# Molecular Systems Design & Engineering

Accepted Manuscript






ROYAL SOCIETY OF CHEMISTRY

rsc.li/molecular-engineering

View Article Online
DOI: 10.1039/C9ME00043G

# Design, System, Application Statement

Traditionally, researchers need reference values in order to detect oxide states in data. However, this technique can automatically detect oxide states without the use of reference values. This offers the possibility of detecting oxide states for new materials that do not have reference values available for use.

View Article Online
DOI: 10.1039/C9ME00043G

Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

# Automatic oxidation threshold recognition of XAFS data using supervised machine learning†

Itsuki Miyazato,*[a,b] Lauren Takahashi,[a,b] and Keisuke Takahashi[a,b,c]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Oxidation state of materials are characterized by the X-ray absorption near edge structure(XANES) region om X-ray absorption spectroscopy(XAS). However, the challenge in identifying oxides states are strong depending on researchers judgment based on shift change between measured XAS and reference spectra data. Here, automatic oxidation threshold recognition is performed using machine learning and experimental XAS spectra. In particular, workflow from experimental data collection, data preprocessing and prediction using machine learning are proposed. 10 descriptors for recognizing the oxide state in XAS spectra is discovered. More importantly, oxide state of unknown experimental XAS spectra is identified using trained machine. Proposed approach thus allows for the machine learning to automatically recognize the oxidation threshold of a given XAS spectra without the presence of reference data, leading to the fast analysis of XAS spectra.

## Introduction

X-ray absorption spectroscopy (XAS) is a technique commonly used for characterizing materials due to its selectivity of the target element and its ability to be applied towards a material in any state such as amorphous, and liquid phase[1–5] In particular, the X-ray absorption near edge structure(XANES) region within XAS spectra is key towards understanding properties of target materials since it contains the information of bond states and coordination around the target element.[6,7] In general, the XANES region is typically analyzed through comparison between spectra of the target samples and of reference spectra, either of which can be spectra obtained from experiment or derived from theoretical simulations. Identification of whether materials are oxide or not can be evaluated by comparing the measure XANES peak with reference peak of non-oxide of its material. The challenge in evaluating XANES peak is that the degree of shift from reference XANES peak whether materials are oxide or non oxide are based on human judgment. In other words, there is no solid threshold for deciding oxide or non oxide. This becomes quite problematic for truly characterizing XANES peaks.

Within such circumstance, there have been severals attempts to apply materials informatics for analyzing and understanding the XAFS spectra, in general. In particular, the matching algorithm is reported based on XAS spectra by using FEFF calculation spectra database, but the method is still dependent on the quality of the calculated results.[8,9] THere are cases where three dimensional structures for metal nanoparticles can be identified by using machine learning[10,11]. Yet despite the techniques chosen, all analysis of XAS spectra is still dependent on the availability of reference spectra and of the judgment of the researcher. Without the existence of reference spectra, it is virtually impossible for researchers to judge oxidation states, for example.

Here, automatic oxidation threshold recognition from experimental XANES spectra is proposed using machine learning and data preprocesing of XANES spectra. In particular, the descriptors that affect the changes in XAFS peaks of metal nanoparticles are sought after. By determining the descriptors, it then becomes achievable to employ machine learning which can then learn the hidden trends in XANES specta, thereby leading towards the distinction between oxide and nonoxide states without the presence of reference spectra. The proposed approach therefore establishes a way for a machine learning to automatically detect oxidation thresholds within XAFS spectra without the presence of reference data.

## Method

The proposed approach contains several steps where the workflow is organized and shown in Figure 1. The first step is to collect the experimental XAS spectra and then preprocessed into a format that is readable by a machine learning algorithm. Once

[a] Department of Chemistry, Hokkaido University, Sapporo 060-8510, Japan. E-mail: miyazato@sci.hokudai.ac.jp
[b] Center for Materials research by Information Integration (CMI$^2$),National Institute for Materials Science (NIMS), 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan.
[c] Institute for Catalysis, Hokkaido University, N21, W10, Kita-ku, Sapporo 001-0021, Japan.

† Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 00.0000/00000000.

View Article Online
DOI: 10.1039/C9ME00043G



the data is preprocessed, preprocessed data is trained using machine learning where the machine learns the hidden trends in multi-dimensional space. Upon learning the trends, the trained machine is then used in order to predict whether the XAS spectra is of an oxide or a non-oxide. Note that during the prediction process, the target spectra is acquired outside trained data set therefore, prediction of oxide state in unknown XAS spectra is performed.

