Liquid electrolyte informatics using an exhaustive search with linear regression†

Keitaro Sodeyama,†abc Yasuhiko Igarashi,†abd Tomofumi Nakayama,†d Yoshitaka Tateyama†ace and Masato Okada†ad

Exploring new liquid electrolyte materials is a fundamental target for developing new high-performance lithium-ion batteries. In contrast to solid materials, disordered liquid solution properties have been less studied by data-driven information techniques. Here, we examined the estimation accuracy and efficiency of three information techniques, multiple linear regression (MLR), least absolute shrinkage and selection operator (LASSO), and exhaustive search with linear regression (ES-LIR), by using coordination energy and melting point as test liquid properties. We then confirmed that ES-LIR gives the most accurate estimation among the techniques. We also found that ES-LIR can provide the relationship between the “prediction accuracy” and “calculation cost” of the properties via a weight diagram of descriptors. This technique makes it possible to choose the balance of the “accuracy” and “cost” when the search of a huge amount of new materials was carried out.

1. Introduction

Computational material design with a data-driven information technique has become popular for materials research recently.1 The materials for next-generation lithium-ion batteries (LIBs) are the representative targets. Future LIBs require a higher voltage, a higher capacity, and a longer cycle life and need to be safer.2,3 For such properties, a variety of new “electrode” materials have been reported.4–6 However, new “electrolyte” materials, typically consisting of liquid solvents and Li-salts, have not appeared since 1991 for commercial use. This is because the search for liquid materials is more difficult compared to that for solid materials due to the disordered structure of liquid. Exploring new liquid materials with desirable properties is a challenging issue.7–9

In order to discover new liquid electrolytes with desirable properties, virtual screening with a data-driven information technique is one possible option. In this screening, a database of the features of materials called descriptors is first constructed with data from first-principles calculations or molecular dynamics simulations and/or experiments. Next, we determine the estimation rule (fitting equation) to predict the target properties based on the selected descriptors in the database by using the information techniques. Finally, we handle a huge number of candidate materials under the rule. Several applications of virtual screening to explore new LIB materials have been reported, though most of them are limited to solid materials research.10–13 Only a few applications have been reported for the liquid materials.14–16

To extract the estimation rule for predicting the target properties, we have to select descriptors using data-driven techniques. It is called the variable selection problem. In general, multiple linear regression (MLR),17 in which all the descriptors are used for the estimation, is the most standard treatment for the estimation of the properties of materials. However, irrelevant and redundant descriptors from data do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model. Thus, we have to remove these descriptors. Moreover, fewer descriptors are desirable because it reduces the complexity of the model, and a simpler model is simpler to understand and explain.

When there are $N$ explanatory variables, the simplest variable selection method is a search for all combinations of the variables which requires $2^N - 1 = x_1C_1 + x_2C_2 + \ldots + x_NC_N$ times of estimations.18 We called this naïve method the exhaustive search (ES) method.19–21 Although the ES method comes at the expense of computational complexity of at least $O(2^N)$, we can
use the ES method within the compass of \( N = 30 \), and the ES method can select the best descriptors for predicting the target properties. In this study, we apply the ES method for linear regression and propose a set of descriptor combinations that can produce better estimations. For comparison, we also apply least absolute shrinkage and selection operator (LASSO)\(^{22} \) using an L1-norm regularization term as a standard approximate method for the sparse variable selection, for which the computational complexity is \( O(N^3) \).

In the search for LIB liquid electrolytes, the evaluation of the properties of ion transport and electrochemical stability is indispensable. For the transport, solvation and desolvation from Li-ions at the electrolyte/electrode interface plays a crucial role, and thus the coordination energy of the solvent to Li-ions is an important measure. In order to keep the liquid state for the fast Li-ion transport, the melting point of the electrolyte is also a fundamental property. For the electrochemical stability, the quantities such as ionization potential and electron affinity are significant. Here, however, we focus on the quantities related to the Li-ion transport as the first target.

