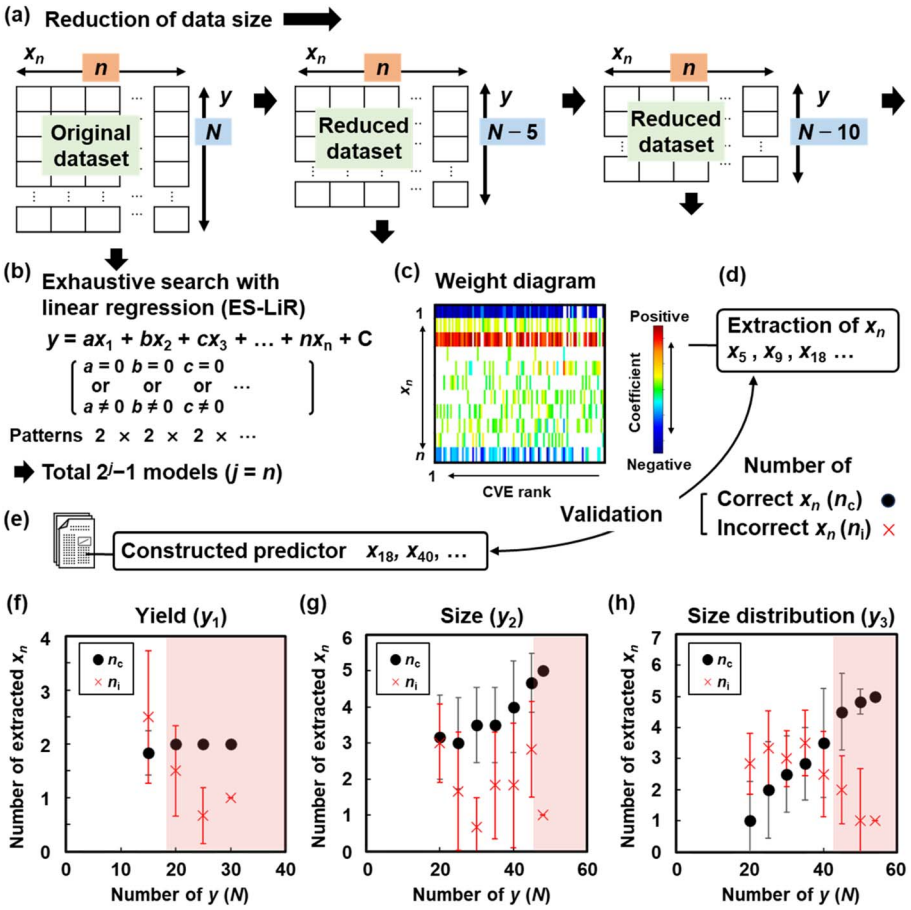

Fig. 2 Effect of data size on variable selection. (a) Reduction of the data size, the number of y (N). (b) Scheme of ES-LiR. (c) Weight diagram visually representing the contribution of each x_n. (d) Extraction of x_n from the weight diagrams with reducing N as shown in panels (a–c). (e) Counting n_c and n_i based on comparison of the extracted x_n with that in the already constructed predictors in our previous works as the correct models. (f–h) Summary of n_c and n_i with reducing N for y_1 (f), y_2 (g), and y_3 (h).

from the weight diagram. As $N_{min} < N_0$ was achieved for y_1, y_2, and y_3, the original datasets already had a sufficient data size for the model construction. Therefore, the generalizable x_n could be extracted from the weight diagrams even based on the small dataset at $N_{min} < N$. These results support the prediction models for the yield, size, and lateral size, and so eqn (1)–(3) were constructed with a sufficient size of data. Moreover, this data-reduction method can be applied to validate the sufficiency of the data size for the variable selection in small data.

2.3. Visualization method of weight diagrams

As discussed in Section 2.1, in SpM-S, the coefficients of x_n are visualized in the weight diagram based on the CVE values (Fig. 3a–c). In contrast, the other ML algorithms for SpM, such as LASSO and minimax concave penalty and penalized linear unbiased selection (MCP), only focus on a certain model with the smallest CVE values. Such algorithms raise concerns about the false extraction of descriptors, particularly in small data. ES-LiR visualizes a large number of models using the weight diagram in the ascending order of the CVE values. However, the preparation method of the weight diagram is somewhat arbitrary; in particular, the visualizing range of the CVE rank in the horizontal axis is arbitrary (Fig. 3b and c). If the appearance of the weight diagram is changed depending on the visualizing range, then different x_n can be extracted. Here the range displaying CVE rank in the horizontal axis was changed to study the effects on the appearance of the weight diagram for extractability of the descriptors (Fig. 3b and c).

