Batteries & Supercaps

Research Article
doi.org/10.1002/batt.202500288

Chemistry Europe
European Chemical Societies Publishing

www.batteries-supercaps.org

Performance Prediction Models with Improved Accuracy and Generalizability for Organic Cathode-Active Materials of Lithium-Ion Battery

Rika Yamamoto, Yasuhiko Igarashi, Hiroaki Imai, Taisei Sakata, Shuntaro Miyakawa, Shino Yoshizaki, Takaya Saito, and Yuya Oaki*

Development of organic energy storage requires enhancing performances of active materials. In particular, reaction potential and specific capacity of cathode-active materials have significant impact on energy density of organic lithium-ion battery. However, discovery of new compounds for active materials based on professional experience and intuition meets the limitation of huge search space of organic molecules. The performance predictors enable efficient discovery of new potential compounds. Although the predictors of potential, capacity, and energy density (models G1) are prepared in the previous work, these become older and have problems. In the present work, the updated models G2 have been constructed to improve the accuracy, usability, and generalizability. The models G2 are prepared by sparse modeling for small data combining machine learning and chemical insight on the training data set with adding new data. The updated models are validated using a new test data set and data-scientific methods. The improved predictors contribute to efficient exploration of new cathode-active materials to realize high-performance batteries.

1. Introduction

Batteries based on organic materials have advantages, such as lightweight and metal-resource free, as a future energy storage device.[1–13] In the future, flying devices, such as drone and high-attitude platform station, need to board battery with both lightweight and high energy density. Organic cathode-active materials with higher reaction potential and specific capacity are required to achieve higher energy density. Various types of redox-active moieties have been studied for cathode, such as organic radicals, disulfides, and quinones.[14–16] In general, such active materials have been found and designed by professional experience and intuition. Synthesis and electrochemical characterization of the targeted compounds are carried out with trial and error. Such conventional manual exploration of new compounds with consumption of time, cost, and effort is not a realistic method in huge chemical space of organic molecules. If the performances, such as potential and specific capacity, are predicted before the experiments based on the molecular structures, new potential compounds can be found more efficiently. In the present work, the new updated predictors for potential, specific capacity, and energy density of organic cathode-active materials (models G2) were constructed to enable efficient exploration (**Figure 1**).

Data-driven methods have been widely applied to process optimization and materials exploration in recent years.[17–24] Machine learning (ML) was used for the development of materials related to organic energy storage.[25–32] In previous works, ML was combined with computational chemistry because sufficient big data was available based on calculation. The physicochemical properties of electrolytes were predicted with the assistance of calculation and ML.[33–40] The redox potentials of cathode-active materials were estimated from the energy levels of molecular orbitals.[41–44] On the other hand, the specific capacity was not predicted in previous works. If the redox-active moieties with lithium ion (Li^+) are known, the theoretical specific capacity can be calculated from the molecular structures. However, the theoretical specific capacity of new molecules including unknown redox-active moieties cannot be calculated from the structures. Therefore, the search space is generally limited to compounds with known redox-active moieties. In addition, it is not easy to estimate the actual specific capacity from the theoretical one because various factors, such as particle size and conductivity, have influence on the utilization rate of the active materials. The prediction of the actual specific capacity, that is, the utilization rate, still remains a challenge. The actual specific capacity in the literature can be used as data for ML. However, the data size is not sufficient for conventional algorithms. Our group has studied

R. Yamamoto, H. Imai, Y. Oaki
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

Y. Igarashi
Faculty of Engineering, Information and Systems
University of Tsukuba
1-1-1 Tennodai, Tsukuba 305-8573, Japan

T. Sakata, S. Miyakawa, S. Yoshizaki, T. Saito
Advanced Battery Development Section
Advanced HAPS Research Department
Advanced Business Division
Research Institute of Advanced Technology
SoftBank Corp, 1-7-1 Kaigan, Minato-ku, Tokyo 105-7529, Japan

Supporting information for this article is available on the WWW under https://doi.org/10.1002/batt.202500288

Research Article
doi.org/10.1002/batt.202500288
Chemistry Europe
European Chemical Societies Publishing

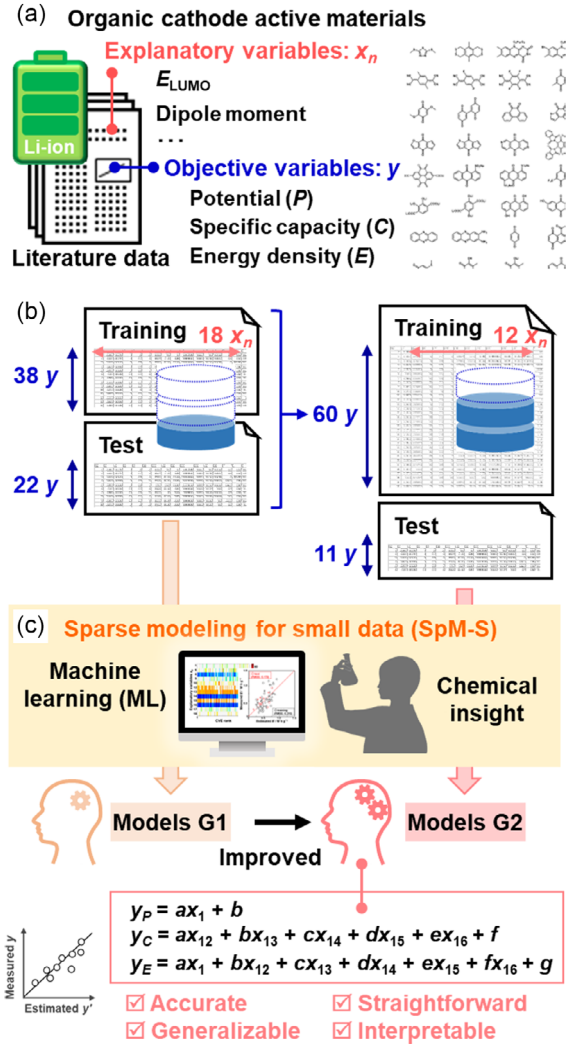


Figure 1. Schematic illustration of the present work. a) Preparation of the data sets including the explanatory (x_n) and objective (y) variables from literature data. b) Data sizes for the previous models G1 (left) and present models G2 (right). c) Model construction method (upper) and improvement of the models from G1 to G2 (lower).

sparse modeling for small data (SpM-S) by combining ML and chemical insight.[24,45–49] The performance predictors for the potential, specific capacity, and energy density (models G1) were constructed in a previous work.[49] The predictor for the actual capacity was prepared on the small literature data on the assumption that the highest specific capacity was reported with optimization of all the related conditions. However, the models G1 have low prediction accuracy and generalizability to unknown test data. In particular, as the number of the reaction sites with Li^+, corresponding to the theoretical specific capacity, is used as a descriptor of specific capacity, the model is not applied to new molecules including unknown redox moieties. In the present work, the updated models G2 were constructed to improve the accuracy and change the descriptors enabling exploration of new compounds for cathode-active materials.

