Data integration for multiple alkali metals in predicting coordination energies based on Bayesian inference

Koki Obinata\textsuperscript{a,b}, Tomofumi Nakayama\textsuperscript{a}, Atsushi Ishikawa\textsuperscript{a}, Keitaro Sodeyama\textsuperscript{a,e}, Kenji Nagata\textsuperscript{a,e}, Yasuhiko Igarashi\textsuperscript{a} and Masato Okada\textsuperscript{a,ag}

\textsuperscript{a}Graduate School of Science, The University of Tokyo, Bunkyo, Japan; \textsuperscript{b}Research Fellow of Japan Society for the Promotion of Science, Chiyoda, Japan; \textsuperscript{c}KARAKURI Inc., Chuo, Japan; \textsuperscript{d}Center for Green Research on Energy and Environmental Materials (GREEN), and International Center for Materials Nanoarchitectonics, National Institute for Materials Science (NIMS), Tsukuba, Japan; \textsuperscript{e}Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science (NIMS), Tsukuba, Japan; \textsuperscript{f}Graduate School of System and Information Engineering, University of Tsukuba, Tsukuba, Japan; \textsuperscript{g}Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Japan

**ABSTRACT**

Building machine learning models using a dataset calculated by first principles calculations is an important approach to explore the next-generation batteries. In previous studies, the rechargeable secondary battery dataset was constructed for Li ion by density functional theory (DFT) calculation, and after that, extended to five alkali metal ions. This dataset can be regarded as consisting of five alkali metal ion groups, and it is one of the interests to know which approach is preferred to build individual models specialized for each alkali metal ion or build a single model by integrating the datasets. We quantitatively evaluate the possibility of data integration in the framework of Bayesian model selection and show that the integration of datasets is suitable. In addition, extracting new knowledge using feature selection is also important in exploring next-generation batteries. In order to further advance the knowledge extraction, the reliability of the selected features should be considered to avoid misinterpretation. We evaluate the confidence level of feature selection by calculating the posterior probabilities of features using Bayesian model averaging (BMA). We found that the confidence level of feature selection increases when the number of data increases.

1. Introduction

Li-ion batteries are widely used in electric vehicles and stationary batteries. The electrolyte is an important component of these batteries, but its development is more difficult than the development of the electrodes because of the disordered structure of the liquid. The development of next-generation batteries using ions other than Li is progressing, but the predominance of Li ion batteries still continues \cite{1,2}.

A data-driven scientific approach to materials exploration using machine learning has been investigated to explore the next-generation batteries \cite{3–6}. Sodeyama et al. \cite{3} predicted the coordination energy of Li-ion to the solvent using machine learning as an
important measure of ion transport performance. They also extracted important features by feature selection. After that, Ishikawa et al. [4] calculated the coordination energies for five alkali metal ions, including Li, to construct a broader dataset, and compared several machine learning methods.

The rechargeable secondary battery dataset constructed by Ishikawa et al. consists of combinations of five alkali metal ions and 66 electrolyte solvents and can be regarded as consisting of ion-grouped datasets. It is of natural interest to compare an approach that builds individual models which specialize for each alkali metal ion and an approach that builds a single model for the entire dataset by integrating the datasets. Ishikawa et al. built only an integrated model for the dataset, and no comparison was made with the approach that builds models individually. Therefore, we quantitatively evaluate the possibility of integrating the dataset using the Bayesian model selection framework [7,8].

In exploring the next-generation batteries by data-driven science, it is also expected to extract new knowledge by machine learning. One of the important techniques is a feature selection. Machine learning models consisting of a small number of features selected by the feature selection are highly interpretable and support experts in considering the relationship between features and the target properties. Extracting knowledge using feature selection is widely studied in the field of materials informatics [3–14]. Since the results of feature selection become the basis of experts’ consideration, the reliability of the results is also important to avoid misinterpretation. In the previous battery researches [3,4], LASSO and an exhaustive search method have been used and the reliability of feature selection results has been evaluated qualitatively based on the weight diagram of the exhaustive search, but it has not been evaluated quantitatively. In this study, we evaluate the confidence level of feature selection by the marginal posterior probability of each feature based on Bayesian inference [15–17]. We calculate the marginal posterior probability based on the BMA framework [18]. Reliability of machine learning model is widely researched, including the field of materials exploration and materials informatics, and it helps researchers to properly and effectively use machine learning considering its risk [19–31].