The relationship between the CVE rank and CVE value was determined to visualize the increasing trend of the CVE value (Fig. 3b). Then, the weight diagrams were prepared within the different CVE ranks as the thresholds (Fig. 3c). The CVE values gradually increased with lowering the rank and then jumped near the bottom (Fig. 3d, h and l). As the datasets contained 11, 18, and 15 x_n for y_1, y_2, and y_3, respectively, the total number of exhaustively constructed models ($2^j - 1$ combinations) was 2.0×10^3 for y_1, 2.6×10^5 for y_2, and 3.3×10^4 for y_3. Whereas the correct x_n ($n = 18, 40$) were clearly visible in the weight diagram within the CVE rank 1.0×10^2 for y_1 (Fig. 3e), the weight diagrams became unclear in the ranks 1.0×10^3 and 2.0×10^3 (Fig. 3f and g). The correct x_n ($n = 18, 40$) could not be extracted from these unclear weight diagrams. Clear weight diagrams were observed in the ranks 1.0×10^4 for y_2 and 1.0×10^3 for y_3 (Fig. 3i, j and m). The visibility was lowered in the weight

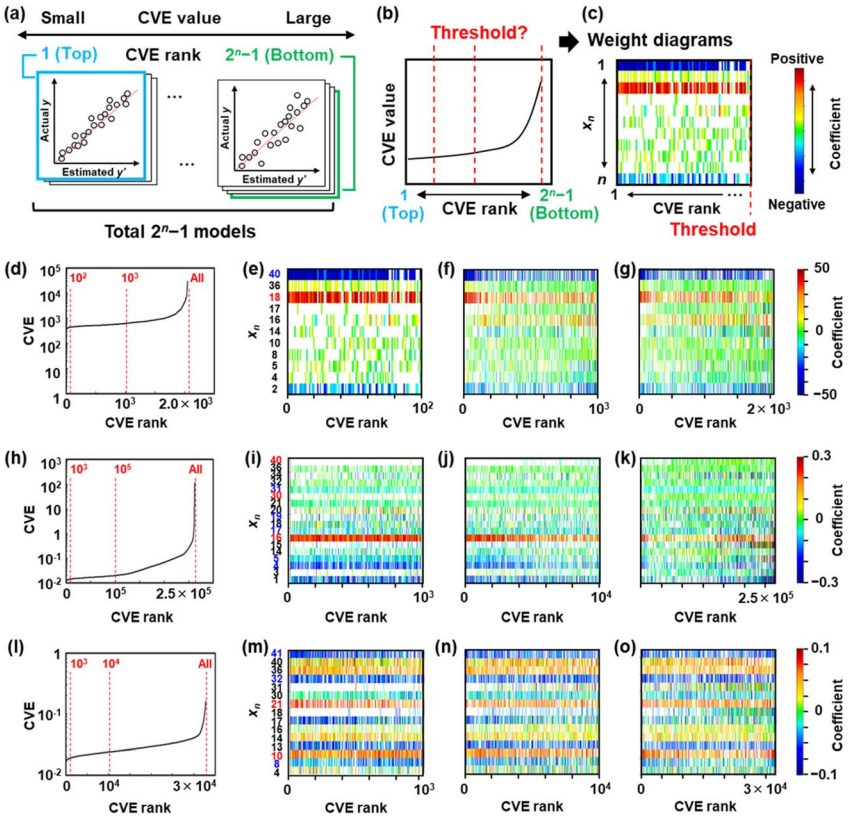

Fig. 3 Visualization method of weight diagrams. (a) All possible linear regression models, $2^n - 1$ patterns, sorted in an ascending order of the CVE rank. (b) Relationship between the CVE rank and value. (c) Preparation of the weight diagrams with the different threshold of the CVE rank. (d, h and l) Relationship between the CVE rank and values for y_1 (d), y_2 (h), y_3 (l). (e–g) Weight diagrams of y_1 within the top CVE rank 1.0×10^2 (e), 1.0×10^3 (f), and 2.0×10^3 (g). (i–k) Weight diagrams of y_2 within the top CVE rank 1.0×10^3 (i), 1.0×10^4 (j), and 2.6×10^5 (k). (m–o) Weight diagrams of y_3 within the top CVE rank 1.0×10^3 (m), 1.0×10^4 (n), and 3.3×10^4 (o).

diagrams in the CVE ranks lower than 1.0×10^5 for y_2 and 1.0×10^4 for y_3 (Fig. 3k, n and o).