In this study, we investigated the estimation accuracy of the MLR, LASSO, and ES-LiR techniques in the search for liquid electrolyte materials. We estimated the coordination energies and melting points as the required properties of the LIB liquid electrolytes and discussed the extracted descriptors by LASSO and ES with linear regression (ES-LiR). The strategy of the ES-LiR method will be useful and applicable in the search for liquid electrolytes with other desired properties.

2. Computational details

2.1. Database

To predict novel LIB liquid electrolytes with desired properties by the information techniques, we constructed a database of known liquid electrolytes. We selected 103 solvent molecules which were commercialized as battery grade materials from KISHIDA Chemical Co., Ltd.\(^{23} \) We adopted the values of melting point, boiling point, flash point, density of solvent, and molecular weight from the catalogue data. Representative solvent molecules are shown in Scheme 1 and the complete list is shown in Scheme S1 of the ESI.\(^{†} \)

2.2. Cluster model calculations

To make the database of the electrolytes more substantial, we added the following values obtained by density functional theory (DFT) calculations of the molecular systems using the Gaussian 09 code:\(^{24} \) the coordination energy between a Li-ion and a solvent molecule, the Mulliken charge of the atom (typically oxygen atom) that is coordinated to a Li-ion, the distance between a Li-ion and the coordinated atom (typically Li-O distance) \( R(\text{Li-O}) \), the HOMO energy, the LUMO energy, and the dipole moment values of the 103 solvent molecules. The calculated data of the representative solvent molecules are shown in Table 1, and the complete data are listed in Table S1 in the ESI.\(^{†} \) The coordination energies \( E_{\text{coord}} \) are evaluated by the difference between the “total energy of a Li–solvent complex” and “the total energies of a solvent molecule and that of a Li-ion”

\[
E_{\text{coord}} = E(\text{Li–solvent}) - \{E(\text{solvent}) + E(\text{Li-ion})\}.
\]

We adopted the B3LYP functional\(^{25} \) with cc-pVDZ basis sets.\(^{26} \) The Mulliken charges and the dipole moments are obtained from the DFT calculations of pure solvent molecules without Li-ions. Geometry optimizations of the Li–solvent complexes and the pure solvent molecules were also carried out. In this study, totally 10 descriptors (explanation variables) were adopted for the database. There are several missing data in the catalogue. We omitted them for the prediction. When the data have no specific value but a range of values, we averaged them.

2.3. Data-driven information techniques

We applied the data-driven information techniques of MLR, LASSO, and ES-LiR to the electrolyte materials search. MLR is a typical supervised machine learning technique to predict certain values of the properties. The method tries to represent the relationship between the set of the given values of the properties, called explanation variables, and the target values for the prediction, called dependent variable, by constructing a model of the linear equation. We set a target value and an \( i \)-th explanation variable as \( z \) and \( x_i \) (\( i = 1, \ldots, 10 \)), respectively. We then assume that the relationship between them is linear and derive it from minimizing eqn (1),

\[
E = \sum_{\mu=1}^{103} \left( z^\mu - \sum_{i=1}^{10} w_i x_i^\mu \right)^2.
\]
This article is licensed under a Creative Commons Attribution 3.0 Unported License.