2. Results and Discussion

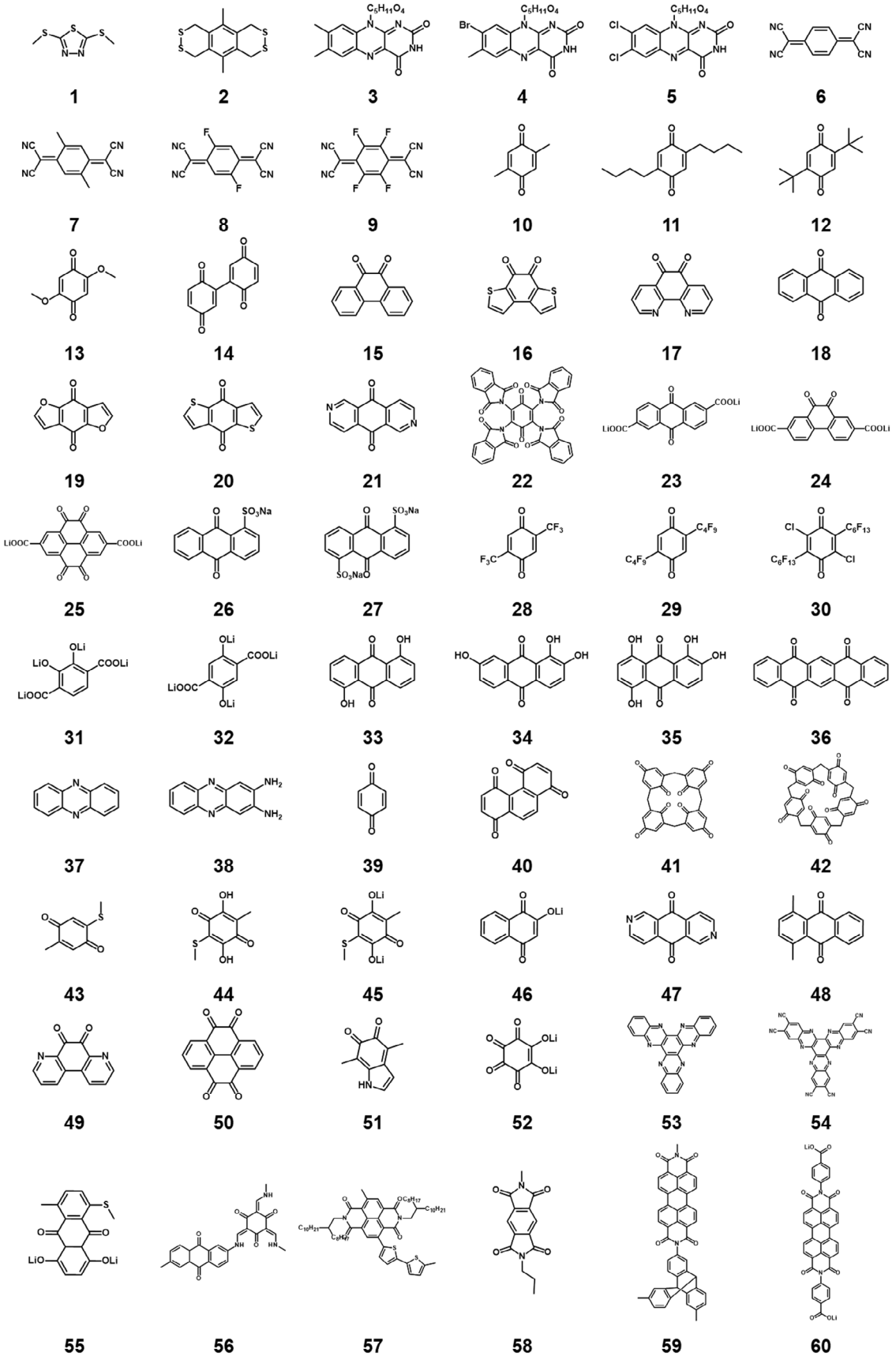
2.1. Preparation of New Training and Test Data Sets

Reaction potential (P/V vs. Li/Li^+) and specific capacity (C/mA h g^{-1}) of cathode-active materials were set as the objective variables (y) y_P and y_C, respectively (Figure 1a). The product of P and C was defined as the apparent energy density (E/W h kg^{-1}) only for the cathode-active material and set as y_E. These reported values for 60 compounds (**1**–**60**), such as quinones and imines, were read out from 36 literature data (**Scheme 1**).[50–85] In the present work, a new training data set was prepared by merging the training and test data sets in our previous work (Figure 1b and Table S1, Supporting Information).[49] In addition, a new test data set was collected from literature containing 11 compounds (**61**–**71**) (**Scheme 2** and Table S2, Supporting Information).[57,73,86–90]

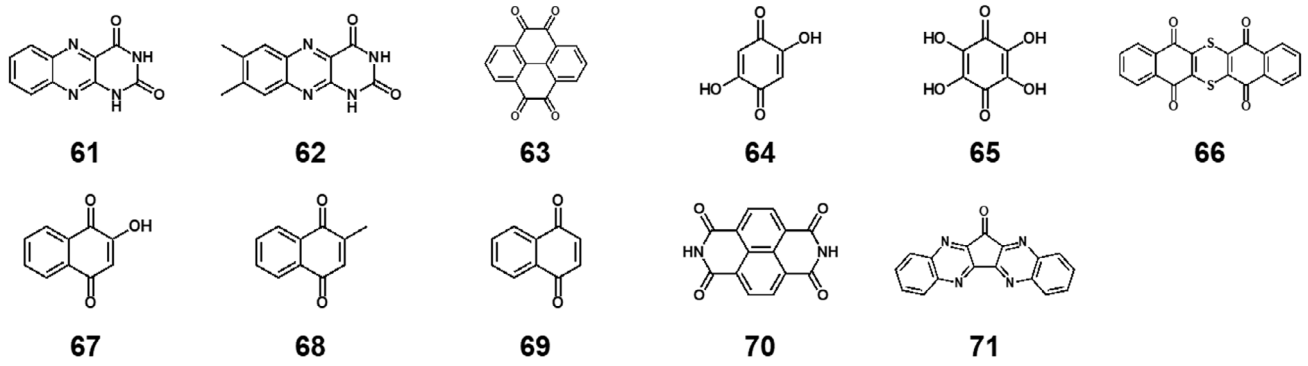
The following explanatory variables (x_n: $n = 1$–18), such as physicochemical parameters of the molecules, were selected based on our chemical insight (**Table 1**):[49] the energy levels of the lowest unoccupied molecular orbital (LUMO) (E_{LUMO0}, x_1/eV), four energy levels higher than E_{LUMO0} (E_{LUMOi}, x_{i+1}/eV ($i = 1$–4)), molecular weight (x_6/g mol^{-1}), the expected number of Li^+ reacting with one molecule (x_7/−, Figure S1, Supporting Information), the number of the conjugated carbon (x_8/−), the number of the unoccupied orbitals (N_{orb}) with the energy level (E_L) lower than the work function of lithium (N_{orb}, $E_L < \Phi_{Li}$, x_9/−) and $E_L = 0$ (N_{orb}, $E_L < 0$, x_{10}/−), the sum of absolute values of E_L in the range of $E_{LUMO0} - 0$ ($\Sigma|E_L|$, $E_{LUMO0} \leq E_L < 0$, x_{11}/eV), each term of Hansen solubility (similarity) parameter (HSP), namely dispersion (δD), polarity (δP), and hydrogen-bonding (δH) (x_{12-14}/−, respectively), dipole moment (x_{15}/Debye), the maximum and minimum values of the partial charge density (x_{16} and x_{17}/−, respectively), and the ratio of the number of the heteroatoms to the total number of atoms except hydrogen (R_{hetero}, x_{18}/−). All these x_n ($n = 1$–18) were used for the construction of the previous models G1. In the present work, x_n ($n = 2$–5, 7, 11) was removed for the construction of the models G2. As the correlation was large for x_2–x_5, x_{10}, and x_{11} (Figure S2, Supporting Information), x_2–x_5 and x_{11} were removed to avoid the multicollinearity (Table 1). The search space is limited when the exploration is performed using the models G1. A descriptor in the former model for the specific capacity included x_7, which is the number of Li^+ reacting with one molecule. As the reaction sites are manually counted based on the molecular structures, x_7 is calculated for the already known reaction sites in the molecules. In other words, x_7 is not calculated for new molecules without the reported reaction sites. In the present work, x_7 was not used to construct the more generalizable models G2 toward exploration in wider search space. In this manner, the data sets were comprised of 60 y and 12 x_n for training and 11 y and 12 x_n for test (Figure 1b).

Research Article
doi.org/10.1002/batt.202500288

Chemistry Europe
European Chemical Societies Publishing



Scheme 1. Molecular structure of **1**–**60** for the training data set.[50–85]



Scheme 2. Molecular structure of **61**–**71** for the test data set.[57,73,86–90]

Table 1. List of explanatory variables used for the models G_1 and G_2 (x_n: $n = 1$–18).

n/–	Parameter	Unit	G_1	G_2
1[a)]	E_{LUMO0}	eV	G_1	G_2
2[a)]	E_{LUMO1}	eV	G_1	–
3[a)]	E_{LUMO2}	eV	G_1	–
4[a)]	E_{LUMO3}	eV	G_1	–
5[a)]	E_{LUMO4}	eV	G_1	–
6	Molecular weight (M_w)	g mol^{-1}	G_1	G_2
7	Expected n_{Li}/molecule	–	G_1	–
8	Number of conjugated carbons	–	G_1	G_2
9[a)]	N_{orb}, $E < \Phi_{Li}$	–	G_1	G_2
10[a)]	N_{orb}, $E < 0$	–	G_1	G_2
11[a)]	$\Sigma \mid E \mid$, $E < 0$	eV	G_1	–
12[b)]	HSP-δD	–	G_1	G_2
13[b)]	HSP-δP	–	G_1	G_2
14[b)]	HSP-δH	–	G_1	G_2
15[a)]	Diploe moment	Debye	G_1	G_2
16[a)]	Maximum of partial charge density	–	G_1	G_2
17[a)]	Minimum of partial charge density	–	G_1	G_2
18	R_{hetero}	–	G_1	G_2

[a)]DFT calculation; [b)]HSP calculation.