Based on these results, it could be seen that the visibility of the weight diagrams and extractability of the descriptors were changed by the range of the CVE rank. The weight diagrams within the CVE ranks about the top 10%, namely 10^2 for y_1, 10^4 for y_2, and 10^3 for y_3, allowed a clear extraction of the descriptors. The CVE ranks achieving one standard error rule were calculated to be 2.3×10^2 for y_1, 4.6×10^4 for y_2 and 2.0×10^3 for y_3. In the present work, the top 10% of the CVE rank was coincident with one standard error rule. Here, one standard error rule was used to estimate the range of the CVE ranks for visualization. The scheme means that all models having a CVE within one standard deviation of the minimum CVE were considered, resulting in the selection of approximately the top 10% of the CVE rankings. However, this coincide was not necessary. In general, the one standard error rule is used to optimize the regularization parameter (λ) in ML.[60] When a larger penalty term is set, the one standard error rule is used to optimize λ instead of the minimum CVE value. These facts imply that a similar scheme can be applied to estimate the threshold for distinctly increasing the CVE by one standard error rule. In contrast, the visibility becomes unclear when the range was expanded to the CVE rank over the top 50%. Whereas the correct descriptors could be extracted from the clear weight diagram, the unclear weight diagram caused an extraction of the wrong descriptors and an oversight of the correct ones. As clear weight diagrams with the top 10% rank were used in our previous works,[16–18] appropriate descriptors were extracted for the construction of the models in eqn (1)–(3).

2.4. Effects of the domain knowledge on the variable selection

As demonstrated in Sections 2.1 and 2.2, the appearance of the weight diagrams, such as the color density and intensity, easily varied by the data size and noise resulting from the small data. Therefore, our chemical insight as experimental scientists was used for the further selection and rejection of the descriptors in addition to the weight diagrams. Such domain knowledge facilitates robust modeling based on small data. As a reference, here an exhaustive search with Bayesian model averaging (ES-BMA) was used to extract the descriptors without the assistance of our domain knowledge. The variable selection using ES-BMA was then compared with that using ES-LiR combined with our domain knowledge.

In ES-BMA, the probability of a descriptor (p) being the significant descriptor was 0.5 for all x_n at the initial state (Fig. 4a). The descriptors were not extracted because $p = 0.5$.

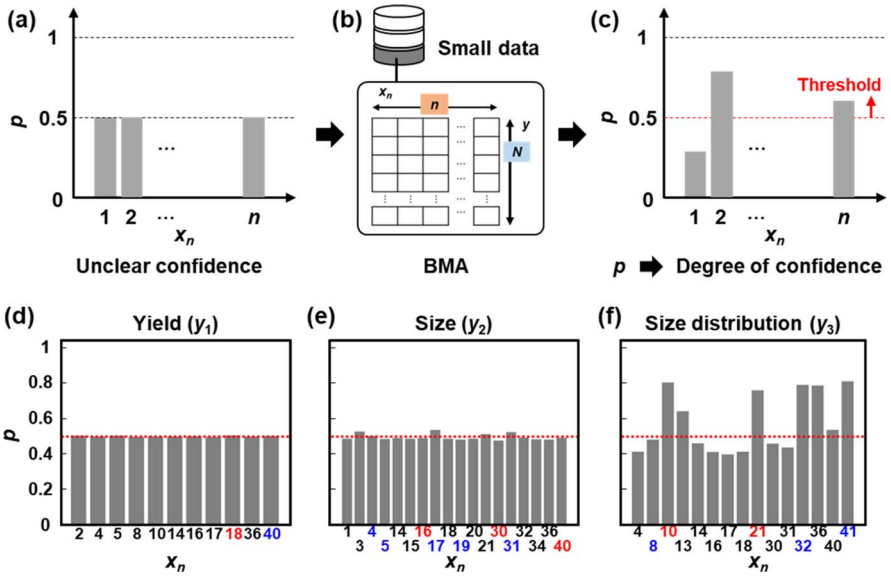

Fig. 4 Variable selection using ES-BMA. (a) Probability (p) before ES-BMA. (b) BMA based on small data. (c) p after ES-BMA. (d) p of each x_n for y_1 (d), y_2 (e), y_3 (f) after ES-BMA.