Our main contributions are as follows.

(1) We evaluate the possibility of data integration based on Bayesian model selection, and we show that the integrated model is more suitable than the individual models.

(2) We evaluate the confidence level of the feature selection by Bayesian model averaging. We show that features considered to be important in previous researches have high confidence level, and the confidence level increases when the number of data increases.

The code used in this paper is available at https://github.com/okada-lab/exhbma.

2. Formulation

2.1. Bayesian linear regression and exhaustive search method

We formulate the linear regression model on the basis of Bayesian inference [32]. Each data is represented by a p-dimensional feature vector \(x = (x^{(1)}, \ldots, x^{(p)})^T \in \mathbb{R}^p\), and a target value is denoted as \(y \in \mathbb{R}\). When we observe \(N\) samples, the whole dataset is denoted as \(X = (x_1, \ldots, x_N)^T \in \mathbb{R}^{N \times p}\) and \(y = (y_1, \ldots, y_N)^T \in \mathbb{R}^N\).

For simplicity, we assume that the observed feature vectors and target values are centered and, in particular, the feature vectors are standardized; thus, the next relationships hold for the target value \(y\) and every \(i\)-th feature \(x^{(i)}\).

\[
\sum_{n=1}^{N} y_n = 0 \quad (1)
\]

\[
\sum_{n=1}^{N} x_n^{(i)} = 0 \quad (2)
\]

\[
\sum_{n=1}^{N} x_n^{(i)2} = 1 \quad (3)
\]

We do not assume the standardization of the target values considering the model selection discussed in Section 2.3.

Suppose that a true relationship is modeled as a linear function of the feature \(x\). When we observe the target value, an observation noise is always included. We model the target value \(y\) as the sum of the linear model and the observation noise \(\varepsilon\). The noise variable \(\varepsilon\) follows a normal distribution with a mean of 0 and a variance of \(\sigma^2\). This observation process is formulated as

\[
p(y) = \mathcal{N}(y|w_0 + x^Tw, \sigma^2). \quad (4)
\]

When \(N\) observed data are independent, the likelihood of the entire observation process is expressed as

\[
p(y|X, w_0, w, \sigma) = \mathcal{N}(y|w_01_N + Xw, \sigma^2I_N), \quad (5)
\]

where \(1_N\) denotes the \(N\)-dimensional vector with all elements being one and \(I_N\) denotes the \(N \times N\) identity matrix.

For the prior distribution of the coefficients \(w\), we set the normal distribution with a mean of 0 and a variance of \(\sigma^2_w\) for each coefficient independently.
\[ p(w|\sigma_w) = \mathcal{N}(w|0, \sigma_w^2 I_p). \]  

This prior distribution expresses the prior knowledge that the magnitude of the coefficients will be around \( \sigma_w \).

We set the following prior distribution for intercept \( w_0 \)
\[
p(w_0) = \mathcal{N}(w_0|0, \sigma_y^2) \tag{7}
\]
where \( \sigma_y^2 \) is the variance of the target value \( y \), not a hyperparameter like \( \sigma_c \) and \( \sigma_w \). Since we consider the centering preprocessing described in Equations (1) and (2), the mean value of the intercept can be considered as 0 [33]. We use the variance of the target value \( \sigma_y \), which is a data-dependent value, as a proxy for the scale of prior uncertainty of intercept \( w_0 \) [18]. We omit \( \sigma_y \) from the notation \( p(w_0) \) to avoid confusion that \( \sigma_y \) is a hyperparameter.