2.2. Construction of Models G2

2.2.1. Construction of Model G2—Potential (P)

The descriptors of *P*, *C*, and *E* were extracted using an exhaustive search with linear regression (ES-LiR), a method of ML (**Figure** 1c and **2** and Table S1, Supporting Information).[24,91] Our experience and chemical insight were used for the selection of the descriptors. In ES-LiR, multiple linear regression models were exhaustively constructed for all the possible combinations of x_n(12 x_n, $n = 1,6,8$–$10,12$–18); total 2^{12}–1(=2.6×10^5) models were prepared using the training data set with cross-validation. The models were ranked in the ascending order of cross-validation error (CVE). The weight diagram summarizes the color-coded coefficients of each x_n for the top 10^2 in descending order of the CVE rank (Figure 2a,c). The densely colored x_n is frequently used in the regression models. The significant descriptors were extracted from the weight diagram based on the color density and our chemical insight. The linear regression model was prepared using the selected x_n. Each x_n was normalized based on mean (0) and standard deviation (1) to represent the relative weight to *y*. One descriptor x_1 was extracted from the weight diagram for *P* (Figure 2a). The predictor for *P* ($y_P{}'$) was described by Equation (1) with root mean square error (RMSE) 0.255 V (black circles in Figure 2b).

$$y_P{}' = -0.239x_1 + 2.48 \quad (1)$$

In fivefold cross validation, the average RMSE was 0.254 ± 0.010 V for the training and 0.266 ± 0.040 V for the test in the training data set containing compounds **1**–**60** (Figure S3, Supporting Information). The validation using the test data set containing compounds **61**–**71** showed RMSE 0.152 V (red circles in Figure 2b and Scheme 2 and Table S2, Supporting Information). The results indicate the sufficient accuracy of the constructed model (Equation (1)). Tenfold cross validation was performed by merging the original training and test data sets (Figure S4, Supporting Information). If the constructed model has the specificity to the original training data set, RMSE increased in this 10-fold cross validation. The average RMSE was 0.241 ± 0.004 V for the training and 0.249 ± 0.039 V for the test (Figure S3, Supporting Information). The RMSE values were not increased by changing the data sets. In addition, the correlation between E_{LUMO0} and redox potential was reported in the previous calculational studies[92,93] These results indicate the validity of the constructed model Equation (1).

2.2.2. Construction of Model G2—Specific Capacity (C)

The weight diagram for *C* indicates the strong correlation of x_8, x_{12}, x_{13}, x_{14}, x_{15}, and x_{16} based on the color density (Figure 2c). Here x_8 was not selected as the descriptors because the manual count is required to estimate the number of conjugated carbons. The specific capacity prediction model ($y_C{}'$) Equation (2) using x_{12}, x_{13}, x_{14}, x_{15}, and x_{16} had RMSE 81.5 Ma h g^{-1} (black circles in Figure 2d).

$$y_C{}' = 10.7x_{12} + 36.2x_{13} - 45.5x_{14} + 30.8x_{15} - 57.6x_{16} + 222 \quad (2)$$

Research Article
doi.org/10.1002/batt.202500288
Chemistry Europe
European Chemical Societies Publishing

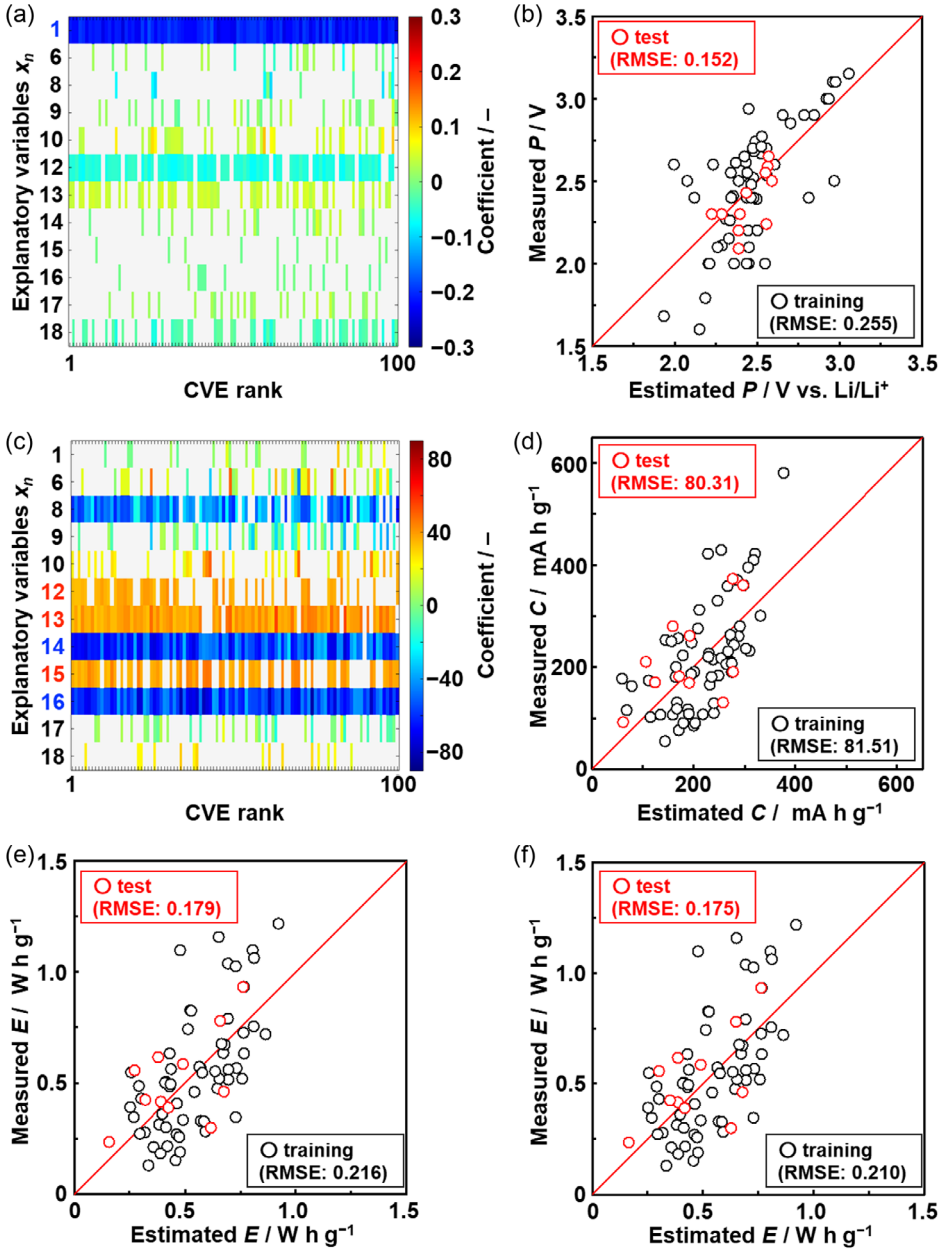


Figure 2. Construction of the models G2. a) Weight diagram representing the coefficients of each x_n for *P*. b) Relationship between the estimated and measured *P* to the training (black) and test (red) data sets. c) Weight diagram for *C*. d) Relationship between the estimated and measured *C* to the training (black) and test (red) data sets. e,f) Relationship between the estimated and measured *E* calculated using e) y_{E1}' (=$y_P' \times y_C'$, Equation (3)) and f) y_{E2}' (linear model, Equation (4)) to the training (black) and test (red) data sets.

The average RMSE values were 80.5 ± 3.38 mA h g^{-1} for the training and 89.2 ± 14.8 mA h g^{-1} for the test in the fivefold cross validation using the original training data set (Figure S5, Supporting Information). RMSE for the test data set was 80.3 mA h g^{-1} (red circles in Figure 2d). The 10-fold cross validation using the merged data set showed the average RMSE 79.8 ± 2.41 mA h g^{-1} for the training and 84.8 ± 21.5 mA h g^{-1} for the test (Figure S5, Supporting Information). These RMSE values support the validity of the constructed model.