The probability p of each x_n was calculated using the prediction accuracy and coefficient of the $2^n - 1$ models prepared by ES-LiR (Fig. 4b and c). In ES-BMA, we consider the uncertainty for all 2^n combinations of variables and introduce a method for quantitatively evaluating the confidence level of variable selection using a weighted average of the model posterior probabilities, which is called Bayesian model averaging (BMA) (Fig. 4b).[61] This method enables evaluating the confidence level of the variable selection and quantifying the importance evaluation of the features, whereas the processes quantitatively depend on the visibility of the weight diagram for ES-LiR. Furthermore, this approach quantitatively assesses the plausibility of the descriptors under the assumption of uniform prior knowledge without relying on the expertise of chemists. The summation over all combinations of indicator vectors can be calculated using the result of the exhaustive search, which is called ES-BMA.[61]

Fig. 4d–f shows the p of each x_n. The descriptors for y_1 and y_2 were not extracted from the probability of each x_n because p was almost 0.5 (Fig. 4d and e). On the other hand, x_{10}, x_{13}, x_{21}, x_{32}, x_{36}, x_{40}, and x_{41} for y_3 showed $p > 0.5$ (Fig. 4f). Four descriptors x_{10}, x_{21}, x_{32}, and x_{41} of five correct x_n in eqn (3) were extractable by the p value based on ES-BMA. These results imply that the data size was insufficient to extract the true descriptors only using ES-BMA without the domain knowledge. The combination of ES-LiR and domain knowledge facilitated the extraction of the descriptors. In SpM-S, the domain knowledge contributed to being able to extract the descriptors and construct the models.

In SpM-S, professional experience and chemical insights, as domain knowledge, are mainly used in the processes of variable selection based on the weight diagram. Although the weight diagram indicated the strong contribution of certain variables, some variables were not used as the descriptors for modeling based on our chemical insight. For example, this scheme provides a more accurate prediction model for the specific capacity of organic anode active materials of lithium-ion batteries.[53] On the other hand, some chemically significant descriptors were not extractable only from the weight diagram. In such a case, the descriptors were manually added for the model. For example, the yield prediction model was constructed by this scheme.[16] However, it is not easy to quantify the physical significance, not just the correlation, of the variables based on chemical insights. In ES-BMA, such physical meaning is represented by the probability value (p). However, p is not estimated from the small data, as shown in Fig. 4d and e. The development of a new quantitative method is required to extract and select the more significant descriptors quantitatively.

2.5. Advantages of SpM-S compared with other algorithms

In linear regression models, ES-LiR contributes to providing accurate models using weight diagrams compared with other algorithms, such as LASSO with variable selection and multiple linear regression (MLR) without variable selection. In ES-LiR, linear regression models are exhaustively prepared using all the possible combinations of x_n. The potential models are selected based on weight diagrams visualizing the contribution of each coefficient. As MLR-based models include the irrelevant x_n, there may be an overfitting that causes a lowering of the prediction accuracy and generalizability. In the present study, the variable selection problem involved a relatively small number of variables. It is feasible to arduously search all the possible combinations and directly identify the optimal model.[47] As other algorithms with variable selection construct specific models with certain CVE values based on approximations, the appropriate variables may not be extracted particularly in noisy and small training data. In our previous works,[45,52,53] prediction models were constructed by LASSO to

compare the prediction accuracy, which was calculated by five-fold CVE. The models using LASSO generally showed larger RMSE values to the test data compared with those using SpM-S, even though smaller RMSE values were achieved from some training data. The other modeling techniques use certain approximations to reduce the calculation cost. On the other hand, ES-LiR is applicable to small data at a realistic calculation cost. More accurate descriptors can be selected with our chemical insights from all the possible regression models.

