AI-based secondary battery material discovery research

Sohyun An

thgus4425@snu.ac.kr

Abstract

This study presents a method to explore materials for organic batteries to overcome the limitations of existing Li batteries. In order to use such an organic battery as a next-generation battery, it is essential to search for a suitable material. Analyzing the correlation between the molecular structure of organic materials and their performance as a battery material can be one way to do this. There are IE (Ionization Energy) and EA (Electronic Affinity) as the material properties that determine voltage of the battery system, which is one of the most important performance as a battery material. To understand this, empirical methods and theoretical methods such as DFT are used, but in the former case, only a limited number of chemical species were studied, and in the latter case, the calculation cost is very high. If Deep Learning technique is used to identify IE and EA from the molecular structure of organic matter, it is possible to solve the above limitation because it is simplified to understand the performance of the battery from the molecular structure. That is, it becomes possible to select candidate substances in a wide chemical space only with information on the molecular structure of organic substances. For this, supervised learning with Fully Connected Neural Network is used in this research. In this process, Extended-Connectivity Fingerprints are used to make the molecular structure of organic matter into a computer-understand vector.

1. Introduction

With the development of smartphones and electric vehicles, the demand for batteries with high capacity and high energy density is increasing. However, in the case of transition metal oxides such as LCO ($LiCoO_2$), which is a typical material used as a cathode material for existing Li-ion batteries, the specific capacity limit exists due to the use of heavy transition metal elements, and the harmful effects of transition metals cannot be ignored. none. In addition, since high temperature is required in the synthesis process of these electrode materials, CO2 emission in the synthesis process is also a problem. Graphite used as an anode material has the advantage of maintaining a stable structure, but it is disadvantageous in terms of energy density because it can accommodate 1 Li atom per 6 C atoms.

Organic electrode materials are being researched as next-generation battery electrode materials that can compensate for these shortcomings of existing batteries. The organic electrode material uses an organic material that maintains structural stability despite oxidation and reduction as a battery material. By tuning the structure of the organic material, energy density and theoretical specific capacity can be adjusted, and since it can be synthesized at room temperature, CO_2 emission can be reduced.

In order to use such an organic battery as a next-generation battery, it is essential to search for a suitable material. Therefore, it is necessary to understand the correlation between the molecular

structure of the organic material and the performance as a battery material. There are IE (Ionization Energy) and EA (Electronic Affinity) as the material properties that determine voltage, which is one of the most important performance as a battery material. To understand this, empirical methods and theoretical methods such as DFT (Density Functional Theory) are used, but in the former case, only a limited number of chemical species were studied, and in the latter case, the calculation cost is very high.

If the DL (Deep Learning) technique is used to identify IE and EA from the molecular structure of organic matter, it is possible to solve the above limitation because it is simplified to understand the performance of the battery from the molecular structure. That is, since it becomes possible to select a candidate group material only with information on the molecular structure of the organic material, increased efficiency can be expected compared to the conventional method.

2. Background

2.1. Machine Learning

Meaning of Machine Learning[1]

ML is a methodology that gives the computer the ability to learn without explicit programming. It started with the idea that general algorithms exist to find patterns for specific problems in various data sets. Using this, in identifying IE and EA from the molecular structure of organic materials, if the ML method is used instead of the theoretical method such as DFT, which has a large calculation cost, it is possible to simplify the understanding of the performance of the battery from the molecular structure. Such ML is largely classified into supervised learning, unsupervised learning, and reinforcement learning. In this study, the supervised learning technique is used.

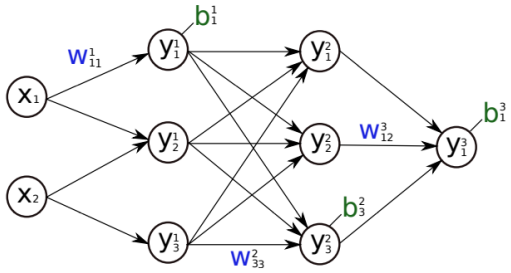

Figure 1: A fully-connected feed-forward neural network

Various characteristics of neural networks

Figure 1 shows an example of a fully-connected feed-forward neural network, which is a representative example of an artificial neural network, and consists of two hidden layers and one output layer. Node(y) is indicated by a circle, and the arrow indicates the direction of information and also the connection of neurons. Each connection has a weight value, and each node is represented by output y and bias b. Hidden neurons exist for the purpose of representing the functional form of NN (Neural Network).