The use of x_n as the descriptors is consistent with our chemical insight. Three descriptors x_{12}, x_{13}, and x_{14} as each term of HSP can be regarded as a simple molecular fingerprint. The positive correlation of x_{12} (HSP-δD) indicates that the molecules with the larger van der Waals interaction enable the higher specific capacity. The larger conjugated moiety serves as the stable framework for the redox-active functional groups. The positive correlations of x_{13} (HSP-δP) and x_{15} (dipole moment) indicate that the more charge-localized states in the molecule exhibit higher reactivity to Li^+. The maximum value of the partial charge density (x_{16}) had negative correlation to the specific capacity. The charge negativity of the molecule promotes the reactivity with Li^+. The negative correlation of x_{14} (HSP-δH) means that the high specific capacity is achieved by molecules with the smaller number of hydrogen-bonding donor and acceptor moieties. As carbonyl in the

quinone moiety acts as the active site, the fact is not simply consistent with our chemical insight. In the training data, the compounds with the much redox-active sites on the smaller conjugated framework exhibited lower specific capacity: examples of this type are 162 mA h g^{-1} for **28** and 117 mA h g^{-1} for **31** (Scheme 1 and Table S1, Supporting Information). The balance between the expansion of the conjugated framework and number of the reaction sites is significant to achieve high specific capacity. In this manner, the selected descriptors are interpretable based on our chemical insight. As x_7, the number of the reacted Li^+, was used in the model G1, the specific capacity was not predicted for new molecules with unknown redox-active sites. Our new models G2 afford the performance prediction of unknown compounds because all the descriptors are simply calculated based on the molecular structure. The models G2 can be widely applied to predict the specific capacity without prior knowledge. In SpM-S, all the possible regression models are exhaustively prepared and visualized in the weight diagram. The more significant descriptors can be selected based on our chemical insight. This variable selection method combining ML and our insight can avoid overfitting small data and selecting less significant descriptors.

2.2.3. Construction of Model G2—Apparent Energy Density (E)

E was calculated by the product of *P* and *C* in the original data sets. The predictor for *E* (y_{E1}') was constructed by the product of the already constructed predictors y_P' and y_C' (Equation (3)) with RMSE 0.216 W h g^{-1} (black circles in Figure 2e).

$$y_{E1}' = y_P' \times y_C' \tag{3}$$

In addition, a linear-regression model (y_{E2}') was prepared based on the original training data set using the same six descriptors x_1, x_{12}, x_{13}, x_{14}, x_{15}, and x_{16} (Equation (4)).

$$y_{E2}' = -0.0408x_1 + 0.0174x_{12} + 0.0997x_{13} - 0.112x_{14} + 0.0654x_{15} - 0.139x_{16} + 0.550 \tag{4}$$

RMSE of this linear predictor was 0.210 W h g^{-1} (black circles in Figure 2f), which is comparable to that of Equation (3). The validation using the test data sets showed RMSE 0.179 W h g^{-1} for the model Equation (3) and 0.175 W h g^{-1} for the model Equation (4) (red circles in Figure 2e,f). The results indicate that both the models have the same prediction accuracy. The present work selected the model Equation (4) because the contribution of each x_n is represented by the coefficients of the linear formula more clearly. The fivefold cross validation of the model Equation (4) showed the average RMSE 0.207 ± 0.011 W h g^{-1} for the training and 0.242 ± 0.040 W h g^{-1} for the test (Figure S6, Supporting Information). In addition, the similar RMSE values, 0.202 ± 0.007 W h g^{-1} for the training and 0.218 ± 0.065 W h g^{-1} for the test, were obtained by the 10-fold cross validation using the merged data sets (Figure S6, Supporting Information). As the predictor of *E* is comprised of the descriptors same as those of *P* and *C*, the model Equation (4) has interpretability and usability.

In our previous works, the constructed models were used to predict the performances of compounds as new active materials.[46,48] The potential compounds can be listed based on our experience and molecular design. The larger number of compounds is prepared using some database and automatic generation of molecules. The charge–discharge measurement is prioritized based on the predicted specific capacity. In this manner, discovery of new cathode-active materials can be accelerated using the resultant model.

2.3. Improved Prediction Accuracy of Models G2

The prediction accuracy of the models G2 was improved with addition of training data (**Figure 3**). The relationship between the estimated and actual values was summarized with RMSE of the models G1 to the previous smaller training data set (compounds **1**–**38**) (black circles in Figure 3a–c). The validation of the models G1 was carried out using the new test data set (compounds **61**–**71**) (red circles in Figure 3a–c). In Figure 3d–f, RMSE values to the original training and new present test data sets were represented by gray and colored bars, respectively. Although RMSE to the training data set slightly increased for the models G2 (dashed gray arrows in Figure 3d–f), that to the test data set decreased (colored arrows in Figure 3d–f). RMSE of the models G2 decreased to 78% for *P*, 38% for *C*, and 34% for *E* compared with those of the models G1. The relationship between the estimated and actual values showed that a couple of plots were deviated from the diagonal line, particularly in the larger region of *C* and *E* (dashed red circles in Figure 3a–c). The models G1 showed the overfitting to the training data and the low generalizability to unknown test data. The results indicate that the model G2 had the improved accuracy and generalizability to predict *P*, *C*, and *E*, particularly for the larger values of *C* and *E*. Two disulfide compounds in the literature were used for further validation: tetrathiotetracene (TTT) and hexathiapentacene (HTP).[94] The estimated *P* was 2.19 V for TTT and 2.47 V for HTP, whereas the reported *P* was 1.85 V for TTT and 1.65 V for HTP. The estimated *C* was 205 mA h g^{-1} for TTT and 102 mA h g^{-1} for HTP, whereas the reported *C* was 257 mA h g^{-1} for TTT and 265 mA h g^{-1} for HTP. The RMSE values 0.63 V for *P* and 121 mA h g^{-1} for *C* were larger than those of the test data containing quinones and imines as displayed in Figure 2b,d. The performance of the disulfides was not accurately predicted because only two disulfides (**1** and **2**) were included in the training data set. The fact implies that further improvement is required to construct the more generalizable predictors. The improved predictors can be prepared by adding the data of other compounds, such as disulfides.

2.4. Effect of Data Size on Model Construction

2.4.1. Data Size for ES-LiR

The data size was increased for the construction of the model G2. If the data size is sufficient, the same x_n can be extracted from the weight diagram with a slight decrease in the data size. After the

Research Article
doi.org/10.1002/batt.202500288
Chemistry Europe
European Chemical Societies Publishing