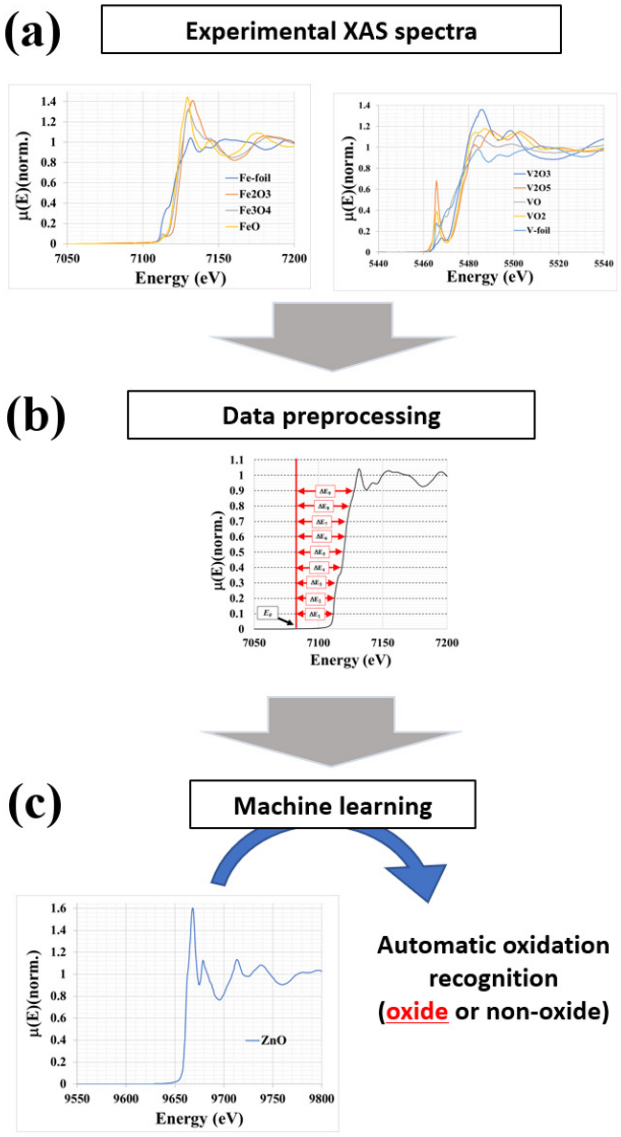


**Fig. 1** Work-flow of Automatic oxidation threshold recognition. (a) Collection of experimental XAS spectra, (b) XAS spectra data preprocessing, (c) Train machine learning with preprocessed data and prediction of oxide state. Note that learned XAS data (a) is not included in XAS spectra data for automatic oxidation recognition (c).

The X-ray absorption spectroscopy(XAS) reference data is collected from the Hokkaido University XAFS database and the Center for Advanced Radiation Sources(CARS) XAFS spectra library.[12,13] 23 K-edge X-ray absorption spectra are collected of the following targets: oxides and pristine foil of silver(Ag), cobalt(Co), iron(Fe), molybdenum(Mo), palladium(Pd), rhodium(Rh), titanium(Ti) and vanadium(V). The collected data is then processed using "Athena", a XAS data processing software, in order to obtain normalized spectra.[14] Please see the supporting information for 23 XAS train data. The energy of the spectra is decomposed into 9 components as shown in Figure 2 where the normalized absorption ($\mu(E)$) classified into $0((E)_0)$, $0.1((E)_{0.1})$, $0.2((E)_{0.2})$, $0.3((E)_{0.3})$, $0.4((E)_{0.4})$, $0.5((E)_{0.5})$, $0.6((E)_{0.6})$, $0.7((E)_{0.7})$, $0.8((E)_{0.8})$ and $0.9((E)_{0.9})$ within the calibrated spectra. The lower energy is taken where the pre-edge appears in the spectra(See Figure 2). The energy shift difference($\Delta\mu(E)_x$; $x$ in the normalized absorption is defined as the following equation (1) in order to standardize the energy shift in all spectra:

$$\Delta\mu(E)_x = E_x - E_0 \quad (1)$$

where $x$ is defined as 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9. Figure 2 illustrates the relationships of the energy shift differences of the spectra, where measurements are taken between starting point $E_0$ and the curve of the particular spectra. These measurements make the data, therefore, readable by a machine learning algorithm.