Table 1  Calculated values of the coordination energy ($E_{\text{coord}}$), the HOMO energy, the LUMO energy, the dipole moment, the Mulliken charge of the oxygen (nitrogen) atom, and the distance between the Li-ion and the oxygen (nitrogen) atom ($R(\text{Li-O})$) of 25 solvent molecules for the database.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Solvent name</th>
<th>Chemical formula</th>
<th>$E_{\text{coord}}$ (kcal mol$^{-1}$)</th>
<th>HOMO (eV)</th>
<th>LUMO (eV)</th>
<th>Dipole moment (Debye)</th>
<th>Mulliken charge</th>
<th>$R(\text{Li-O})$ (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>Propylene carbonate</td>
<td>C$_3$H$_6$O$_2$</td>
<td>−57.4</td>
<td>−7.93</td>
<td>0.946</td>
<td>5.255</td>
<td>−0.243</td>
<td>1.747</td>
</tr>
<tr>
<td>EC</td>
<td>Ethylene carbonate</td>
<td>C$_2$H$_4$O$_2$</td>
<td>−55.9</td>
<td>−8.017</td>
<td>0.919</td>
<td>5.07</td>
<td>−0.24</td>
<td>1.752</td>
</tr>
<tr>
<td>VC</td>
<td>Vinylene carbonate</td>
<td>C$_3$H$_6$O$_2$</td>
<td>−51.7</td>
<td>−6.973</td>
<td>−0.137</td>
<td>4.365</td>
<td>−0.231</td>
<td>1.76</td>
</tr>
<tr>
<td>FEC</td>
<td>Fluoroethylene carbonate</td>
<td>C$_3$H$_4$O$_2$F</td>
<td>−51.2</td>
<td>−8.468</td>
<td>0.493</td>
<td>4.487</td>
<td>−0.222</td>
<td>1.763</td>
</tr>
<tr>
<td>DMC</td>
<td>Dimethil carbonate</td>
<td>C$_2$H$_6$O$_2$</td>
<td>−50.0</td>
<td>−7.774</td>
<td>1.115</td>
<td>0.342</td>
<td>−0.306</td>
<td>1.747</td>
</tr>
<tr>
<td>DEC</td>
<td>Diethyl carbonate</td>
<td>C$_3$H$_6$O$_2$</td>
<td>−52.6</td>
<td>−7.654</td>
<td>1.217</td>
<td>0.613</td>
<td>−0.308</td>
<td>1.74</td>
</tr>
<tr>
<td>EMC</td>
<td>Ethyl methyl carbonate</td>
<td>C$_2$H$_5$O$_2$</td>
<td>−51.3</td>
<td>−7.713</td>
<td>1.168</td>
<td>0.514</td>
<td>−0.307</td>
<td>1.744</td>
</tr>
<tr>
<td>DAC</td>
<td>Dialkyl carbonate</td>
<td>C$_3$H$_6$O$_2$</td>
<td>−31.7</td>
<td>−4.191</td>
<td>0.238</td>
<td>0.494</td>
<td>−0.306</td>
<td>1.76</td>
</tr>
<tr>
<td>Furan</td>
<td>Furan</td>
<td>C$_2$H$_4$O</td>
<td>−48.7</td>
<td>−6.265</td>
<td>0.296</td>
<td>0.511</td>
<td>−0.17</td>
<td>1.866</td>
</tr>
<tr>
<td>THF</td>
<td>Tetrahydrofuran</td>
<td>C$_2$H$_4$O</td>
<td>−47.2</td>
<td>−6.832</td>
<td>1.38</td>
<td>1.434</td>
<td>−0.323</td>
<td>1.808</td>
</tr>
<tr>
<td>THP</td>
<td>Tetrahydropyran</td>
<td>C$_3$H$_6$O</td>
<td>−43.2</td>
<td>−6.711</td>
<td>1.537</td>
<td>1.301</td>
<td>−0.324</td>
<td>1.804</td>
</tr>
<tr>
<td>DOL</td>
<td>1,3-Dioxolane</td>
<td>C$_2$H$_4$O$_2$</td>
<td>−64.4</td>
<td>−6.955</td>
<td>1.493</td>
<td>1.324</td>
<td>−0.315</td>
<td>1.818</td>
</tr>
<tr>
<td>DMM</td>
<td>Dimethoxy methane</td>
<td>C$_3$H$_6$O$_2$</td>
<td>−52.0</td>
<td>−6.846</td>
<td>1.459</td>
<td>2.165</td>
<td>−0.298</td>
<td>1.905</td>
</tr>
<tr>
<td>MA</td>
<td>Methyl acetate</td>
<td>C$_2$H$_5$O$_2$</td>
<td>−53.3</td>
<td>−7.371</td>
<td>0.339</td>
<td>1.733</td>
<td>−0.265</td>
<td>1.755</td>
</tr>
<tr>
<td>EP</td>
<td>Ethyl propionate</td>
<td>C$_3$H$_6$O$_2$</td>
<td>−58.6</td>
<td>−7.31</td>
<td>0.414</td>
<td>1.763</td>
<td>−0.269</td>
<td>1.787</td>
</tr>
<tr>
<td>GBL</td>
<td>g-Butyrolactone</td>
<td>C$_3$H$_6$O</td>
<td>−54.7</td>
<td>−7.269</td>
<td>0.254</td>
<td>4.296</td>
<td>−0.237</td>
<td>1.758</td>
</tr>
<tr>
<td>TMP</td>
<td>Trimethyl phosphate</td>
<td>C$_3$H$_6$O$_2$P</td>
<td>−56.8</td>
<td>−7.765</td>
<td>1.112</td>
<td>3.356</td>
<td>−0.467</td>
<td>1.71</td>
</tr>
<tr>
<td>NMP</td>
<td>N-Methyl-2-pyrrolidone</td>
<td>C$_3$H$_6$ON</td>
<td>−65.1</td>
<td>−6.421</td>
<td>0.842</td>
<td>3.609</td>
<td>−0.299</td>
<td>1.724</td>
</tr>
<tr>
<td>ES</td>
<td>Ethylene sulfite</td>
<td>C$_2$H$_4$O$_2$S</td>
<td>−63.9</td>
<td>−7.725</td>
<td>−0.823</td>
<td>3.123</td>
<td>−0.423</td>
<td>1.758</td>
</tr>
<tr>
<td>SL</td>
<td>Sulfolane</td>
<td>C$_4$H$_8$O$_2$S</td>
<td>−63.7</td>
<td>−7.383</td>
<td>0.826</td>
<td>5.087</td>
<td>−0.459</td>
<td>2.014</td>
</tr>
<tr>
<td>PS</td>
<td>1,3-Propanediol</td>
<td>C$_3$H$_6$O$_2$S</td>
<td>−57.3</td>
<td>−7.917</td>
<td>0.549</td>
<td>5.468</td>
<td>−0.426</td>
<td>2.034</td>
</tr>
<tr>
<td>DSMO</td>
<td>Dimethyl sulfide</td>
<td>C$_3$H$_6$OS</td>
<td>−67.8</td>
<td>−6.01</td>
<td>0.963</td>
<td>3.821</td>
<td>−0.542</td>
<td>1.718</td>
</tr>
<tr>
<td>AN</td>
<td>Acetonitrile</td>
<td>CH$_3$N</td>
<td>−47.0</td>
<td>−8.933</td>
<td>0.898</td>
<td>3.743</td>
<td>−0.181</td>
<td>1.92</td>
</tr>
<tr>
<td>PN</td>
<td>Propionitrile</td>
<td>CH$_3$N</td>
<td>−48.4</td>
<td>−8.802</td>
<td>0.587</td>
<td>3.826</td>
<td>−0.185</td>
<td>1.914</td>
</tr>
<tr>
<td>MEK</td>
<td>Methyl ethyl ketone</td>
<td>C$_3$H$_6$O</td>
<td>−53.0</td>
<td>−6.601</td>
<td>−0.386</td>
<td>2.771</td>
<td>−0.225</td>
<td>1.759</td>
</tr>
</tbody>
</table>