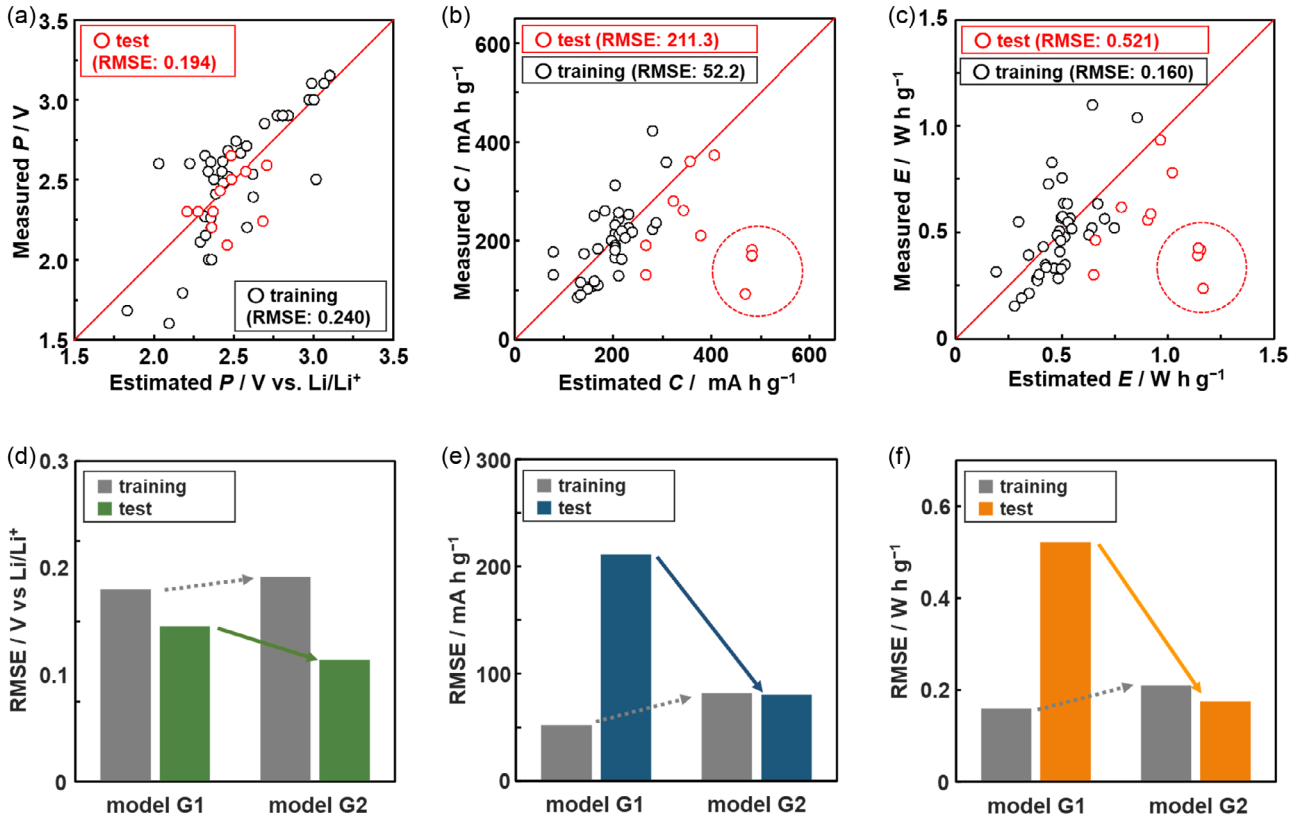


Figure 3. Prediction accuracy of the models G1 and G2. a–c) Relationship between the estimated and measured a) *P*, b) *C*, and c) *E* to the previous original training data set comprised of the compounds **1**–**38** (black) and new present test data set comprised of the compounds **61**–**71** (red) using the models G1. The dashed circles indicate the test data with the particularly large RMSE. d–f) RMSE of the models G1 (left, gray) and G2 (right, gray) to the original training data sets containing 38 *y* and 60 *y*, respectively. RMSE of the models G1 (left, colored) and G2 (right, colored) to the same new test data set containing 11 *y*.

original data set containing 60 *y* was randomly decreased to that containing 55, 50, 45, and 40 *y*, the weight diagrams were prepared by ES-LiR using the reduced data sets (**Figure 4**). The extractability of the descriptors for the specific capacity prediction was studied using the reduced data sets. In the original training data set (60 *y*), x_8, x_{10}, x_{12}, x_{13}, x_{14}, x_{15}, and x_{16} were extractable in the original weight diagram (Figure 2c). The same x_n were visible in the weight diagram prepared using the reduced data set containing 55 *y* (Figure 4a). Although the same descriptors were extractable for the reduced data set containing 50 *y*, x_{10} showed the intensified colors in the weight diagram (Figure 4b). When the data size was further decreased to 45 and 40 *y*, the same descriptors were not visually extracted from the weight diagrams (Figure 4c,d). The color of x_{13} and x_{15} faded and the color of x_{10} intensified. The results support that the model G2 was constructed with a sufficient size of the training data containing 60 *y*.

2.4.2. Data Size for Exhaustive Search with Bayesian Model Averaging (ES-BMA)

In ES-LiR, descriptors are visually selected by researchers using the weight diagram. Our visual recognition and domain knowledge assist the variable selection. Here ES-BMA was applied to extract the descriptors without our chemical insight. In ES-BMA,[95] the likelihood of each linear regression model is represented by a probability value of being a descriptor (*p*). A method for quantitatively evaluating the confidence level of variable selection is introduced to estimate the uncertainty for all the combinations of variables. All the possible models for each x_n are integrated with weighting by the posterior probabilities. Whereas the initial *p* is 0.5 for all x_n, ES-BMA provides *p* as a quantitative metric for the extraction of the descriptors.

The selected descriptors x_{12}, x_{13}, x_{14}, and x_{16} in the model G2 have *p* larger than 0.6 (**Figure 5**a). As a reference, the reduced training data set G1′ with the same data size for the model G1 (38 *y*) was prepared by a random sampling from new training dataset for the models G2. When the reduced data set G1′ was used for ES-BMA, *p* had no significant differences (Figure 5b). Therefore, the specific x_n was not extracted only using ES-BMA from the reduced data set because of the small data size (38 *y*). The domain knowledge assisted the variable selection for the models G1. On the other hand, most of the descriptors were extractable by ES-BMA because of the sufficient data size. ES-BMA supports the validity of the model G2 in terms of the selected descriptors and data size. Here it is not easy to define the threshold of *p* in ES-BMA theoretically. A previous report using another algorithm empirically set the threshold at $p = 0.9$ under the sufficient data size.[96] When the data size is small, the distinct differences cannot appear in the *p* value. In the present work, the deviation from the tentatively defined threshold $p = 0.6$

Research Article
doi.org/10.1002/batt.202500288
Chemistry Europe
European Chemical Societies Publishing

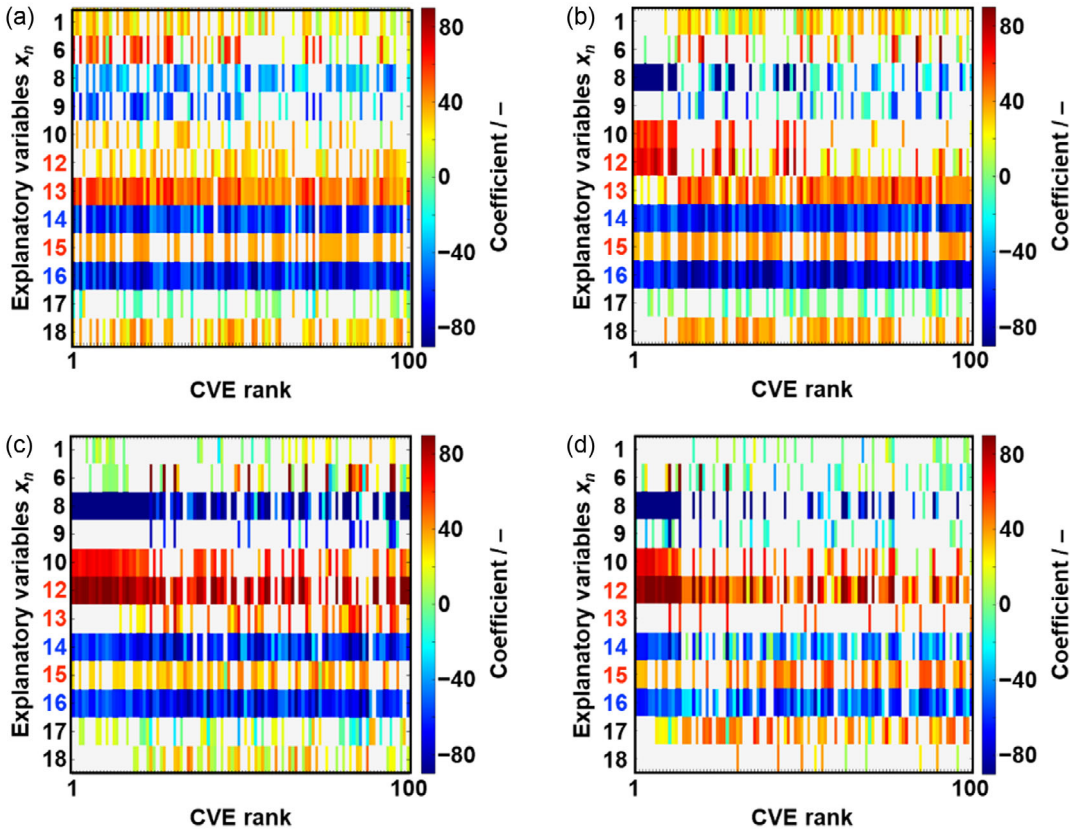


Figure 4. Weight diagrams for *C* prepared with random reducing the size of the training data set to a) 55 *y*, b) 50 *y*, c) 45 *y*, and d) 40 *y*. The colored x_n indicates the selected descriptors in the model Equation (2).