From Bayes’ theorem
\[
p(A|B) = \frac{p(B|A)p(A)}{p(B)}, \tag{8}
\]
the joint posterior distribution of intercept \( w_0 \) and coefficients \( w \) given by data \( X \) and \( y \) is calculated as
\[
p(w_0, w|X, y, \sigma_c, \sigma_w) = \frac{p(y|X, w_0, w, \sigma_c)p(w|\sigma_w)p(w_0)}{p(y|X, \sigma_c, \sigma_w)} \tag{9}
\]
where the denominator is the likelihood of the prediction model with fixed values of \( \sigma_c \) and \( \sigma_w \)
\[
p(y|X, \sigma_c, \sigma_w) = \int p(y|X, w_0, w, \sigma_c)p(w|\sigma_w)p(w_0)dwdw_0 \tag{10}
\]
\[
= \mathcal{N}(y|0, \sigma_c^2 I_N + \sigma_w^2 X X^T) \sqrt{\frac{\sigma_c^2}{N \sigma_c^2 + \sigma_w^2}} \tag{11}
\]
Following the maximum a posteriori (MAP) estimation, the intercept \( w_0 \) and coefficients \( w \) are estimated by maximizing the posterior distribution expressed as Equation (9); in this case, this estimation method is equivalent to the Ridge regression [33]. By using the Equations (5)–(7), the posterior distribution of the intercept \( w_0 \) and coefficients \( w \) is calculated as
\[
p(w_0, w|X, y, \sigma_c, \sigma_w) = \mathcal{N}(w_0|0, (N/\sigma_c^2 + 1/\sigma_w^2)^{-1}), \mathcal{N}(w|\mu, \Lambda) \tag{12}
\]
\[
\Lambda = \frac{1}{\sigma_c^2} X^T X + \frac{1}{\sigma_w^2} I_p \tag{13}
\]
\[
\mu = \Lambda^{-1} \frac{1}{\sigma_c^2} X^T y \tag{14}
\]
where \( \Lambda \) is the precision matrix and \( \mu \) is the mean vector. From this result, the MAP solution for the intercept \( w_0 \) is 0 and that for the coefficients \( w \) is \( \mu \).

We considered the prior distribution for the hyper-parameters \( \sigma_c \) and \( \sigma_w \) and the final likelihood of the model is calculated as
\[
p(y|X) = \int d\sigma_c \sigma_w p(\sigma_c, \sigma_w) N(y|0, \sigma_c^2 I_N + \sigma_w^2 X X^T) \sqrt{\frac{\sigma_c^2}{N \sigma_c^2 + \sigma_w^2}} \tag{15}
\]
where \( p(\sigma_c, \sigma_w) \) is properly defined at the interval \((0, \infty)\). This marginalization over \( \sigma_c \) and \( \sigma_w \) cannot be performed analytically, so a numerical integration is performed. Note that \( \sigma_y \) in Equation (7) is not a hyperparameter, so marginalization is not performed.

We have used all the \( p \) features, but we can consider submodels by choosing which features to include in the model and compare their performance. To specify a submodel, we introduce an indicator vector \( c \) as
\[
c = (c_1, \ldots, c_p) \in \{0, 1\}^p. \tag{16}
\]
The \( i \)-th feature \( x^{(i)} \) is included in the model when \( c_i = 1 \) and is not included when \( c_i = 0 \). By introducing a submatrix \( X_t \) that extracts column \( i \) from \( X \) where \( c_i = 1 \) and substituting \( X \) with \( X_t \) in Equation (15), we can obtain the marginal likelihood for \( p(y|c, X) \).

An exhaustive search method, which compares all \( 2^p \) models, is often carried out to select the model with the best evaluation metrics [3–36]. In the previous study by Ishikawa et al. [4], for example, the cross-validation error [33] is used as the evaluation metric. A null model, a constant prediction model where no feature is considered, is sometimes excluded from the search set depending on the analysis. However, since we do not have strong support for excluding the null model, we include the null model in the following formulation.

When we apply the exhaustive search method to the Bayesian linear regression model formulated above, we can compare models by the posterior probability of the model
\[
p(c|X, y) = \frac{p(y|c, X)p(c)}{p(y|X)}, \tag{17}
\]
where the denominator
\[
p(y|X) = \sum_c p(y|c, X)p(c) \tag{18}
\]
is a likelihood for the entire model set. We use \( \sum_c (\cdot) \) to express the summation over all combinations of the indicator vector. For the prior distribution of the model specified by the indicator vector, \( p(c) \), we set the Bernoulli distribution for each element of the indicator vector as
\[ p(c|\alpha) = \prod_{j=1}^{P} \alpha^c (1 - \alpha)^{1-c}, \quad 0 < \alpha < 1. \]  

The hyperparameter \( \alpha \) represents the proportion that each feature is included in the model in the prior distribution, and the expected number of features included in the model is \( \mathbb{E}[\sum_i c_i] = \alpha P \). In this paper, we fix \( \alpha = 0.5 \) and omit \( \alpha \) from the notation of Equation (19). With this setting, all the models have equal prior probabilities.