Assuming that all nodes in the same layer have the same activation function, the output value for node i in layer l is as follows.

$$y_i^l = f_l\left(\sum_{i=1}^{N_{i-1}} w_{ij}^l u_j^{l-1} + b_i^l\right)$$

If the output value is calculated for all nodes of a layer, the calculation for the next layer can be performed in the same way. This is expressed as follows.

[1] John-Anders Stende, 『Constructing high-dimensional neural network potentials for molecular dynamics』, Faculty of Mathematics and Natural Sciences University of Oslo, September 2017, pp.19-43.

$$y_i^{l+1} = f_{l+1}\left[\sum_{j=1}^{N_l} w_{ij}^{l+1} f_l\left(\sum_{k=1}^{N_{l-1}} w_{jk}^{l} f_{l-1}\left(\cdots f_1\left(\sum_{n=1}^{N_0} w_{mn}^{1} x_n + b_m^1\right)\cdots\right) + b_j^l\right) + b_i^{l+1}\right]$$

Next, we will look at the overall model training. NN learns through the process of repeatedly performing data feeding to the network. NN finds patterns in data and improves accuracy by adjusting parameters such as weights and biases through a learning algorithm. Our purpose is to construct a NN representing a function that maps IE or EA to output through input data (ECFPs) representing the molecular structure. That is, it is called "training" to repeatedly modify randomly initialized weights and bias values, and to adjust the values until the desired output value is obtained to minimize the error. And the error, which is a measure of the result for training, is defined by what is called a loss function or cost function. A commonly used loss function in regression problems is the Mean Square Error, as follows.

$$\Gamma = \frac{1}{2N}\sum_{i=1}^{N}\left(Y_i - y_i\right)^2$$

(N : *the number of output nodes*, Y_i : *desired output value*, y_i : *predicted output value*)

In addition, Mean Absolute Error is the average of all absolute errors as follows.

$$\Gamma = \frac{1}{N}\sum_{i=1}^{N}\left|Y_i - y_i\right|$$

Another characteristic of NN is the optimization algorithm. Various algorithms are used to obtain the NN parameter set to minimize the loss function. Among the various update rules, in particular, gradient descent-based methods are widely used in NN research. There are Stochastic Gradient Descent(SGD), Momentum, Adagrad, RMSProp, and Adam, etc. It is important to find an appropriate learning rate for both of them. If the learning rate is too small, the convergence will be very slow, and if the learning rate is too large, the loss function may fluctuate near the minimum or even cause divergence.

2.2. Extended-Connectivity FingerPrints[2]

ECFPs can be said to be topological fingerprints that express molecular characteristics. In particular, fingerprints developed with an emphasis on molecular structure-activity modeling. ECFPs can be quickly calculated for a particular molecule and can exhibit a myriad of different molecular properties.

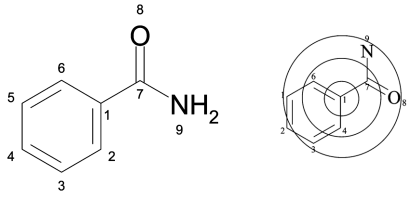

Figure 2: Benzoic acid amide atom numbering (of non-hydrogen atoms).

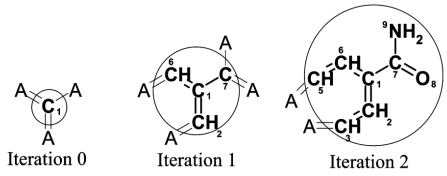

Figure 3: Illustration of the effect of iterative updating on the information represented by an atom identifier

The process of conversion from molecular structure to ECFPs is as follows.

1. Initial assignment stage

Each atom in a molecule is assigned an integer identifier. For example, each atom may be assigned an atomic number. These initial atom identifiers become the initial fingerprint set.