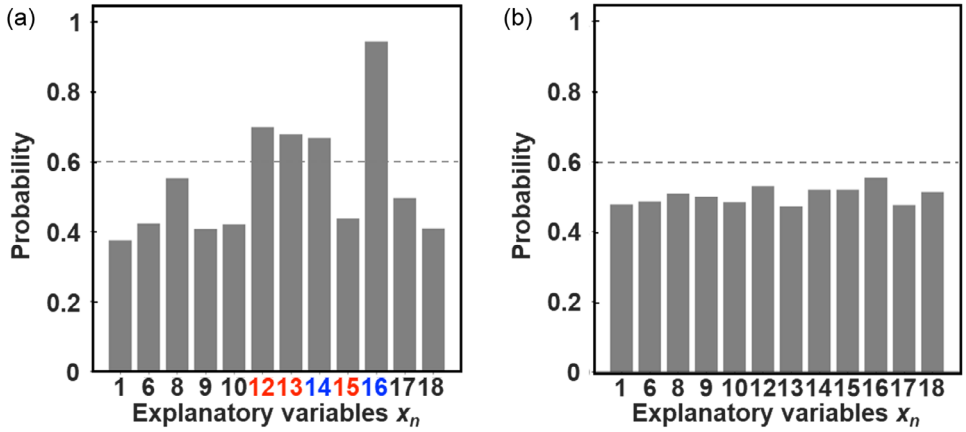


Figure 5. *p* value calculated by ES-BMA using the a) original training data set (60 *y*) and b) randomly reduced data set (38 *y*).

was used in the selection of the descriptors. The calculation method for the *p* value was studied based on statistical inference.[97] Combination of the method and ES-BMA is a current challenge to derive an accurate *p* value as the appropriate threshold.

2.5. Prediction Based on Other ML Algorithms

SpM-S has been studied as a ML approach advantageous for small data.[24,45,47,49] The accuracy of the model G2 was compared with that of the models constructed by other ML algorithms (**Figure 6**). The same training and test data sets were used for this demonstration (Table S1 and S2, Supporting Information). Least absolute shrinkage and selection operate (LASSO) and multiple linear regression (MLR) with and without the variable selection were used as linear regression models, respectively. The following nonlinear regression modeling methods were used: support vector regression (SV-R), random forest regression (RF-R), and neural network regression (NN-R). RMSE to the training and test data sets were calculated as a metric for the

Research Article
doi.org/10.1002/batt.202500288
Chemistry Europe
European Chemical Societies Publishing

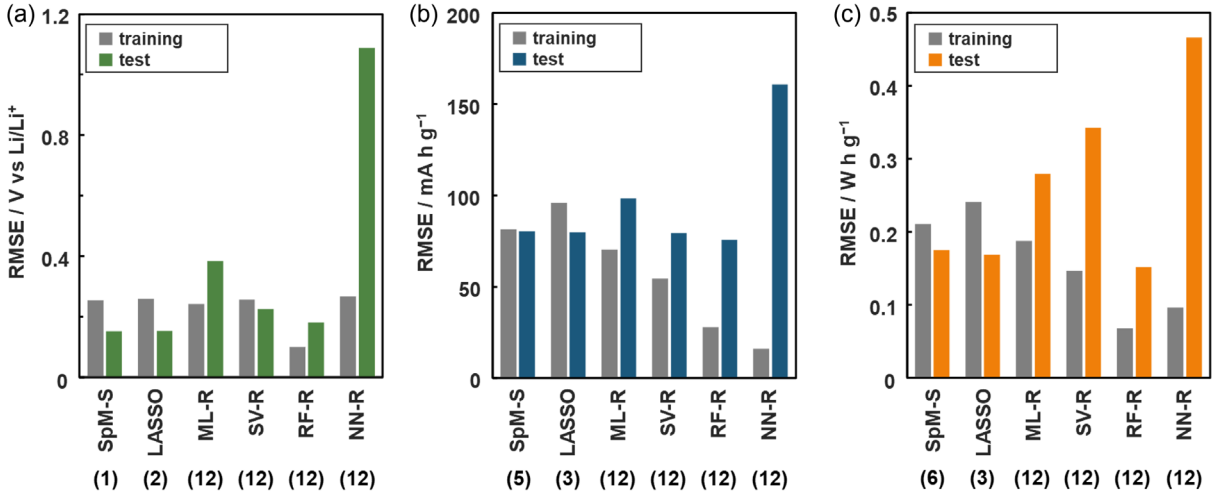


Figure 6. RMSE of the models for a) *P*, b) *C*, and c) *E* constructed using the different ML algorithms to the original training (gray) and test (colored) data sets. The number of the used descriptors was listed below the name of the algorithms.

comparison. The detailed method of these algorithms was referred to in our previous work.[49]

RMSE of the models G2 using SpM-S was smaller than that using LASSO to the training data sets (gray bars in Figure 6). For the test data sets, RMSE had no distinct differences in these two modeling methods (colored bars in Figure 6). ML-R exhibited larger RMSE values in the test data set compared with SpM-S, even though a smaller RMSE was achieved in the training data set. This trend means that ML-R causes overfitting to the training data and lowers the generalizability. In the nonlinear algorithms, RMSE to the test data set was quite larger than that to the training data set, except SV-R for *P*. As these nonlinear models cause the overtraining of the training data, the generalizability to unknown test data is lowered. Based on these results, the more generalizable models are constructed by SpM-S compared with the other algorithms. A limited number of the significant descriptors are extracted by SpM-S compared with the other linear and nonlinear models. The contribution of each descriptor is interpreted with the coefficient in the linear regression model. It is not easy to interpret the contribution and weight of all the used descriptors for the nonlinear models. Therefore, the interpretability of SpM-S is effective in future material design combined with professional experience and insight.

The same demonstration was carried out using the merged data set of the original training and test ones (Figure S7, Supporting Information). The merged data set was divided into five groups; four groups were used for training and the remaining one was assigned to test. The average RMSE values were calculated with changes in the assignment of the test for five different patterns using linear and nonlinear modeling methods. The same trend displayed in Figure 6 was observed in this demonstration. The results support the advantages of SpM-S: straightforwardness, generalizability, and interpretability. SpM-S provides the linear regression model comprising a limited number of significant descriptors with combination of ML and our chemical insight. The generalizability is improved by avoiding overtraining. Moreover, the contribution of each descriptor is easily interpretable in the linear model compared with the other modeling methods. Although sparse modeling with various algorithms was used in materials science and chemistry,[22,33] the domain knowledge, such as experience and chemical insight, was not fully combined in the model construction processes. Our group has studied combining experience, insight, and intuition with sparse modeling.[24,45–49] A recent report indicates that combination of scientific curiosity and ML is significant for unprecedented findings.[98] In this manner, collaboration of ML and scientists can be significant for the accelerated discovery of new materials and optimization of processes.

In recent reports, graph neural network (GNN) was used for design of polymer materials, such as electrolyte.[99,100] Cycle stability of batteries was predicted using new methods, such as GNN and transfer learning.[101–103] If graph and node as features are set in appropriate manner, GNN can be effectively used under sufficient amount of data. Domain knowledge is significant for the appropriate definition of graph and node. However, neural network algorithm requires sufficient data size and lowers the interpretability of the model. Figure 6 implies that the current small data is insufficient for neural network. Although small data related to the targets is utilized by transfer learning, the prior data with sufficient quality and quantity is required for training before the transfer. As SpM-S constructs a linear regression model using a limited number of descriptors, the straightforwardness and interpretability have advantage compared with these recent algorithms.

3. Conclusions

New updated performance predictors were constructed for organic cathode-active materials of lithium-ion battery. The prediction models for the potential, specific capacity, and apparent energy density were prepared based on the small data by SpM-S combining ML and our chemical insight. The models G1 in our previous works had low accuracy and generalizability because

of the lack of data size. In addition, a descriptor corresponding to theoretical specific capacity was used in the previous predictor. The specific capacity of new compounds with unknown redox-active moieties was not predicted by model G1. The models G2 were obtained with the addition of training data and validated using new test data. The specific capacity predictor was constructed using the descriptors alternative to theoretical specific capacity. When the validation was carried out using the same test data set, the RMSE of the models G2 decreased to 78% for *P*, 38% for *C*, and 34% for *E* compared with those of the models G1. The improved prediction accuracy was achieved for the models G2 to the test data set. The data scientific validation supports that the models G2 were constructed in a sufficient size of data. Moreover, our SpM-S provided the linear prediction models with straightforwardness, interpretability, and generalizability compared with the models constructed using the other algorithms. Potential new compounds are designed from the known redox-active molecules based on professional experience. In recent years, automatic molecular generation has provided a large number of potential compounds. The updated models G2 can be applied to prioritize these potential compounds for the experiments. In this manner, the resultant models G2 can be applied to explore new organic cathode materials in a wide search space of organic compounds efficiently.

Acknowledgements

This work was supported by JST PRESTO (Y.O., JPMJPR16N2 and Y.I. JPMJPR17N2), JST CREST (Y.I., JPMJCR21O1), JSPS-KAKENHI (JP22K19071, Y.O.), the New Energy and Industrial Technology Development Organization (NEDO) (JPNP14004), and Mitsubishi Foundation (Y.O., 202410007).

Conflict of Interest

The authors declare no conflict of interest.