In the exhaustive search method, it is also possible to evaluate the importance of a feature using a weight diagram, which plots the values of the coefficients arranged in the order of model performance [3–5].

2.2. BMA and feature selection

In this study, we consider the uncertainty of each submodel specified by the indicator vector \( c \) and introduce a method of quantitatively evaluating the confidence level of feature selection using a weighted average of the model posterior probabilities, which is called BMA [18]. The overview of this method is shown in Figure 1. This method makes it possible to evaluate the confidence level of feature selection and quantify the importance evaluation of features, which has been a qualitative one when using the weight diagram of the exhaustive search method.

For a random variable \( r \) that one wants to predict, such as the target value \( y \), the coefficient of the feature \( w \), and the element of the indicator vector \( c \), the prediction by BMA is formulated as

\[ p(r|X, y) = \sum_c p(r|c, X, y)p(c|X, y). \]  

The summation over all combinations of indicator vectors can be calculated using the result of the exhaustive search. Therefore, we refer to the prediction with this form as ES-BMA.

When applied to the prediction of the target value \( y \) for new data \( x \), Equation (20) is expressed as

\[ p(y|x, X, y) = \sum_c p(y|x, c, X, y)p(c|X, y). \]  

This prediction by BMA integrates the predictions of the submodels considering their uncertainty \( p(c|X, y) \). The prediction by BMA is superior to the prediction using any single model specified by \( c \) in predicting new data [18].

When applied to feature selection, we estimate the probability that the \( i \)-th feature is included in the model by marginalizing over the model with \( c_i = 1 \):

\[ p(c_i = 1|X, y) = \sum_{c_{-i}} p(c_i = 1|c_{-i}, X, y)p(c_{-i}|X, y) \]

\[ = \sum_{\{c_i = 1\}} p(c|X, y), \]  

where \( c_{-i} \) is an indicator vector excluding the \( c_i \) element. This posterior probability is well calibrated to the precision of feature selection [15], and we use this probability value as the confidence level of feature selection.

Here, we briefly explain the confidence evaluation for feature selection in contrast to regression and classification. In the case of regression, we can evaluate the confidence of the target variable prediction by estimating the confidence interval [19–21]. In the case of classification, when using a discriminant function such as Fisher’s linear discriminant, confidence evaluation is not performed. However, we can evaluate the confidence level of the estimated class by the model’s output probability value using a discriminative model that models the class posterior probability, such as logistic

![Figure 1. Overview of Bayesian model averaging and exhaustive search method.](image-url)
regression [32]. The case of feature selection is similar to classification. When features are selected by parameter optimization, such as in LASSO, confidence evaluation is not performed. However, we can evaluate the confidence level of the selection result by the posterior probability value based on Bayesian inference [15–17].

In addition to model interpretability, feature selection also has the advantage of reducing the computational cost. The prediction by BMA using Equation (21) requires all the features and all the submodels to be used in the prediction phase, and this is computationally expensive. To make a better prediction with a single model using fewer features, features whose posterior probability calculated by Equation (23) is above 0.5, not features that are in the model with the highest posterior probability \( p(c|X, y) \), should be selected [37].

### 2.3. Model selection for different model building approaches

When we have multiple approaches in building a model and want to choose the best one, model selection is performed. In Bayesian inference, model selection is performed based on the negative log likelihood of the model, also known as the Bayesian free energy [7,8]. For simplicity, we use the log likelihood for model comparison and choose the approach with the larger log likelihood.

In this paper, we consider a case where a dataset is regarded as consisting of subgroup(s) and we want to compare an approach that builds individual models for subgroup(s) and an approach that builds a single model for the entire dataset by integrating subgroup(s). In general, as the number of data increases, the model accuracy and the stability of the learning process increases; therefore, integrating datasets can improve the model accuracy. On the other hand, integrating datasets with significantly different mechanisms may negatively affect model learning. Therefore, we need to compare these two approaches.