where $w_i$ ($i = 1, \ldots, 10$) is the coefficient of the $i$th explanation variable.

As descriptors $x_n$, we adopted the following sets of features, $x_1$ = boiling point, $x_2$ = density, $x_3$ = dipole moment, $x_4$ = flash point, $x_5$ = HOMO, $x_6$ = LUMO, $x_7$ = melting point, $x_8$ = molecular weight, $x_9$ = Mulliken charge, and $x_{10}$ = distance between the Li-ion and the coordinated oxygen atom for the prediction of the coordination energies. In the case of the melting point prediction, $x_2$ is redefined to the coordination energy and the other descriptors are the same as in the former case.

LASSO is also the supervised machine learning method. The linear equation of the fitting is the same as that of the MLR method, while LASSO involves a penalty term as expressed in the second term of eqn (2).

$$ E = \sum_{i=1}^{103} \left( z^a - \sum_{j=1}^{10} w_i x_i^a \right)^2 + \lambda \sum_{j=1}^{10} |w_j| $$  \hspace{1cm} (2)

In eqn (2), $\lambda$ is the penalty parameter and the order of the penalty term is linear. This method is a sparse estimation technique and can minimize the error function with extracted descriptor sets. If $\lambda$ is sufficiently large, some of the coefficients are driven to zero, leading to a sparse model in which the corresponding coefficients play no role. On the other hand, in the case where $\lambda = 0$, the results are the same as the results of MLR. The penalty term allows complex models to be trained on the data sets of limited size without severe over-fitting.