The accuracy, generalizability, and interpretability of the sparse linear models based on small data were compared with those of other nonlinear algorithms in our previous works.[45] The results imply that nonlinear models have concerns about overtraining linked to the training data and a lowering of the generalizability, particularly in the case of small data. In such a case, linear models are preferable to describe the whole trend of the data. We recognize the importance of frameworks, such as sure independent screening and sparsifying operator (SISSO),[62] which integrates sure independent screening (SIS) with LASSO-based variable selection to efficiently manage ultra-high-dimensional descriptor spaces. However, in our current study, the dimensionality of the descriptor space was limited to several tens of dimensions, enabling an exhaustive search approach rather than requiring dimensionality reduction using SIS. Furthermore, as demonstrated in a recent work,[63] a Monte Carlo-based approximate exhaustive search method could be employed for moderately high-dimensional scenarios. Our ongoing research efforts are directed toward integrating SIS and exhaustive search strategies to enhance the descriptor selection efficiency and effectiveness, particularly in high-dimensional and correlated descriptor spaces.

3. Conclusions

Linear regression prediction models for the yield, size, and size distribution of exfoliated nanosheets were constructed by SpM-S combining ES-LiR and domain knowledge on small data. The present work validated the model construction method and process, such as the significance of the data size, visualization method, and use of chemical insights for the variable selection. Weight diagrams were constructed that visualized the significance of the variables by color. Then, the significant descriptors were selected with the assistance of our chemical insight. The appearance of the weight diagram was changed with reducing the data size. The data size had a specific lower limit to extract the same appropriate descriptors. This method can be widely applied to validate whether the data size is sufficient or not. Whereas conventional ML algorithms with variable selection focus on a certain model with the minimum CVE value, the weight diagram of ES-LiR overviews a large number of models in ascending order of the CVE values. The visualization range of the CVE values had an effect on the appearance of the weight diagram leading to the extraction of the descriptors. A clear weight diagram suitable for the variable selection was obtained within about top 10% of the CVE rank, which was consistent with the one standard error rule. When the variables were selected using the probability of ES-BMA without the assistance of our domain knowledge, the descriptors could not be extracted only from the probability value. The fact implies that the domain knowledge could be effectively used for the variable selection to supplement the deficiency of data. The present study supports the validity of the prediction models for the yield, size, and size distribution of exfoliated nanosheets reported in our previous works. Moreover, SpM-S combining ES-LiR and chemical insight is a suitable method for small data in a variety of fields. In the present work, the models were constructed based on the data about the exfoliation of layered inorganic–organic composites into surface-modified nanosheets, as shown in Fig. 1a and b. The yield prediction model has also been applied to the exfoliation of other layered compounds, such as graphite and layered organic polymers.[64] As the chemical features of the layer surface and dispersion media are used as the descriptors, the models can be applied to other types of the layered materials.

Data availability

The data supporting this article have been included as part of the ESI.†

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was partially supported by JST PRESTO (Y. O., JPMJPR16N2 and Y. I. JPMJPR17N2), JST CREST (Y. I., JPMJCR21O1), JSPS-KAKENHI (Y. O. JP22H02148, JP23K23416, JP25H00425).

Notes and references

1 V. Nicolosi, M. Chhowalla, M. G. Kanatzidis, M. S. Strano and J. N. Coleman, *Science*, 2013, **340**, 1226419.
2 H. P. Cong, J. F. Chen and S. H. Yu, *Chem. Soc. Rev.*, 2014, **43**, 7295.
3 X. Jin, T.-H. Gu, N. H. Kwon and S.-J. Hwang, *Adv. Mater.*, 2021, **33**, 2005922.
4 Y. Oaki, *Chem. Lett.*, 2021, **50**, 305.
5 M. Osada and T. Sasaki, *Adv. Mater.*, 2012, **24**, 210.
6 M. A. Timmerman, R. Xia, P. T. P. Le, Y. Wang and J. E. ten Elshof, *Chem.–Eur. J.*, 2020, **26**, 9084.
7 X. An and D. Yang, *Nanoscale*, 2025, **17**, 4212.
8 U. Celano, D. Schmidt, C. Beitia, G. Orji, A. V. Davydov and Y. Obeng, *Nanoscale Adv.*, 2024, **6**, 2260.
9 M. Rahman and M. S. Al Mamun, *Nanoscale Adv.*, 2024, **6**, 367.
10 Y. Oaki and Y. Igarashi, *Bull. Chem. Soc. Jpn.*, 2021, **94**, 2410.
11 B. Ryu, L. Wang, H. Pu, M. K. Y. Chan and J. Chen, *Chem. Soc. Rev.*, 2022, **51**, 1899.
12 H. He, Y. Wang, Y. Qi, Z. Xu, Y. Li and Y. Wang, *Nano Energy*, 2023, **118**, 108965.