Keywords: machine learning · organic cathode-active materials · performance predictors · small data · sparse modeling

[1] P. Novák, K. Müller, K. S. V. Santhanam, O. Hass, *Chem. Rev.* **1997**, *97*, 207.
[2] H. Nishide, K. Oyaizu, *Science* **2008**, *319*, 737.
[3] Z. Song, H. Zhou, *Energy Environ. Sci.* **2013**, *6*, 2280.
[4] S. Muench, A. Wild, C. Friebe, B. Häupler, T. Janoschka, U. S. Schubert, *Chem. Rev.* **2016**, *116*, 9438.
[5] S. Lee, J. Hong, K. Kang, *Adv. Energy Mater.* **2020**, *10*, 2001445.
[6] J. J. Shea, C. Luo, *ACS Appl. Mater. Interfaces* **2020**, *12*, 5361.
[7] P. Poizot, J. Gaubicher, S. Renault, L. Dubois, Y. Liang, Y. Yao, *Chem. Rev.* **2020**, *120*, 6490.
[8] Y. Chen, C. Wang, *Acc. Chem. Res.* **2020**, *53*, 2636.
[9] K. Hatakeyama-Sato, K. Oyaizu, *Chem. Rev.* **2023**, *123*, 11336.
[10] H. Guo, C. Wang, *ChemSusChem* **2024**, *17*, e202301586.
[11] M. R. Raj, G. Lee, M. V. Reddy, K. Zaghib, *ACS Appl. Energy Mater.* **2024**, *7*, 8196.
[12] R. Dantas, C. Ribeiro, M. Souto, *Chem. Commun.* **2024**, *60*, 138.
[13] T. Banerjee, R. Kundu, *Energy Fuels* **2024**, *38*, 12487.
[14] D. Xu, M. Liang, S. Qi, W. Sun, L.-P. Lv, F.-H. Du, B. Wang, S. Chen, Y. Wang, Y. Yu, *ACS Nano* **2021**, *15*, 47.
[15] K. Zou, W. Deng, D. S. Silverster, G. Zou, H. Hou, C. E. Banks, L. Li, J. Hu, X. Ji, *ACS Nano* **2024**, *18*, 19950.
[16] X. Peng, J. Guo, D. Huang, B. Ouyang, Y. Du, H. Yang, *ChemSusChem* **2024**, *17*, e202401975.
[17] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, O. Levy, *Nat. Mater.* **2013**, *12*, 191.
[18] K. Rajan, *Annu. Rev. Mater. Res.* **2015**, *45*, 153.
[19] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360.
[20] K. T. Butler, D. W. Davis, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.
[21] L. Himanen, A. Geurts, A. S. Foster, P. Rinke, *Adv. Sci.* **2019**, *6*, 1900808.
[22] Y. Miyake, A. Saeki, *J. Phys. Chem. Lett.* **2021**, *12*, 12391.
[23] K. Terayama, M. Sumita, R. Tamura, K. Tsuda, *Acc. Chem. Res.* **2021**, *54*, 1334.
[24] Y. Oaki, Y. Igarashi, *Bull. Chem. Soc. Jpn.* **2021**, *94*, 2410.
[25] G. Huang, F. Huang, W. Dong, *Chem. Eng. J.* **2024**, *492*, 52294.
[26] A. Chen, X. Zhang, Z. Zhou, *InfoMat* **2020**, *2*, 553.
[27] Y. Shao, L. Knijff, F. M. Dietrich, K. Hermansson, C. Zhang, *Batteries Supercaps* **2021**, *4*, 585.
[28] P. Thakkar, S. Khatri, D. Dobariya, D. Paterl, B. Dey, A. K. Singh, *J. Energy Storage* **2024**, *81*, 110452.
[29] S. Haghi, M. F. V. Hidalgo, M. F. Niri, R. Daub, J. Marco, *Batteries Supercaps* **2023**, *6*, e202300046.
[30] Y. Xu, J. Ge, C. W. Ju, *Energy Adv.* **2023**, *2*, 896.
[31] R. Fan, B. Xue, P. Tian, X. Zhang, X. Yuan, H. Zhang, *Chem. Commun.* **2024**, *60*, 14303.
[32] T. T. Wu, G. L. Dai, J. J. Xu, F. Cao, X. H. Zhang, Y. Zhao, Y. M. Qian, *Rare Met.* **2023**, *42*, 3269.
[33] K. Sodeyama, Y. Igarashi, T. Nakayama, Y. Tateyama, M. Okada, *Phys. Chem. Chem. Phys.* **2018**, *20*, 22585.
[34] K. Hatakeyama-Sato, T. Kashikawa, K. Kimura, K. Oyaizu, *Adv. Intel. Syst.* **2021**, *3*, 2000209.
[35] S. Jeschke, P. Johansson, *Batteries Supercaps* **2021**, *4*, 1156.
[36] K. Li, J. Wang, Y. Song, Y. Wang, *Nat. Commun.* **2023**, *14*, 2789.
[37] S. S. Manna, S. Manna, B. Pathak, *J. Mater. Chem. A* **2023**, *11*, 21702.
[38] W. Sukmas, J. Qin, R. Chanajaree, *Adv. Theory Simul.* **2024**, *8*, 2401048.
[39] H. D. Jagad, J. Fu, W. R. Fullerton, C. Li, E. Detsi, Y. Qi, *J. Electrochem. Soc.* **2024**, *171*, 060516.
[40] T. Ootahara, K. Hatakeyama-Sato, M. I. Thomas, Y. Takeoka, M. Rikukawa, M. Yoshizawa-Fujita, *ACS Appl. Electron. Mater.* **2024**, *6*, 5866.
[41] S. Xu, J. Liang, Y. Yu, R. Liu, Y. Xu, X. Zhu, Y. Zhao, *J. Phys. Chem. C* **2021**, *125*, 21352.
[42] R. P. Carvalho, C. F. N. Marchiori, D. Brandell, C. M. Araujo, *Energy Storage Mater.* **2022**, *44*, 313.
[43] X. Zhou, A. Khetan, J. Zheng, M. Huijben, R. A. Janssen, S. Er, *Digital Discovery* **2023**, *2*, 1016.
[44] J. Du, J. Guo, Q. Sun, W. Liu, T. Liu, G. Huang, X. Zhang, *J. Mater. Chem. A* **2024**, *12*, 12034.
[45] Y. Haraguchi, Y. Igarashi, H. Imai, Y. Oaki, *Digital Discovery* **2022**, *1*, 26.
[46] T. Komura, K. Sakano, Y. Igarashi, H. Numazawa, H. Imai, Y. Oaki, *ACS Appl. Energy Mater.* **2022**, *5*, 8990.
[47] H. Tobita, Y. Namiuchi, T. Komura, H. Imai, K. Obinata, M. Okada, Y. Igarashi, Y. Oaki, *Energy Adv.* **2023**, *2*, 1014.
[48] W. Hamada, M. Hishida, R. Sugiura, H. Tobita, H. Imai, Y. Igarashi, Y. Oaki, *J. Mater. Chem. A* **2024**, *12*, 3294.
[49] K. Sakano, Y. Igarashi, H. Imai, S. Miyakawa, T. Saito, Y. Takayanagi, K. Nishiyama, Y. Oaki, *ACS Appl. Energy Mater.* **2022**, *5*, 2074.
[50] M. M. Doeff, S. J. Visco, L. C. De Jonghe, *J. Appl. Elctrochem.* **1992**, *22*, 307.
[51] S. R. Deng, L. B. Kong, G. Q. Hu, T. Wu, D. Li, Y. H. Zhou, Z. Y. Li, *Electrochim. Acta* **2006**, *51*, 2589.
[52] M. Lee, J. Hong, D. H. Seo, D. H. Nam, K. T. Nam, K. Kang, C. B. Park, *Angew. Chem. Int. Ed.* **2013**, *52*, 8322.
[53] S. Nishida, Y. Yamamoto, T. Takui, Y. Morita, *ChemSusChem* **2013**, *6*, 794.
[54] T. Yokoji, Y. Kameyama, S. Sakaida, N. Maruyama, M. Satoh, H. Matsubara, *Chem. Lett.* **2015**, *44*, 1726.
[55] M. Yao, H. Senoh, S. Yamazaki, Z. Siroma, T. Sakai, K. Yasuda, *J. Power Sources* **2010**, *195*, 8336.
[56] T. Yokoji, Y. Kameyama, N. Maruyama, H. Matsubara, *J. Mater. Chem. A* **2016**, *4*, 5457.
[57] Y. Liang, P. Zhang, J. Chen, *Chem. Sci.* **2013**, *4*, 1330.
[58] Y. Liang, P. Zhang, S. Yang, Z. Tao, J. Chen, *Adv. Energy Mater.* **2013**, *3*, 600.
[59] Z. Luo, L. Liu, Q. Zhao, F. Li, J. Chen, *Angew. Chem. Int. Ed.* **2017**, *56*, 12561.
[60] A. Shimizu, H. Kuramoto, Y. Tsujii, T. Nokami, Y. Inatomi, N. Hojo, H. Suzuki, J. Yoshida, *J. Power Sources* **2014**, *260*, 211.