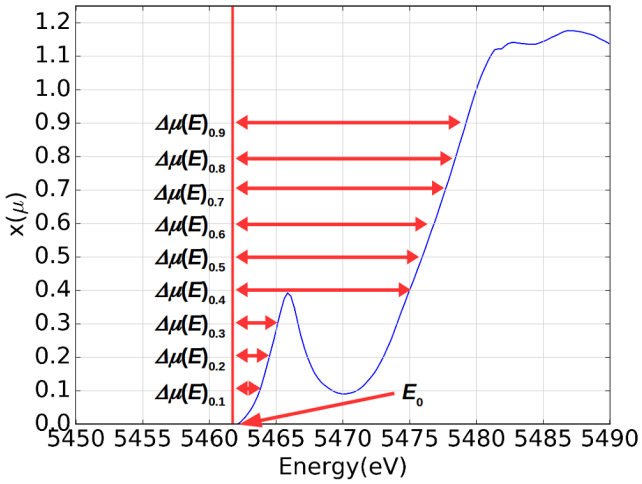


**Fig. 2** The scheme of definition of descriptors $\Delta\mu(E)_x$($x$= 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9) from XAS spectra. Blue represents experimental XAFS peaks while red represents peak measurements made manually.

Scikit–learn is used for machine learning process.[15] In particular, eight different supervised machine learning classification algorithms are applied and evaluated during the search for an appropriate machine learning algorithm and descriptors for oxide states in XAS spectra: Logistic Regression, Support vector machine, Random forest, Extra–trees, Decision trees, Gausian naive bayes, Multinomial naive bayes, Bernoulli naive bayes and k-nearest neighbors.[16–22] Cross validation is implemented for evaluating the accuracy of trained machine where train data set is randomly split into 90% train data and 10% test. Average score of 10 random 90% train data and 10% test is taken and evaluated.

View Article Online
DOI: 10.1039/C9ME00043G



## Results and discussions

The search for descriptors determining the oxides states within XANES analysis is performed using machine learning techniques. Machine learning reveals the following 10 descriptors that determine the oxidation threshold within XAFS spectra: $\Delta\mu(E)_x$($x$= 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9) and atom number of target element. As can be seen in Table 1, these descriptors have high cross validation scores which vary according to the machine learning algorithm, where the support vector classifier is found to return the highest average score of 83%. These descriptors are seen to affect the XANES peaks that are often used when judging if the target material is an oxide or non-oxide. As can be seen in the analyzed XANES spectra listed within Supporting Information, the differences in peaks can be attributed to the difference in energy shift between oxide and nonoxide targets. In general, the XANES peak in metal is shifted more dramatically towards a higher $\Delta\mu(E)$ region as can be seen in the analyzed XANES spectra (see Supporting Information). This tendency has also been reported in previous studies for cases such as Mn, Fe, and Cu.[23–25] Thus, the listed descriptors has good agreement with how the energy shift has been traditionally understood; therefore, machine learning can reproduce researcher judgment.

**Table 1** The results of cross validation with 8 different classification algorithms against oxide states in 2 classes(oxide or non-oxide).

| Algorithm[1] | Average score (%) | Median (%) | Standard deviation |
|---|---|---|---|
| LR[2] | 73 | 66 | 0.24 |
| ETC[3 18] | 70 | 66 | 0.27 |
| DTC[2 19] | 73 | 66 | 24 |
| RFC[3 17] | 73 | 66 | 0.24 |
| SVC[4 16] | 83 | 1 | 0.22 |
| GNB[2 20] | 50 | 33 | 0.3 |
| MNB[2 21] | 33 | 33 | 0.33 |
| BNB[2 21] | 73 | 66 | 0.2 |
| KNN[2 22] | 80 | 83 | 0.2 |

[1] LR : Logistic Regression, ETC : Extra Trees Classifier, DTC : Decision Tree Classifier, RFC : Random Forest Classifier, SVC : Support Vector Classifier, GNB : Gaussian Naive Bayes, MNB : Multinomial Naive Bayes, BNB : Bernoulli Naive Bayes, and KNN : k-nearest neighbors.
[2] Default setting is implemented.
[3] Random states of 1000 and estimators of 3 are implemented.
[4] RBF kernel is implemented.