2. Iterative updating stage

[2] David Rogers and Mathew Hahn, 『Extended-Connectivity Fingerprints』, J. Chem. Inf. Model., 50, 742-754, February 4, 2010, pp.742-754.

Each atom collects its own identifier and the identifiers of its immediately neighboring atoms, into an array. In this case, the neighbors are ordered using their identifiers, and the order of the attaching bonds, to avoid order-dependence. When all atoms in the molecule have generated their new identifiers, they replace their old identifiers with their new identifiers. And these new identifiers are added to the fingerprint set. In this step, this iteration is repeated a predetermined number of times.

3. Duplicate identifier removal stage

When all iterations of the set number of iterations are completed, repeated identifiers are removed and the remaining integer identifier set becomes ECFPs.

All these processes are well summarized in Figure 2 and Figure 3. Figure 2 shows an example of atom numbering of benzoic acid amide, and the results for the iteration effect on atom 1 are shown in Figure 3. As can be seen in Figure 3, initial atom identifiers in iteration 0 indicate only each atom and attached bond. When iteration 1 is reached, information on neighboring atoms immediately next to atom 1 is also included. In Iteration 2, it is possible to include information about the amide group as well as a significant portion of the aromatic ring of this molecule.

2.3. Ionization Energy and Electron Affinity of Organic Molecules[3]

IE (Ionization Energy) and EA (Electron Affinity) of organic molecules can be said to be inherent fundamentals of physico-chemical properties. Therefore, it is an important characteristic in determining the energy level in the study of organic electronics materials.

IE (EA) means the energy difference between the neutral N-electron system and the positively-(negatively-)charged N−1 (N+1)-electron system.

These IE and EA were measured through direct photoemission and inverse photoemission techniques, respectively. That is, the energy of holes or electrons injected into the sample was measured. Further from this method, computational quantum mechanical modeling called DFT, which was introduced by Enrico Fermi and developed by Walter Kohn, is one of the theories for calculating the shape of electrons in molecules and their energy by quantum mechanics. DFT can predict whether a molecule may or may not exist in the world, the shape and properties of a particular molecule, IE, EA, etc. Among scientific calculations using computers, it is one of the most widely used method in the fields of quantum mechanics calculations, but has a disadvantage in that the calculation cost is very large.

2.4. Redox Flow Battery Components and Screening Considerations

RFB (Redox Flow Battery) is an electrochemical storage device that stores electrical energy as chemical energy of the electrolyte in a system that charges and discharges the active material in the electrolyte by oxidation-reduction unlike the existing secondary batteries. The actual electrochemical reaction takes place in the stack and works by continuously circulating the electrolyte inside the stack using a fluid pump. Redox pairs used as active materials include V/V, Zn/Br, Fe/Cr, and Zn/air, among which V/V and Zn/Br redox pairs are the most widely used.

An RFB using V for both positive and negative poles is called VRFB, and with this, the basic structure and charging/discharging principle of RFB is as follows.[4]

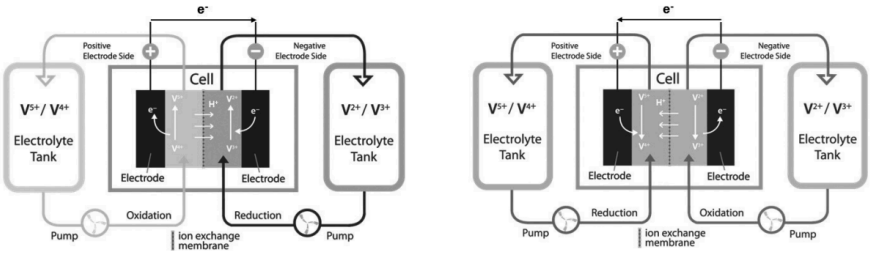

Figure 4. VRFB charging (left) and discharging (right) process

[3] Susumu Yanagisawa, 『Determination of the ionization energy and the electron affinity of organic molecular crystals from first-principles: dependence on the molecular orientation at the surface』, Department of Physics and Earth Sciences, Faculty of Science, University of the Ryukyus, 1 Senbaru, Nishihara, Okinawa 903-0213, Japan, November 27, 2019, pp.1-9.