Let \( k \) be a subscript representing a subgroup(s). The dataset for subgroup(s) \( k \) is denoted as \( X^{(k)}, y^{(k)} \), the set of parameters in the corresponding model is denoted as \((w^{(k)}, w_0^{(k)}, \sigma_x^{(k)}, \sigma_w^{(k)})\), and the indicator vector for the model is denoted as \( c^{(k)} \). For simplicity, we use \( \theta^{(k)} = (w^{(k)}, w_0^{(k)}, \sigma_x^{(k)}, \sigma_w^{(k)}) \) in the following formulations. From Equation (18), the likelihood of the model is calculated as

\[
    p(y^{(k)}|X^{(k)}) = \sum_{c^{(k)}} \int d\theta^{(k)} p(y^{(k)}, \theta^{(k)}|c^{(k)}, X^{(k)}) p(c^{(k)}).
\]

When considering individual models for all subgroups, the likelihood is

\[
    p_{\text{individual}}(y|X) = \prod_k p(y^{(k)}|X^{(k)}).
\]

Here, we assume that the models for each subgroup are independent.

On the other hand, if all datasets are integrated and considered as a single dataset, then the likelihood is calculated by replacing \( k \)-dependent parameters in Equation (24),

\[
    p_{\text{integrated}}(y|X) = \sum_{c^{\text{all}}} \int d\theta^{\text{all}} p(y, \theta^{\text{all}}|c^{\text{all}}, X)p(c^{\text{all}}).
\]

Model selection between the individual models approach and the integrated model approach is performed by comparing Equations (25) and (26).

Since Equations (25) and (26) are probability distributions of the target variable \( y \), they are affected by the scaling operation on \( y \). For example, when \( y \) is multiplied by \( a \), \( p(y|X) \) changes by a factor of \( 1/a \). If we perform a standardization operation on \( y \), the individual and integrated models will have different variable transformations; therefore, the comparison of probability distributions will not be correct. To prevent this effect, only the centering operation is performed on the target variable \( y \), and the standardizing operation is not performed.

### 3. Results and discussion

#### 3.1. Dataset

We use the rechargeable secondary battery dataset used by Ishikawa et al. [4] to show the effectiveness of the ES-BMA. In this dataset, coordination energy is used as the target property since it can be considered as a good indicator for battery performance. Coordination energy is obtained by DFT calculation. The coordination energy dataset consists of 330 data, which are combinations of five alkali metal ions (Li, Na, K, Rb, and Cs) and 66 electrolyte solvents. Seventy electrolyte solvents were used in their study, but four were duplicated, so we exclude them. The feature vector consists of 13 elements, three of which are derived from alkali metal ions and 10 from electrolyte solvents. Of the 13 features, six are also obtained by DFT calculation.

#### 3.2. Experimental setting

As mentioned in Sect. 2.2, we use a Bernoulli distribution, Equation (19), with the hyperparameter \( \alpha = 0.5 \) for the prior distribution of the model specified by the indicator vector.

To integrate over the hyperparameters \( \sigma_x \) and \( \sigma_w \) in the likelihood Equation (15), numerical integration is
performed using the values calculated for the grid within a finite range.

For the prior distribution of the hyperparameters $\sigma_e$ and $\sigma_w$, we use a distribution $p(x) \propto x^{-1}$ defined within the finite range, which is a noninformative prior for a scale-invariant parameter [32], to reflect the case that we have little prior knowledge. Prior distribution $p(x) \propto x^{-1}$ itself is an improper prior, which cannot be normalized at the interval $(0, \infty)$. However, it can be expressed asymptotically by a proper distribution, for example, the gamma distribution $g(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1}e^{-x/\theta}$ with shape parameter $k \to 0$ and scale parameter $\theta \to \infty$.

### 3.3. Model selection between individual models and an integrated model

The rechargeable secondary battery dataset consists of combinations of five alkali metal ions and 66 electrolyte solvents and can be regarded as consisting of ion-grouped datasets. Therefore, it is of natural interest to compare an approach that builds individual models which specialize for each alkali metal ion and an approach that builds a single model for the entire dataset by integrating the datasets.

By applying the model selection framework shown in Sect. 2.3, we compare the two approaches. In this case, subgroup index $k$ introduced in Sect. 2.3 corresponds to an alkali metal ion $(k = \text{Li, Na, K, Rb, Cs})$. Table 1 shows the likelihoods of the individual models and the integrated model. The likelihoods of each alkali metal ion model are also listed for the individual models. From this result, we can integrate all the datasets and regard that they follow the identical distribution.