To determine a suitable value of the penalty parameter, $\lambda$, we use cross validation (CV), which approximately extract the prediction error from the limited data. For the CV, the given data from the database are divided into training data and validating data to evaluate the prediction accuracy. After the iteration of this training and validating process with different dividing positions, the CV error is obtained with less variability. We carried out the 10-fold (10 times iterations) cross validation and choose an optimal based on when the CV error was at its minimum. In this study, the CV error of LASSO is derived from the coefficients in eqn (2), which are affected by the optimal penalty parameter.

We then consider the proposed sparse estimation technique, ES-LiR. Assuming that the coefficients are sparse, namely, the coefficients have a small number of non-zero elements, we estimate which coefficient of the explanatory variable is non-zero. To be more precise, let us consider that the number of explanatory variables is $N$. In ES-LiR, in contrast to LASSO, whether each coefficient is zero or not is determined by exhaustively evaluating all combinations of $N$ explanatory variables, $2^N - 1$. To evaluate each combination, each value of the non-zero coefficient is determined by the least squares method and we calculate the CVE for each combination. Finally, we obtain optimal non-zero elements. This approach requires a longer calculation time compared with MLR and LASSO. In this study, the size of the data is not large and we can easily apply the ES-LiR method for the estimation.

We formulate exhaustive search for the linear regression problem (ES-LiR) by using an indicator variable that represents a combination of non-zero explanatory variables. The indicator is defined as an $N$-dimensional binary vector,

$$ e = (e_1, e_2, \ldots, e_N) \in \{0,1\}^N $$  \hspace{1cm} (3)

Each variable $e_i$ takes 0 or 1: $e_i = 1$ if the $i$-th variable belongs to the combination and $e_i = 0$ if it does not. Using the indicator, $e$,
we can write the linear regression problem by minimizing
\[ E = \sum_{p=1}^{p} \left( z_p - \sum_{i=1}^{N} w_i x_i \right)^2, \]
where \( p \) is the number of samples. This formulation makes the essence of the problem more explicit, and the best \( \epsilon \) for modeling and predicting a target variable, \( z \), is searched by minimizing the CVE in ES-LiR.

It is easy to imagine that the ES method becomes intractable for a large size. To reduce the computational load, it is effective to use sampling methods, such as the Markov chain Monte Carlo (MCMC) method and the replica exchange Monte Carlo (REMC) method. In our previous study,\(^{21} \) to deal with the difficulty, we proposed the approximate exhaustive search (AES) method for linear regression, using the above sampling method.

### 3. Results and discussion

#### 3.1. Coordination energy prediction

The correlation between the calculated coordination energies and estimated ones by MLR, LASSO, and ES-LiR is shown in Fig. 1, and their predicted values are shown in Table 2. In these data, the estimated values have a good correlation with the true values (DFT calculated data). For the samples with the lowest coordination energies of around \(-100 \text{ kcal mol}^{-1} \) (true value), the estimation accuracy is not high. The solvents are 12-crown 4-ether and 18-crown 6-ether as shown in Tables S1 and S2 (ESI\(^+ \)). They coordinate to Li-ions by four or more oxygen atoms of the solvent. Thus, the coordination manner is different from the other solvents, and it can be affected to the low estimation accuracy of the coordination energy.

The CV errors of the MLR, LASSO and ES-LiR methods were calculated to be 10.2, 9.18, and 8.78 kcal mol\(^{-1} \), respectively (Table 3). This suggests that the prediction accuracy of ES-LiR is the best among the three methods. The accuracy is mainly affected by the quality of the descriptor choice and the selection of the data-driven technique. Regarding the choice of descriptors, we can generate the descriptors from first-principles calculation results to improve the prediction accuracy, though too many descriptors may cause over-fitting in some information techniques and decrease the accuracy, especially the MLR case. The ES-LiR method can consider the whole combination patterns of the descriptors, and the over-fitting is easily detected by the result of the less prediction accuracy of the combinations. This indicates that we are not suffered from the selection of the information techniques. Remaining treatment for improving the prediction accuracy is by increasing the amount of descriptors.