[4] Shin Han, Yujong Kim, Jihyang Huh, 『Development of Vanadium Redox Flow Battery and Demonstration in Korea』, Journal of the Electric World / Monthly Magazine, Special Issues 4, June, 2014, p. 50.

As shown in Figure 4, during battery charging, $V^{4+}$are oxidized to V^{5+} at the positive electrode side in the stack, and $V^{3+}$are reduced to V^{2+} at the negative electrode side in the stack. A reduction reaction occurs at the anode and an oxidation reaction occurs at the cathode. This RFB is composed of a stack, an electrolyte, and a pump to circulate the electrolyte. Additionally, there is a current collector for transferring electrons from the anode and cathode to the outside.

When selecting the material for the cathode and anode of RFB, the following are the considerations.

1. Reactivity with other components
2. Solubility in electrolyte
3. Voltage

In particular, in the case of 3. Voltage, in order for a battery to have high energy storage capacity, the potential of the anode material should be low and the potential of the cathode material should be high. However, considering this alone, the voltage cannot be increased unconditionally. In addition to this, solubility in electrolytes must be considered, and it must not react with other components of the battery. Considering these factors comprehensively, a potential window should be set and cathode/anode materials should be screened to have a high voltage within it.

Currently, V/V and Zn/Br, which are the most widely used redox pairs in RFB, are toxic elements, so the danger cannot be ignored. If the redox pair is replaced with an organic molecule, the toxicity problem can be solved and the voltage tunability is increased because there are various organic substances, and there is an advantage that the price is lowered. Therefore, in this study, we screen based on organic materials as candidates for cathode/anode materials.

3. Research Method

The study was implemented using the tensorflow.keras library[5].

3.1. Ionization Energy and Electron Affinity data set collection

First, IE and EA data sets are collected from DFT database. After that, organic molecules are converted into SMILES (Simplified Molecular Input Line Entry System, molecular structure string expression). Create a table that summarizes SMILES, IE, and EA for each organic molecule using the Pandas library. This table becomes a data set to be used for ANN training later. Since input data in NN should be ECFPs of organic molecules, it should be converted from SMILES to ECFPs. For this, RDKIT[6] package was used.

3.2. Build ANN Model

Build an appropriate ANN model for each of IE and EA. For this, after repeating training and validation using a part of the above data set for various ANN model candidates, the ANN model with the smallest MAE (Mean Absolute Error) value is selected.

Variables for various ANN models architecture include the number of radius, the number of bits of ECFP, the number of layers, and the number of neurons in each layer. In order to prevent overfitting that the ANN model is specialized for the training data set and the performance is not increased rather than decreased, L2 regularization and Dropout are additionally performed.

At this time, l2_value = 0.003, dropout_rate = 0.1, learning_rate = 0.003, batch_size = 30000, epochs = 5000, the activation function is the ReLU function and the number of k-fold splits is set to 2.

---

[5] Martin Abadi et al., 『TensorFlow: Large-scale machine learning on heterogeneous systems』, Software available from tensorflow.org, 2015.

[6] Open-Source cheminformatics : http://www.rdkit.org

Table 1: ANN model candidates

index	radius	n_bits of ECFPs	layer 1	layer 2	layer 3
0	1	1024	1024	512	256
1	1	1024	512	256	128
2	1	1024	1024	512	-
3	1	1024	1024	256	-
4	2	1024	1024	512	256
5	2	1024	512	256	128
6	2	1024	1024	512	-
7	2	1024	1024	256	-
8	3	1024	1024	512	256
9	3	1024	512	256	128
10	3	1024	1024	512	-
11	3	1024	1024	256	-
12	1	2048	1024	512	256
13	1	2048	512	256	128
14	1	2048	1024	512	-
15	1	2048	1024	256	-
16	2	2048	1024	512	256
17	2	2048	512	256	128
18	2	2048	1024	512	-
19	2	2048	1024	256	-
20	3	2048	1024	512	256
21	3	2048	512	256	128
22	3	2048	1024	512	-
23	3	2048	1024	256	-
24	1	1024	512	256	-

3.3. Model Training & Validation

We selected index2 model for IE and index24 model for EA, and conducts training and validation with the entire data set. At this time, the number of k-fold splits was set to 8, and the remaining conditions were the same as in step 3.2, but dropout_rate = 0.1. The related python code is attached in the appendix. The best model with the lowest validation MAE value is selected as the final model by examining the k-fold cross validation result.