### Table 1. Likelihood of the individual models and the integrated model.

| model            | log$p(y|x)$ |
|------------------|-------------|
| Li               | 0.31        |
| Na               | 16.4        |
| K                | 26.3        |
| Rb               | 31.2        |
| Cs               | 36.3        |
| Individual Models| 110.5       |
| Integrated Model | 165.8       |

Figure 3 shows the posterior probabilities of features calculated for the integrated model. The posterior probabilities are close to 0 or 1, indicating that the results of feature selection are clearer than those of individual models. It can be considered that the confidence level of feature selection increases with the number of data compared with the individual models.

Features with a posterior probability greater than 0.5 were also selected in the previous study by Ishikawa et al. [4], who used an exhaustive search method.

Figure 4 shows the parity plot between coordination energy predicted by ES-BMA (Equation (21)) and that obtained by DFT calculation. Data which are assigned to test set in each split of 10-fold cross-validation are plotted. The accuracy obtained by the 10-fold cross-validation is 0.131 eV. This accuracy is comparable to that obtained by Ishikawa et al. [4], which directly optimizes the cross-validation error.

Figure 5 shows how the posterior probability of each feature changes for all the data integration patterns. The horizontal axis is the number of integrated alkali metal ion dataset, and the vertical axis is the posterior probability value. The posterior probabilities of the NBO charge of the O atom and the total dipole are close to 1 early in the data integration process. In previous studies, these two features are also considered particularly important [4]. The evaluation of posterior probabilities may contribute to the early identification of essential features in the data collection process.

As we integrate datasets, the posterior probability becomes closer to 0 and 1, and the confidence level becomes clearer. In particular, the lowest unoccupied molecular orbital (LUMO) energy, which has low importance in the individual models, becomes more important as we integrate datasets. Based on model selection result, we can regard that integrated datasets follow the identical distribution, so we consider that this improvement in the confidence level is due to the increase in the number of data.

### 3.4. Application of ES-BMA

From the result of model selection, the integrated model is the main target of analysis. However, to elucidate the effect of data integration, we compare results of both approaches.

Figure 2 shows the posterior probabilities of features for individual models. The posterior probabilities of the natural bond orbital (NBO) charge of the O atom and the total dipole are close to 1, and those of the highest occupied molecular orbital (HOMO) energy and melting point are close to 0, so the selection result for these features can be trusted with high confidence. This result is consistent with that of a previous study in which the NBO charge of the O atom and the total dipole are strongly related to the coordination energy [4]. On the other hand, the posterior probabilities of total energy and density are approximately 0.5, showing that the confidence levels of the selection result for these features are low.

### 3.5. Posterior distribution of hyperparameters

To show that the finite range of the prior distribution for the hyperparameters $\sigma_e$ and $\sigma_w$ is appropriate for numerical integration, we examine the posterior distribution of these two parameters. From Bayes’ theorem, the posterior distribution is obtained by
normalizing the product of the likelihood of the two parameters calculated by Equation (11) and the prior distribution $p(\sigma_e, \sigma_w)$. The result is shown in Figure 6, which reveals that the peak of the posterior distribution is well included in the region defined as the prior distribution. By comparing the posterior distribution of the individual model and that of the integrated model, we find that the peak is sharper in...
the integrated model owing to the increase in the number of data.

For other applications of the posterior distribution of the hyperparameters, it is possible to determine whether the model accuracy has reached the expected level by comparing the distribution of the noise parameters with the accuracy of the target value guaranteed by the experimental or DFT calculation results.
4. Conclusions

In this paper, we introduced a framework for verifying data integration and a feature selection method with confidence level evaluation using the rechargeable secondary battery dataset. This dataset can be regarded as groups consisting of five alkali metal ions. Hence, it is one of the interests which approach is preferred to build individual models which specialize for each alkali metal ion or build a single model by integrating the datasets. We showed the validity of building a single model by performing a model selection in the framework of the Bayesian inference. The results provide quantitative guidance on whether we should build a single integrated model or individual models when multiple datasets are available. Note that in this paper, we performed data integration only within simulation datasets. Recently, data integration of experiment and simulation datasets has been studied [38], so for our future work, we plan to validate the integration of datasets from both simulation and experimental data.