Fig. 2 shows the histogram of the CV errors of descriptor combinations calculated by the ES-LiR method. The histogram can extract not only the optimal solution but all the solutions, which enable us to map the solutions of various machine learning and data-driven methods and scientists' hypotheses. Then, we can evaluate these methods and hypotheses.\(^{21} \) As shown in Fig. 2, the CV errors of MLR and LASSO and the best value of ES-LiR are depicted. This suggests that LASSO, which

\begin{table}[h]
\centering
\caption{Estimated and first-principles calculation values of the coordination energy of solvents (kcal mol\(^{-1} \))}
\begin{tabular}{|c|c|c|c|c|}
\hline
Solvents & True value & MLR & LASSO & ES-LiR \\
\hline
PC & -57.4 & -50.7 & -55.5 & -57.1 \\
EC & -55.9 & -51.5 & -55.6 & -57.6 \\
VC & -51.7 & -54.1 & -53.1 & -53.0 \\
FEC & -51.2 & -49.3 & -53.3 & -55.8 \\
DMC & -50.0 & -55.0 & -53.6 & -53.9 \\
DEC & -52.6 & -51.0 & -52.7 & -53.8 \\
EMC & -51.3 & -52.3 & -54.9 & -54.8 \\
Furan & -31.7 & -48.0 & -48.4 & -46.1 \\
THP & -48.7 & -51.3 & -53.4 & -52.5 \\
THP & -47.2 & -50.1 & -52.0 & -51.9 \\
DOL & -43.2 & -47.0 & -53.6 & -53.6 \\
DMM & -64.4 & -49.7 & -50.6 & -49.2 \\
MA & -52.0 & -50.3 & -51.5 & -51.8 \\
EP & -53.5 & -51.5 & -51.6 & -51.5 \\
MCA & -58.6 & -50.8 & -52.1 & -54.6 \\
VA & -51.7 & -52.0 & -51.0 & -49.6 \\
GBL & -56.8 & -52.5 & -54.5 & -55.5 \\
TMP & -65.1 & -59.7 & -62.8 & -64.8 \\
NMP & -63.9 & -58.7 & -57.3 & -57.7 \\
SL & -63.7 & -56.8 & -61.4 & -66.3 \\
PS & -57.3 & -60.3 & -39.5 & -61.1 \\
DMSO & -67.5 & -68.2 & -64.7 & -67.2 \\
AN & -47.0 & -46.5 & -45.6 & -46.6 \\
PN & -48.4 & -45.3 & -46.4 & -47.2 \\
MEK & -53.0 & -51.9 & -49.3 & -49.4 \\
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Cross-validation errors of the coordination energies and the extracted combination of descriptors of MLR, LASSO, and ES-LiR}
\begin{tabular}{|c|c|c|c|}
\hline
Data-driven technique & Combination of descriptors & CV error \text{(kcal mol}^{-1} \text{)} \\
\hline
MLR & \( x_1 - x_{10} \) & 10.2 \\
LASSO & \( x_4, x_6, x_9, x_{10} \) & 9.18 \\
ES-LiR & \( x_1, x_{10} \) & 8.78 \\
\hline
\end{tabular}
\end{table}

\( x_1 \) = boiling point, \( x_2 \) = density, \( x_3 \) = dipole moment, \( x_4 \) = flash point, \( x_5 \) = HOMO, \( x_6 \) = LUMO, \( x_7 \) = melting point, \( x_8 \) = molecular weight, \( x_9 \) = Mulliken charge, and \( x_{10} = R(\text{Li-O}) \).
has been widely used in recent studies, is not a best prediction method and the extracted descriptors are not a best combination (Table 3) from the combinations of the small CVE data.