3.4. Screening through Inference

Finally, with the selected model and the entire inference dataset, inference the IE and EA values, respectively. After that, plot the learning curve and obtain R^2, MAE and standard deviation values. In consideration of reliability, the anode and cathode materials are screened with a predetermined potential window.

4. Research Results and Discussion

4.1. Inference results

The entire IE and EA datasets were collected on 2020/08/04 from molecular explorer at materialsproject.org. After k-fold cross validation, the final model was selected based on the model that learned best. The results of inference with the final model are shown in the Table 2.

The values indicating whether the learned model correctly infers the label value were 0.723 for the IE prediction model and 0.445 for the EA prediction model (Table 2). Looking at the inference results of IE and EA of the learned model, it can be seen that the IE case predicted better than the EA case.

Table 2: Inference results

	IE	EA
R^2	0.723	0.445
MAE [eV]	0.51	0.528
Standard deviation	0.452	0.478

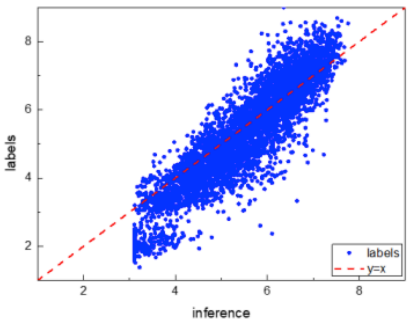

Figure 5: IE Inference result

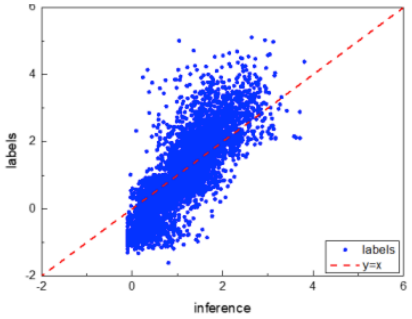

Figure 6: EA Inference result

On the one hand, a not small value of 0.5eV was obtained for MAE. There are two possible reasons for this. First, the input data used for learning in this study is ECFPs of non-oxidized/ reduced molecules, and it may have been difficult to learn IE/EA obtained from oxidation/ reduction processes only with information on non-oxidized/reduced molecules. However, considering the Koopmans' Theorem that molecular IE and EA are closely related to HOMO and LUMO, respectively, this may not be the main cause. For other reasons, the structure of the learning model and the suitability of hyperparameters in the learning process can be considered. It was confirmed that the model trained in Figure 5 and Figure 6 could not make inferences with IE/ EA below a certain value. If hyperparameters such as l2 regularizer and dropout rate are further optimized, it is expected that these problems can be solved.

4.2. Screening

4.2.1. Potential window

There are several factors that determine the potential window when screening the cathode and anode materials of a battery. Among them, the electrode component and the electrolyte component play a decisive role.

Electrolytes commonly used as solvents for battery electrolytes include dimethyl carbonate (DMC), diethyl carbonate (DEC), ethylene carbonate (EC), and propylene carbonate (PC). Among them, screening was conducted based on ACN (Acetonitrile), a solvent often used in RFB system.

Table 3: Redox potential of ACN

E^0 of ACN	Reduction [V]	Oxidization [V]
VS. SHE	-2.6	3.5
VS. Li/Li+	0.44	6.54
Absolute	1.68	7.78

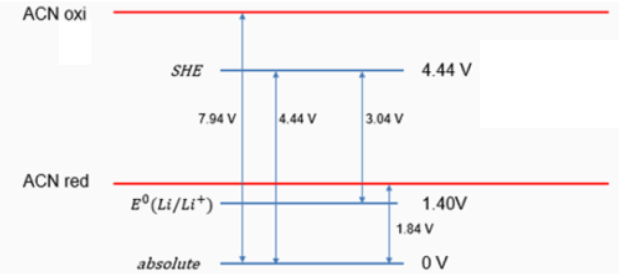

Figure 7: Scheme of Redox potential of ACN

As shown in Table 3 and Figure 7, The redox potential of ACN is reduction 0.44V, oxidation is 6.54V ($VS.\ Li/Li^+$). Also, in a typical battery, the metal used as the current collector of the anode is Aluminum, which is oxidized at 4.7V ($VS.\ Li/Li^+$).