In exploring the next-generation batteries, knowledge extraction through feature selection is also important. In advancing the discussion based on feature selection, it is necessary to consider the reliability of the feature selection results. We evaluated the confidence level of feature selection by the marginal posterior probability of each feature. In the ES-BMA, we use Bayesian model averaging and calculate the posterior probability of features considering the uncertainty of models searched by the exhaustive search. By applying the ES-BMA, selection result of the previous study is confirmed to be highly confident. It is also shown that the confidence level becomes higher by integrating the datasets. In particular, LUMO energy, which has a low posterior probability in individual models, proves to be important by integrating the datasets.

In materials informatics, the number of data is often small. Therefore, data integration framework and uncertainty evaluation method of feature selection becomes important since the former improves model performance by increasing data and the latter helps experts to extract new knowledge considering the data fluctuations due to small data.

One of the challenges of the ES-BMA is that it is computationally expensive for both learning and inference since it involves an exhaustive search and numerical integration. In particular, the computational cost of an exhaustive search increases exponentially with respect to the number of features; thus, it is often computationally prohibitive for real data. To overcome this problem, a sampling approach such as a Monte Carlo method can be used to estimate the result of ES-BMA in a statistically unbiased manner [18,35].

Disclosure statement

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

Funding

This work was supported by JSPS KAKENHI under Grant Number [JP20K21749], and CREST [JPMJCR1761, JPMJCR2101] and PRESTO [JPMJPR17N2] from the Japan Science and Technology Agency (JST).

ORCID

Keitaro Sodeyama @ http://orcid.org/0000-0002-9228-0729
Kenji Nagata @ http://orcid.org/0000-0001-9894-4461
Yasuhiro Igarashi @ http://orcid.org/0000-0003-1042-6657
Masato Okada @ http://orcid.org/0000-0002-9040-8784

Data availability statement

The data that support the findings of this study are openly available in the Materials Data Repository (MDR) at https://doi.org/10.48505/nims.3057. Ionic radii of alkali metal ions are 0.9 (Li), 1.16 (Na), 1.52 (K), 1.66 (Rb), and 1.81 (Cs), which are not included in the data.
References