The ES-LiR method not only minimizes the CVE but also derives the CVE in all combinations, so you can see the whole picture of them. Using the whole pictures, the ES-LiR method can be used to construct the weight diagram, which shows the top 25 best combinations of the descriptors, as shown in Fig. 3. The weight diagram reveals the stability of the important descriptors for the estimation, even if the error is at the same level as the other methods. Each colour represents the fitted coefficient of each descriptor, which shows the importance for the coordination energy prediction. The white-blocks of the map correspond to the descriptors which are not adopted for the prediction. From this data, the Mulliken charge is the significant descriptor for the coordination energy prediction. The distance corresponds to the strength of the interaction between Li and O. On the other hand, the flash point is not a trivial descriptor. It might be a weak relationship between “the oxygen radical reaction for burning” and “the Li cation–solvent interaction”, though the number of the samples should be increased for such a discussion.

In materials informatics, proper combinations of descriptors change depending on the purpose of data analysis. In this paper, our goal is both to accurately predict the coordination energy and to reduce the calculation cost. Using the weight diagram (Fig. 3), we realize our purpose. As shown in Fig. 3, the 11th accurate combination does not include the descriptor of $R$(Li–O). To obtain the distance between Li and oxygen, additional Li–solvent complex calculations are required, though the other descriptors, density, flash point, and Mulliken charge, are obtained by catalogue data and only solvent calculations. The difference in the first and 11th CVE errors is quite small, 0.126 kcal mol$^{-1}$. The value is not a significantly big difference for comparing the coordination energies of various solvents. According to Table 1, the $10^{-1}$ kcal mol$^{-1}$ order is the target accuracy for coordination energies. Then, if we choose the 11th best combination of descriptors (“Flash point” and “Mulliken charge”), we can reduce the calculation cost to a half because the extra calculation for obtaining $R$(Li–O) is omitted. This indicates that we can choose the balance of the “prediction accuracy” and the “calculation cost for obtaining the descriptors” for the combinatorial material search when we employ the ES-LiR method and calculate the histogram and weight diagram.

### 3.2. Melting point

Fig. 4 shows the correlation between the melting point from the catalogue data and the estimated data by MLR, LASSO, and ES-LiR. The CV errors of them were obtained to be 30.06, 29.75, and 28.49 °C, respectively (Table 4). Although the CV error is still large in ES-LiR, the error of ES-LiR is smaller than the LASSO and MLR results. From the extraction of the descriptors by LASSO, density is one of the significant descriptors for the melting point. It matches the chemical intuition because the
The weight diagram of descriptors obtained by ES-LiR is shown in Fig. 6. The most important descriptors are density, dipole moment, and flash point, and the least important descriptors are HOMO and LUMO. The weight values of ES-LiR are 28.49, which is less than that of LASSO and MLR.

3.3. Statistical significance of the proposed methods about the CV error

Let us consider the statistical significance of the difference in the CV errors of MLR, LASSO, and ES-LiR. For the evaluation of the CV errors, we calculated the CV error for each data set in ES-LiR, just like the condition of LASSO. As a result of applying it to the coordination energy prediction, the CV errors of MLR, LASSO, and ES-LiR are respectively 10.20, 9.18, and 6.34. We conducted a paired sample t-test to the data of 10-fold CV errors of “MLR and ES-LiR” and “LASSO and ES-LiR”, and the p value was less than 0.001, which was a significant result.

4. Conclusions

In order to explore new LIB electrolyte materials, we investigated the estimation procedure by data-driven information techniques. We predicted the coordination energies and melting points of solvents by information techniques such as MLR, LASSO, and ES-LiR. ES-LiR reproduced the most accurate estimation of the properties among them. We found that ES-LiR chose the balance of “prediction accuracy” and the “calculation cost to obtain the descriptors” when the combinatorial material search by virtual screening was carried out. This feature is general for all the material exploring studies with virtual screening. This treatment can be a key technique to future material searches.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This research was supported by the JST, PRESTO and NIMS, “Materials research by information” integration initiative. The calculations in this work were carried out on the supercomputer center of NIMS. The work was supported in part by the K computer at the RIKEN AICS through the HPCI System Research Projects (Proposal no. hp160174, hp170198, and hp180134). This work was also supported in part by MEXT KAKENHI (JP15H05701).

References