Therefore, in the case of the anode, the maximum oxidation potential is 4.7V ($VS.\ Li/Li^+$). That is, the potential window becomes 0.44V ($VS.\ Li/Li^+$) for the negative electrode and 4.7V ($VS.\ Li/Li^+$) for the positive electrode.

4.2.2. Screening results

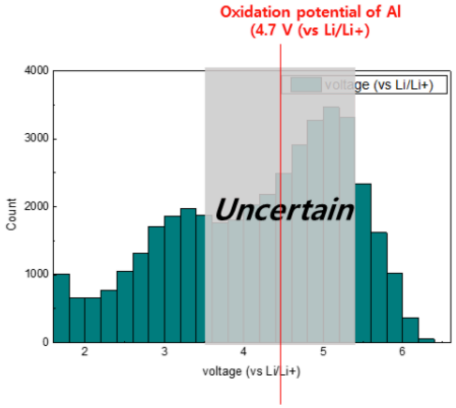

Figure 8: IE Screening result

The (IE value) – 1.40V of a substance is its oxidation potential ($VS.\ Li/Li^+$)[7] [8]. Referring to the IE inference result Figure 8, assuming that the error follows a normal distribution, an uncertain area of 1.41 V is created when estimating with 95% confidence. Most substances can be used with ACN. However, for high energy efficiency, it would be advantageous to use a material in the highest voltage range. From this point of view, if you screen only substances that oxidize in the range between 3.0V and 3.29V, screening results of 2758 substances are obtained.

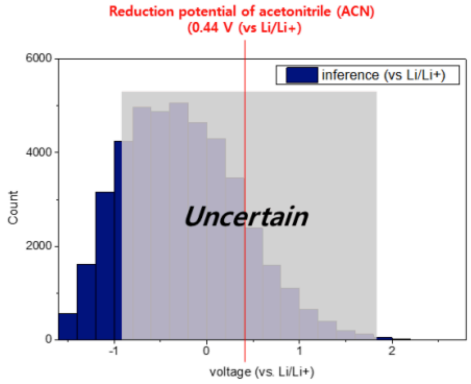

Figure 9: EA Screening result

The (EA value) – 1.40V of a material becomes the reduction potential ($VS.\ Li/Li^+$) of that material. Referring to the EA inference result Figure 9, assuming that the error follows a normal distribution, assuming that the error follows a normal distribution, an uncertain area of 1.484 V is created if it is estimated with 95% confidence. There are 33 substances that can be used with ACN outside this domain.

As such, the number of materials screened as the positive electrode candidate group was 2758 in the 3~3.29 V region, and the number of materials screened as the negative electrode candidate group was 33 in the 1.8 V region or higher. Considering that the voltage range of the materials currently being studied in the organic battery system is from the late 2V to the early 3V at the anode and the mid 1V to the early 2V at the cathode, it is expected that energy efficiency similar to or better than that of the material currently being studied is expected. In addition, in this study,

[7] Y. Ding et al., Chem. Soc. Rev. 47, 69, 2018.

[8] N. Dardenne et al., J. Phys. Chem. C 119, 23373, 2015.

screening was performed assuming ACN for the electrolyte solvent. It is expected that screening can be performed according to the redox potential of the solvent even when other solvents are used.

5. Conclusion

In order to use organic batteries as next-generation batteries to overcome the shortcomings of existing Li batteries, it is necessary to search for suitable materials. Among them, the Deep Learning (DL) technique was applied to identify IE and EA as the material properties that determine the voltage of the battery. The ultimate goal was to construct an ANN that identifies IE and EA from organic molecular structures using DL techniques.