The effect of descriptors is investigated using the RadViz visualizer[26,27]. RadViz visualization allows for multi-dimensional data to be plotted within a two dimensional space where the data is plotted based on its influence across all dimensions. Descriptors $\Delta\mu(E)_{0.3}$, $\Delta\mu(E)_{0.4}$, $\Delta\mu(E)_{0.5}$, and $\Delta\mu(E)_{0.9}$ within in radviz are chosen by VizRank method where k-nearest neighbor classification is implemented to find the descriptor combinations of the highest score[28]. As seen in Figure 3, the energy shift differences of $\Delta\mu(E)_{0.4}$,$\Delta\mu(E)_{0.5}$ and $\Delta\mu(E)_{0.9}$ appear to have a strong affinity towards the oxide states. Figure 3 demonstrates the spectra against multiple factors (in this case, $\Delta\mu(E)_{0.3}$, $\Delta\mu(E)_{0.4}$, $\Delta\mu(E)_{0.5}$ and $\Delta\mu(E)_{0.9}$) where those 4 descriptors classfied the oxides 0 and 1 which denote nonoxide and oxide states, respectively. In particular, spectra that are considered to be oxides appear to favor $\Delta\mu(E)_{0.3}$ and $\Delta\mu(E)_{0.4}$ while non-oxides appear to favor $\Delta\mu(E)_{0.9}$. From this, it can be concluded that $\Delta\mu(E)_{0.3}$, $\Delta\mu(E)_{0.4}$, $\Delta\mu(E)_{0.5}$ and $\Delta\mu(E)_{0.9}$ are strong indicators for determining the oxidation state of the target spectra.

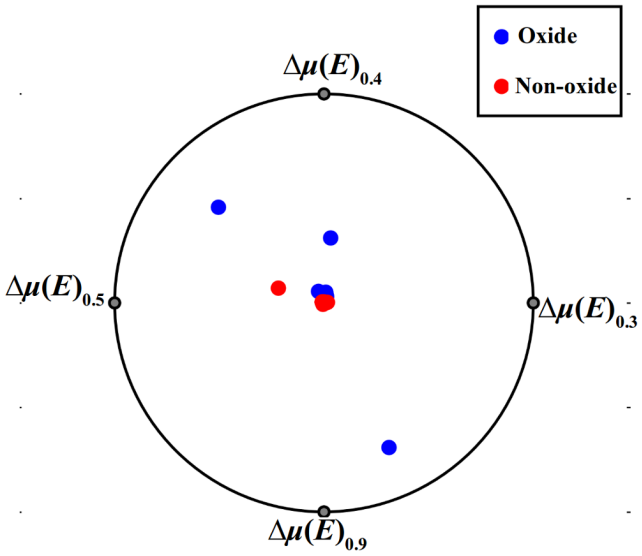


**Fig. 3** The data analysis among descriptors by using RadViz Visualizer.[26,27] Energy shift differences($\Delta\mu(E)_{0.3}$, $\Delta\mu(E)_{0.4}$, $\Delta\mu(E)_{0.5}$ and $\Delta\mu(E)_{0.9}$) has turned out to have strong relationship. Note that nonoxide and oxide states are denoted by 0 and 1, respectively.

Once the descriptors are determined, the inverse problem of prediction of oxide state in unknown XANES spectra is performed. Success with the inverse problem can lead to trained machines becoming proficient in determining the oxidation threshhold without the presence of reference spectra. Here, 10 sets of XAS spectra is collected from the SPring-8 Experimental Data Repository System Portal[29]. Cu, Nb, Zn, and Zr are chosen as the target atoms in order to exclude the atoms that were present during the original machine training phase. Then each of 10 XAS spectra is same data preprocess as taken in Figure 2. The details of preprocessing 10 XAS data is collected in Supporting Information. Once data is proprocessed, oxide state of those 10 XAS spectra is judged by the trained machine. The results are evaluated by the accuracy score as defined by the following equation (Equation 2):

$$P_{accuracy} = \frac{N_{correct}}{N_{all}} \quad (2)$$

where $P_{accuracy}$, $N_{correct}$ and $N_{all}$ are the accuracy score(%), number of correct answers, and number of samples(in that case: 10), respectively.