[61] W. Wan, H. Lee, X. Yu, C. Wang, K. W. Nam, X. Q. Yang, H. Zhou, *RSC Adv.* **2014**, *4*, 19878.
[62] T. Yokoji, H. Matsubara, M. Satoh, *J. Mater. Chem. A* **2014**, *2*, 19347.
[63] S. Gottis, A. L. Barres, F. Dolhem, P. Poizot, *ACS Appl. Mater. Interfaces* **2014**, *6*, 10870.
[64] R. Zeng, L. Xing, Y. Qiu, Y. Wang, W. Huang, W. Li, S. Yang, *Electrochim. Acta* **2014**, *146*, 447.
[65] M. Yao, S. Yamazaki, H. Senoh, T. Sakai, T. Kiyobayashi, *Mater. Sci. Eng. B* **2012**, *177*, 483.
[66] B. Tian, Z. Ding, G. H. Ning, W. Tang, C. Peng, B. Liu, J. Su, C. Su, K. P. Loh, *Chem. Commun.* **2017**, *53*, 2914.
[67] Z. Song, Y. Qian, T. Zhang, M. Otani, H. Zhou, *Adv. Sci.* **2015**, *2*, 1500124.
[68] D. Wu, Z. Xie, Z. Zhou, P. Shen, Z. Chen, *J. Mater. Chem. A* **2015**, *3*, 19137.
[69] W. Huang, Z. Zhu, L. Wang, S. Wang, H. Li, Z. Tao, J. Shi, L. Guan, J. Chen, *Angew. Chem. Int. Ed.* **2013**, *52*, 9162.
[70] Z. Zhu, M. Hong, D. Guo, J. Shi, Z. Tao, J. Chen, *J. Am. Chem. Soc.* **2014**, *136*, 16461.
[71] K. Liu, J. Zheng, G. Zhong, Y. Yang, *J. Mater. Chem.* **2011**, *21*, 4125.
[72] Z. Song, Y. Qian, X. Liu, T. Zhang, Y. Zhu, H. Yu, M. Otani, H. Zhou, *Energy Environ. Sci.* **2014**, *7*, 4077.
[73] J. Lee, M. J. Park, *Adv. Energy Mater.* **2017**, *7*, 1602279.
[74] Z. Song, Y. Qian, M. L. Gordin, D. Tang, T. Xu, M. Otani, H. Zhan, H. Zhou, D. Wang, *Angew. Chem., Int. Ed.* **2015**, *54*, 13947.
[75] A. Shimizu, Y. Tsujii, H. Kuramoto, T. Nokami, Y. Inatomi, N. Hojo, J. Yoshida, *Energy Technol.* **2014**, *2*, 155.
[76] T. Liu, C. K. Kim, B. Lee, Z. Chen, S. Noda, S. S. Jang, S. W. Lee, *Energy Environ. Sci.* **2017**, *10*, 205.
[77] H. Chen, M. Armand, G. Demailly, F. Dolhem, P. Poizot, J. M. Tarascon, *ChemSusChem* **2008**, *1*, 348.
[78] C. Peng, H. G. Ning, J. Su, G. Zhong, W. Tang, B. Tian, C. Su, D. Yu, L. Zu, J. Yang, F. M. Ng, S. Y. Hu, Y. Yang, M. Armand, K. P. Loh, *Nat. Energy* **2017**, *2*, 17074.
[79] Y. Hanyu, T. Sugimoto, Y. Ganbe, A. Masuda, I. Honma, *J. Electrochem. Soc.* **2014**, *161*, A6.
[80] A. Petronico, K. L. Bassett, B. G. Nicolau, A. A. Gewirth, R. G. Nuzzo, *Adv. Energy Mater.* **2018**, *8*, 1700960.
[81] S. Wang, Q. Wang, P. Shao, Y. Han, X. Gao, L. Ma, S. Yuan, X. Ma, J. Zhou, X. Feng, B. Wang, *J. Am. Chem. Soc.* **2017**, *139*, 4258.
[82] Y. Liang, Z. Chen, Y. Jing, Y. Rong, A. Facchetti, Y. Yao, *J. Am. Chem. Soc.* **2015**, *137*, 4956.
[83] X. Fan, F. Wang, X. Ji, R. Wang, T. Gao, S. Hou, J. Chen, T. Deng, X. Li, L. Chen, C. Luo, L. Wang, C. Wang, *Angew. Chem., Int. Ed.* **2018**, *57*, 7146.
[84] T. B. Schon, A. J. Tilley, E. L. Kynaston, D. S. Seferos, *ACS Appl. Mater. Interfaces* **2017**, *9*, 15631.
[85] M. E. Bhosale, K. Krishnamoorthy, *Chem. Mater.* **2015**, *27*, 2121.
[86] J. Hong, M. Lee, B. Lee, D. H. Seo, C. B. Park, K. Kang, *Nat. Commun.* **2014**, *5*, 5335.
[87] Y. Hanyu, Y. Ganbe, I. Honma, *J. Power Sources* **2013**, *221*, 186.
[88] T. Ma, Q. Zhao, J. Wang, Z. Pan, J. Chen, *Angew. Chem. Int. Ed.* **2016**, *55*, 6428.
[89] G. S. Vadehra, R. P. Maloney, M. A. Garcia-Garibay, B. Dunn, *Chem. Mater.* **2014**, *26*, 7151.
[90] C. Peng, G.-H. Ning, J. Su, G. Zhong, W. Tang, B. Tian, C. Su, D. Yu, L. Zu, J. Yang, M.-F. Ng, Y.-S. Hu, Y. Yang, M. Armand, K. P. Loh, *Nat. Energy* **2017**, *2*, 17074.
[91] Y. Igarashi, H. Takenaka, Y. Nakanishi-Ohno, M. Uemura, S. Ikeda, M. Okada, *J. Phys. Soc. Jpn.* **2018**, *87*, 044802.
[92] A. Kuhn, K. G. von Eschwege, J. Conradie, *J. Phys. Org. Chem.* **2012**, *25*, 58.
[93] R. B. Araujo, A. Banerjee, P. Panigrahi, L. Yang, M. Strømme, M. Sjödin, C. M. Araujo, R. Ahuja, *J. Mater. Chem. A* **2017**, *5*, 4430.
[94] P. Hu, X. He, M. Ng, J. Ye, C. Zhao, S. Wang, K. Tan, A. Chaturvedi, H. Jiang, C. Kloc, W. Hu, Y. Lon, *Angew. Chem. Int. Ed.* **2019**, *58*, 13513.
[95] K. Obinata, T. Nakayama, A. Ishikawa, K. Sodeyama, K. Nagata, Y. Igarashi, M. Okada, *Sci. Tech. Adv. Mater. Methods.* **2022**, *2*, 355.
[96] F. R. Bach, in *Proc. 25th Int. Conf. Machine Learning* **2008**.
[97] J. D. Lee, D. L. Sun, Y. Sun, J. E. Taylor, *Ann. Stat.* **2016**, *44*, 907.
[98] L. Bustillo, T. Laino, T. Rodrigues, *Chem. Sci.* **2023**, *14*, 10378.
[99] K. Hatakeyama-Sato, T. Tezuka, M. Umeki, K. Oyaizu, *J. Am. Chem. Soc.* **2020**, *142*, 3301.
[100] R. Gurnani, C. Kuenneth, A. Toland, R. Ramprasad, *Chem. Mater.* **2023**, *35*, 1560.
[101] X. Y. Yao, G. Chen, M. Pecht, B. Chen, *J. Energy Storage* **2023**, *50*, 106437.
[102] K. Q. Zhou, Y. Qin, C. Yuen, *J. Energy Storage* **2024**, *100*, 113502.
[103] T. Lin, S. Chen, S. J. Harris, T. Zhao, Y. Liu, J. Wan, *eScience* **2024**, *4*, 100280.

Manuscript received: April 17, 2025
Revised manuscript received: May 9, 2025
Version of record online: May 12, 2025