To this end, IE/EA values for organic molecules were first collected through the DFT database, and among various candidate ANN models, Training & Validation was repeated, the model architecture was selected, hyperparameter optimization was performed, and the model with the lowest MAE value was selected. Afterwards, inference was performed with the entire dataset for this model. Finally, when the solvent is ACN in RFB, organic molecules that can be used as anode/anode were screened.

As a result, the R^2values were 0.723 for the IE prediction model and 0.445 for the EA prediction model. That is, it could be confirmed that the IE case predicted better than the EA case. On the other hand, in both IE/EA, the MAE was about 0.5eV, which was not small.

The reason why the value was less than 1 and the MAE value was rather large could be analyzed as follows. While the input data are ECFPs of molecules that are not oxidized/reduced, there may be some differences because IE/EA is obtained in the process of oxidizing/reducing molecules. However, considering Koopmans' Theorem, this is probably not the main cause, and the insufficient hyperparameter optimization of ANN may have had an effect. The evidence is that the learned model could not infer with IE/EA below a certain value. If hyperparameter optimization such as L2 regularizer and dropout rate is sufficiently performed, this problem can be solved.

Referring to the inference result, assuming that the error follows a normal distribution, it was estimated with a reliability of 95%, and materials that could be used with ACN were screened. Since the potential window was 0.44 V for the negative electrode and 4.7 V for the positive electrode, based on this, screening as a positive electrode candidate group was 2758 in the 3~3.29 V range, and the material screened as a negative electrode candidate group was 1.8 V or higher in the region. There were 33.

Considering the voltage range of the materials currently being studied in the organic battery system, it was possible to screen molecules that are similar to or have better energy efficiency than the materials currently being studied.

This study has great significance in that it raised the possibility that IE/EA of organic molecules could be identified through DL. However, the uncertain region is quite extensive and the number of candidates for screening is still very large. To solve this problem, it may be possible to increase the accuracy by using ECFPs of oxidized/reduced organic molecules as input data, or to conduct training on more datasets. Moreover, in addition to sufficient hyperparameter optimization, if the model architecture is further diversified and training is conducted for numerous model architectures, a model with better fitting will be found.

References

1. John-Anders Stende, 『Constructing high-dimensional neural network potentials for molecular dynamics』, Faculty of Mathematics and Natural Sciences University of Oslo, September 2017, pp.19-43.
2. David Rogers and Mathew Hahn, 『Extended-Connectivity Fingerprints』, J. Chem. Inf. Model., 50, 742–754, February 4, 2010, pp.742-754.Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc V. Le. Massive exploration of neural machine translation architectures. *CoRR*, abs/1703.03906, 2017.
3. Susumu Yanagisawa, 『Determination of the ionization energy and the electron affinity of organic molecular crystals from first-principles: dependence on the molecular orientation at the surface』, Department of Physics and Earth Sciences, Faculty of Science, University of the Ryukyus, 1 Senbaru, Nishihara, Okinawa 903-0213, Japan, November 27, 2019, pp.1-9.
4. Shin Han, Yujong Kim, Jihyang Huh, 『Development of Vanadium Redox Flow Battery and Demonstration in Korea』, Journal of the Electric World / Monthly Magazine, Special Issues 4, June, 2014, p. 50.
5. Martin Abadi et al., 『TensorFlow: Large-scale machine learning on heterogeneous systems』, Software available from tensorflow.org, 2015.
6. Open-Source cheminformatics : http://www.rdkit.org
7. Y. Ding et al., Chem. Soc. Rev. 47, 69, 2018.
8. N. Dardenne et al., J. Phys. Chem. C 119, 23373, 2015.
9. Sechan Lee, Giyun Kwon, Kyojin Ku, Kyungho Yoon, Sung-Kyun Jung, Hee-Dae Lim, Kisuk Kang, 『Recent progress in organic electrodes for Li and Na rechargeable batteries』, Advanced Materials, Vol 30, 1704682, 2017.

Appendix

https://github.com/cownow4425/SNU_Advanced-Energy-Materials-Lab/blob/0b42e220d8c7c90ba8c454c3848d6ab174f38ec4/entire_process.py