Table 2 lists the predictions of various trained machines against the new dataset of ten XAS spectra with true values, predicted values, and corresponding accuracy scores are listed. Logistic regression (LR) is seen to have an accuracy score of 80%, which is considered accurate. Additionally, it is the only algorithm that can successfully distinguish an oxide state from an unoxidated state. In this case, the two instances where predicted values were not

View Article Online
DOI: 10.1039/C9ME00043G



**Table 2** The results of reverse problem by using trained machine against 10 XAS spectra.

| Target atom | Sample[1] | True value[1,2] | Predicted value | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | LR[2,3] | ETC[2,3] | DTC[2,3] | RFC[2,3] | SVC[2,3] | GNB[2,3] | MNB[2,3] | BNB[2,3] | KNN[2,3] |
| Cu | Cu-foil | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | CuO | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | $Cu_2O$ | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| Nb | Nb-foil | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| | NbO | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 |
| | $NbO_2$ | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 |
| Zn | Zn-foil | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| | ZnO | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| Zr | Zr-foil | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| | $ZrO_2$ | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 |
| Accuracy score (%) | | | 80 | 50 | 60[4] | 60[4] | 60[4] | 20 | 60[4] | 60[4] | 50 |

1 Sample spectra is collected from SPring-8 Experimental Data Repository System Portal[29].
2 0 and 1 are indicated as "non-oxide" and "oxide" ,respectively.
3 LR : Logistic Regression, ETC : Extra Trees Classifier, DTC : Decision Tree Classifier, RFC : Random Forest Classifier, SVC : Support Vector Classifier, GNB : Gaussian Naive Bayes, MNB : Multinomial Naive Bayes, BNB : Bernoulli Naive Bayes and KNN : k-nearest neighbors.
4 Note that all predicted values are "1 (oxide)" so that it cannot adequately capture the data structure in its learning process, such as over-fitting or inappropriate algorithm implementation.

accurate were cases where the oxide was weakly oxidated. This demonstrates that while it may be difficult to determine the exact threshold for cases with weak oxidation, the machine learning can reliably predict the oxidation state of the XAS data with no reference data. More importantly, this proves that it is, indeed, possible for a machine learning to predict the oxide state of the spectra of untrained XAS spectra once it learns from experimentally-reported spectra without the presence of reference spectra. This thereby expands researchers' ability to determine oxidative states of spectra without supplementary date where it was previously impossible due to lack of reference spectra for researchers to use for comparison.

It is also immediately recognizable that five of the eight algorithms are unable to determine the difference in oxidation state despite. Decision tree classifier(DFC),random forest classifier(RFC),support vector classifier (SVC) ,and Bernoulli Naive Bayes(BNB) returned a value of 1 for all test cases; despite its seemingly high cross validation score, these algorithms were unable to distinguish the oxidation states. k-nearest neighbors returned a value of 1 for Cu cases, while other returned a value of 0. Additionally, Multinomial Naive Bayes(MNB) returns a validation score of 20%, the lowest of all eight algorithms, while extra tree classifier(ETC) has an accuracy score of 50%, which is remarkably lower than the first round of reported cross validation scores (as reported in Table 1). These issues could possibly be due to overfitting data processing with the implemented algorithms due to the size of the dataset although score in cross validation is relatively high as shown in Table 1[30] Additionally, one can consider that the learning algorithm is incomplete in its capture of the data structure and therefore leading to errors in predicting the target variable. These results, thus, can also act as an example where initial cross validation scores do not properly evaluate the continued success of an algorithm when applied towards supervised learning.

## Conclusions

In conclusion, descriptors responsible for oxidation states within oxides are determined by applying machine learning towards data derived from XAS data. 23 sets of oxide and pristine samples of XAS spectra are collected and preprocessed for machine learning applications. Various machine learning algorithms are then explored in order to determine the descriptors responsible for determining the oxide state. The following 10 descriptors are thereby determined: $\Delta\mu(E)_x$($x$= 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9) and atom number of the target element where the highest average cross validation score is 83 % for the support vector classifier. The descriptors can be understood to be the energy shift in XANES spectra that differs according to the oxidation state in certain X-ray absorption edge of the element. The inverse problem is then explored using trained machines against a new set of different XAS spectra in order to evaluate the machine's ability to recognize the oxidation threshhold of untrained spectra where the logistic regression algorithm is found to be most successful with a cross validation score of 80%. These results therefore demonstrate that by employing machine learning, it becomes possible to determine the oxidation state of XAS spectra without the use of reference data or spectra, thereby expediting and expanding the scope of XANES analysis using XAS spectra.

## Conflicts of interest

The authors declare no competing financial interests.