13 B. Lu, Y. Xia, Y. Ren, M. Xie, L. Zhou, G. Vinai, S. A. Morton, A. T. S. Wee, W. G. van der Wiel, W. Zhang and P. K. J. Wong, *Adv. Sci.*, 2024, **11**, 2305277.
14 P. Chavalekvirat, W. Hirunpinyopas, K. Deshsorn, K. Jitapunkul and P. Iamprasertkun, *Precis. Chem.*, 2024, **2**, 300.
15 G. Nakada, Y. Igarashi, H. Imai and Y. Oaki, *Adv. Theory Simul.*, 2019, **2**, 1800180.
16 K. Noda, Y. Igarashi, H. Imai and Y. Oaki, *Adv. Theory Simul.*, 2020, **3**, 2000084.
17 R. Mizuguchi, Y. Igarashi, H. Imai and Y. Oaki, *Nanoscale*, 2021, **13**, 3853.
18 Y. Haraguchi, Y. Igarashi, H. Imai and Y. Oaki, *Adv. Theory Simul.*, 2021, **4**, 2100158.
19 S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito and O. Levy, *Nat. Mater.*, 2013, **12**, 191.
20 K. Rajan, *Annu. Rev. Mater. Res.*, 2015, **45**, 153.
21 K. T. Butler, J. M. Frost, J. M. Skelton, K. L. Svanea and A. Walsh, *Chem. Soc. Rev.*, 2016, **45**, 6138.
22 B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360.
23 A. Agrawal and A. Choudhary, *MRS Commun.*, 2019, **9**, 779.
24 L. Himanen, A. Geurts, A. S. Foster and P. Rinke, *Adv. Sci.*, 2019, **6**, 1900808.
25 Y. Miyake and A. Saeki, *J. Phys. Chem. Lett.*, 2021, **12**, 12391.
26 K. Terayama, M. Sumita, R. Tamura and K. Tsuda, *Acc. Chem. Res.*, 2021, **54**, 1334.
27 K. Hatakeyama-Sato, *Polym. J.*, 2023, **55**, 117.
28 R. B. Canty, B. A. Koscher, M. A. KcDonald and K. F. Jensen, *Digital Discovery*, 2023, **2**, 1259.
29 S. Lo, S. G. Baird, J. Schrier, B. Blaiszik, N. Carson, I. Foster, A. Aguilar-Granda, S. V. Kalinin, B. Maruyama, M. Politi, H. Tran, T. D. Sparks and A. Aspuru-Guzik, *Digital Discovery*, 2024, **3**, 842.
30 J. M. Granda, L. Donina, V. Dragone, D. L. Long and L. Cronin, *Nature*, 2018, **559**, 377.
31 R. Shimizu, S. Kobayashi, Y. Watanabe, Y. Ando and T. Hitosugi, *APL Mater.*, 2020, **8**, 111110.
32 B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick and A. I. Cooper, *Nature*, 2020, **583**, 237.
33 A. D. Sendek, B. Ransom, E. D. Cubuk, L. A. Pellouchoud, J. Nanda and E. J. Reed, *Adv. Energy Mater.*, 2022, **12**, 2200553.
34 Y. Zhang and C. Ling, *npj Comput. Mater.*, 2018, **4**, 25.
35 Y. Liu, B. Guo, X. Zou, Y. Li and S. Shi, *Energy Storage Mater.*, 2020, **31**, 434.
36 D. E. P. Vanpoucke, O. S. J. van Knippenberg, K. Hermans, K. V. Bernaerts and S. Mehrkanoon, *J. Appl. Phys.*, 2020, **128**, 054901.
37 P. Xu, X. Ji, M. Li and W. Lu, *npj Comput. Mater.*, 2023, **9**, 42.
38 B. Dou, Z. Zhu, E. Merkurjev, L. Ke, L. Chen, J. Jian, Y. Zhu, J. Liu, B. Zhang and G. W. Wei, *Chem. Rev.*, 2023, **123**, 8736.
39 P. Xu, X. Ji, M. Li and W. Lu, *npj Comput. Mater.*, 2023, **9**, 42.