[1] Xu K. Nonaqueous liquid electrolytes for lithium-based rechargeable batteries. Chem Rev. 2004;104(10):4303–4418.
[2] Goodenough JB, Kim Y. Challenges for rechargeable Li batteries. Chem Mater. 2010;22(3):587–603.
[3] Sodeyama K, Igarashi Y, Nakayama T, et al. Liquid electrolyte informatics using an exhaustive search with linear regression. Phys Chem Chem Phys. 2018;20(35):22585–22591. DOI:10.1039/C7CP08280K
[4] Ishikawa A, Sodeyama K, Igarashi Y, et al. Machine learning prediction of coordination energies for alkali group elements in battery electrolyte solvents. Phys Chem Chem Phys. 2019;21(48):26399–26405. DOI:10.1039/C9CP03679B
[5] Nakayama T, Igarashi Y, Sodeyama K, et al. Material search for Li-ion battery electrolytes through an exhaustive search with a gaussian process. Chem Phys Lett. 2019;731:136622.
[6] Liu Y, Guo B, Zou X, et al. Machine learning assisted materials design and discovery for rechargeable batteries. Energy Stor Mater. 2020;31:434–450.
[7] Akaike H. Likelihood and the bayes procedure. In: Parzen E, Tanabe K, Kitagawa G. Selected papers of hirotugu akaike. New York: Springer; 1998. p. 309–332.
[8] Nagata K, Sugiya S, Okada M. Bayesian spectral deconvolution with the exchange monte carlo method. Neural Netw. 2012;28:82–89.
[9] Ghiringhelli LM, Vybiral J, Levchenko SV, et al. Big data of materials science: critical role of the descriptor. Phys Rev Lett. 2015;114(10):105503. DOI:10.1103/PhysRevLett.114.105503
[10] Pilania G, Mannodi-Kanakkithodi A, Ueberuga B, et al. Machine learning bandgaps of double perovskites. Sci Rep. 2016;6(1):1–10. DOI:10.1038/srep19375
[11] Janet JP, Kulik HJ. Resolving transition metal chemical space: feature selection for machine learning and structure–property relationships. J Phys Chem A. 2017;121(46):8939–8954.
[12] Ouyang R, Curtarolo S, Ahmetcik E, et al. Sisso: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. Phys Rev Mater. 2018;2(8):083802. DOI:10.1103/PhysRevMaterials.2.083802
[13] Yada A, Nagata K, Ando Y, et al. Machine learning approach for prediction of reaction yield with simulated catalyst parameters. Chem Lett. 2018;47(3):284–287. DOI:10.1246/cl.171130
[14] Mangal A, Holm EA. A comparative study of feature selection methods for stress hotspot classification in materials. Integr Mater Manuf Innov. 2018;7(3):87–95.
[15] Viallefont V, Raftery AE, Richardson S. Variable selection and bayesian model averaging in case-control studies. Stat Med. 2001;20(21):3215–3230.
[16] Hans C. Model uncertainty and variable selection in bayesian lasso regression. Stat Comput. 2010;20(2):221–229.
[17] Forte A, Garcia-Donato G, Sted M, Methods and tools for bayesian variable selection and model averaging in normal linear regression. Int Stat Rev. 2018;86(2):237–258.
[18] Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. J Am Stat Assoc. 1997;92(437):179–191.
[19] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv Neural Inf Process Syst. 2017;30.
[20] Gal Y, Hron J, Kendall A. Concrete dropout. Adv Neural Inf Process Syst. 2017;30.
[21] Kuleshov V, Fenner N, Ermon S. Accurate uncertainties for deep learning using calibrated regression. Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden. PMLR; 2018. p. 2796–2804.
[22] Platt J, et al., Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv in large margin classif. 1999;10(3):61–74.
[23] Guo C, Pleiss G, Sun Y, et al. On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR; 2017. p. 1321–1330.
[24] Peterson AA, Christensen R, Khoshidi A. Addressing uncertainty in atomistic machine learning. Phys Chem Chem Phys. 2017;19(18):10978–10985.
[25] Kaikkoura B, Gallagher B, Kim S, et al. Reliable and explainable machine-learning methods for accelerated material discovery. Npj Comput Mater. 2019;5(1):1–9. DOI:10.1038/s41524-019-0248-2
[26] Musil F, Willatt MJ, Langovoy MA, et al. Fast and accurate uncertainty estimation in chemical machine learning. J Chem Theory Comput. 2019;15(2):906–915. DOI:10.1021/acs.jctc.8b00959
[27] Janet JP, Duan C, Yang T, et al. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. Chem Sci. 2019;10(34):7913–7922. DOI:10.1039/C9SC02298H
[28] Tran K, Neiswanger W, Yoon J, et al. Methods for comparing uncertainty quantifications for material property predictions. Mach Learn Sci Technol. 2020;1(2):025006. DOI:10.1088/2632-2153/ab7c1a
[29] Scalía G, Grambow CA, Pernici B, et al. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. J Chem Inf Model. 2020;60(6):2697–2717. DOI:10.1021/acs.jcim.9b00975
[30] Tavazza F, DeCost B, Choudhary K. Uncertainty prediction for machine learning models of material properties. ACS Omega. 2021;6(48):32431–32440.
[31] Zhang J, Kaikkoura B, Han TY. Leveraging uncertainty from deep learning for trustworthy material discovery workflows. ACS Omega. 2021;6(19):12711–12721.
[32] Bishop CM, Nasrabadi NM. Pattern recognition and machine learning. Vol. 4. New York: Springer; 2006.
[33] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2009.
[34] Igarashi Y, Ichikawa H, Nakanshi-Ohno Y, et al. Es-dos: exhaustve search and density-of-states estimation as a general framework for sparse variable selection. Proc J Phys. 2018;1056012001. IOP Publishing.
[35] Igarashi Y, Takenaka H, Nakanshi-Ohno Y, et al. Exhaustive search for sparse variable selection in linear regression. J Phys Soc Jpn. 2018;87(4):044802. DOI:10.7566/JPSJ.87.044802
[36] Nagata K, Kitazono J, Nakajima S, et al. An exhaustive search and stability of sparse estimation for feature selection problem. IPSJ Online Trans. 2015;8:25–32.
[37] Barbieri MM, Berger JO. Optimal predictive model selection. Ann Stat. 2004;32(3):870–897.
[38] Harashima Y, Tamai K, Doi S, et al. Data assimilation method for experimental and first-principles data: finite-temperature magnetization of (nd, pr, la, ce) 2 (fe, co, ni) 14 b. Phys Rev Mater. 2021;5(1):013806. DOI:10.1103/PhysRevMaterials.5.013806