40 S. J. Pan and Q. Yang, *IEEE Trans. Knowl. Data Eng.*, 2010, **22**, 1345.
41 S. G. Espley, E. H. E. Farrar, D. Buttar, S. Tomasi and M. N. Grayson, *Digital Discovery*, 2023, **2**, 941.
42 Y. Kim, Y. Kim, C. Yang, K. Park, G. X. Gu and S. Ryu, *npj Comput. Mater.*, 2021, **7**, 140.
43 A. U. Mahmood, M. M. Ghelardini, J. B. Tracy and Y. G. Yingling, *Chem. Mater.*, 2024, **36**, 9330.
44 L. Bustillo, T. Laino and T. Rodrigues, *Chem. Sci.*, 2023, **14**, 10378.
45 Y. Haraguchi, Y. Igarashi, H. Imai and Y. Oaki, *Digital Discovery*, 2022, **1**, 26.
46 R. Tibshirani, M. Wainwright and T. Hastie, *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman and Hall/CRC, Philadelphia, PA, 2015.
47 Y. Igarashi, H. Takenaka, Y. Nakanishi-Ohno, M. Uemura, S. Ikeda and M. Okada, *J. Phys. Soc. Jpn.*, 2018, **87**, 044802.
48 Y. Igarashi, K. Nagata, T. Kuwatani, T. Omori, Y. Nakanishi-Ohno and M. Okada, *J. Phys. Conf. Ser.*, 2016, **699**, 012001.
49 E. Candès and J. Romberg, *Inverse Probl.*, 2007, **23**, 969.
50 K. Sodeyama, Y. Igarashi, T. Nakayama, Y. Tateyama and M. Okada, *Phys. Chem. Chem. Phys.*, 2018, **20**, 22585.
51 T. Komura, K. Sakano, Y. Igarashi, H. Numazawa, H. Imai and Y. Oaki, *ACS Appl. Energy Mater.*, 2022, **5**, 8990.
52 K. Sakano, Y. Igarashi, H. Imai, S. Miyakawa, T. Saito, Y. Takayanagi, K. Nishiyama and Y. Oaki, *ACS Appl. Energy Mater.*, 2022, **5**, 2074.
53 H. Tobita, Y. Namiuchi, T. Komura, H. Imai, K. Obinata, M. Okada, Y. Igarashi and Y. Oaki, *Energy Adv.*, 2023, **2**, 1014.
54 W. Hamada, M. Hishida, R. Sugiura, H. Tobita, H. Imai, Y. Igarashi and Y. Oaki, *J. Mater. Chem. A*, 2024, **12**, 3294.
55 T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer, 2001, ch. 7.
56 I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, The Morgan Kaufmann Series in Data Management Systems, 2nd edn, 2005, pp. 149–151.
57 R. Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1143.
58 V. Vakharia, M. Shan, P. Nair, H. Borade, P. Sahlot and V. Wankhede, *Batteries*, 2023, **9**, 125.
59 P. Nair, V. Vakharia, M. Shah, Y. Kumar, M. Woźniak, J. Shafi and M. F. Ijaz, *Int. J. Intll. Syst.*, 2024, 8185044.
60 K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT press, 2012.
61 K. Obinata, T. Nakayama, A. Ishikawa, K. Sodeyama, K. Nagata, Y. Igarashi and M. Okada, *Sci. Technol. Adv. Mater.: Methods*, 2022, **2**, 355.
62 R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler and L. M. Ghiringhelli, *Phys. Rev. Mater.*, 2018, **2**, 083802.
63 K. Obinata, Y. Igarashi, K. Nagata, K. Sodeyama and M. Okada, *APL Mach. Learn.*, 2025, **3**, 016108.
64 K. Noda, Y. Igarashi, H. Imai and Y. Oaki, *Chem. Commun.*, 2021, **57**, 5921.