## Acknowledgements

This work is funded by Japan Science and Technology Agency(JST) CREST Grant Number JPMJCR17P2, JSPS KAKENHI Grant–in–Aid for Young Scientists (B) Grant Number JP17K14803, and Materials research by Information Integration ($MI^2I$) project of the Support Program for Starting Up Innovation Hub from Japan Science and Technology Agency (JST). The work

View Article Online
DOI: 10.1039/C9ME00043G



was carried out under the cooperative research program of Institute for Catalysis at Hokkaido University (No2018-001). The authors greatly thank Prof. Dr. Kiyotaka Asakura and Dr. Hajime Imura for good discussion in data preprocessing and descriptors selections.

## Notes and references

1 G. Malta, S. A. Kondrat, S. J. Freakley, C. J. Davies, L. Lu, S. Dawson, A. Thetford, E. K. Gibson, D. J. Morgan, W. Jones *et al.*, *Science*, 2017, **355**, 1399–1403.

2 A. Martini, E. Borfecchia, K. Lomachenko, I. Pankin, C. Negri, G. Berlier, P. Beato, H. Falsig, S. Bordiga and C. Lamberti, *Chemical science*, 2017, **8**, 6836–6851.

3 M. L. Whittaker, W. Sun, K. A. DeRocher, S. Jayaraman, G. Ceder and D. Joester, *Advanced Functional Materials*, 2018, **28**, year.

4 J. Vega-Castillo, G. Buvat, G. Corbel, A. Kassiba, P. Lacorre and A. Caneiro, *Dalton Transactions*, 2017, **46**, 7273–7283.

5 D. Wakabayashi, N. Funamori, T. Kikegawa, K. Watanabe, S. Kohara, H. Nitani, Y. Niwa, Y. Takeichi, H. Abe and M. Kimura, *Phys. Rev. B*, 2017, **96**, 024105.

6 C. S. Spanjers, P. Guillo, T. D. Tilley, M. J. Janik and R. M. Rioux, *The Journal of Physical Chemistry A*, 2016, **121**, 162–167.

7 P. D'Angelo and V. Migliorati, *The Journal of Physical Chemistry B*, 2015, **119**, 4061–4067.

8 C. Zheng, K. Mathew, C. Chen, Y. Chen, H. Tang, A. Dozier, J. J. Kas, F. D. Vila, J. J. Rehr, L. F. Piper *et al.*, *npj Computational Materials*, 2018, **4**, 12.

9 K. Mathew, C. Zheng, D. Winston, C. Chen, A. Dozier, J. J. Rehr, S. P. Ong and K. A. Persson, *Scientific data*, 2018, **5**, 180151.

10 J. Timoshenko, D. Lu, Y. Lin and A. I. Frenkel, *The Journal of Physical Chemistry Letters*, 2017, **8**, 5091–5098.

11 J. Timoshenko, A. Halder, B. Yang, S. Seifert, M. J. Pellin, S. Vajda and A. I. Frenkel, *The Journal of Physical Chemistry C*, 2018, **122**, 21686–21693.

12 *XAFS database, Insutitute for Catalysis, Hokkaido University*, `https://www.cat.hokudai.ac.jp/catdb/index.php?action=xafs_login_form&opnid=2`, (Accessed on 01/15/2018).

13 *XAS Spectra Library (beta)*, `http://cars.uchicago.edu/xaslib/search`, (Accessed on 01/15/2018).

14 B. Ravel and M. Newville, *Journal of Synchrotron Radiation*, 2005, **12**, 537–541.

15 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay.

16 C. Cortes and V. Vapnik, *Machine Learning*, 1995, **20**, 273–297.

17 L. Breiman, *Machine Learning*, 2001, **45**, 5–32.

18 P. Geurts, D. Ernst and L. Wehenkel, *Machine Learning*, 2006, **63**, 3–42.

19 L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, 1984.

20 T. F. Chan, G. H. Golub and R. J. LeVeque, COMPSTAT 1982 5th Symposium held at Toulouse 1982, 1982, pp. 30–41.

21 C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to information retrieval*, Cambridge university press Cambridge, 2008, vol. 1, pp. 234–265.

22 N. S. Altman, *The American Statistician*, 1992, **46**, 175–185.

23 M. Belli, A. Scafati, A. Bianconi, S. Mobilio, L. Palladino, A. Reale and E. Burattini, *Solid State Communications*, 1980, **35**, 355–361.

24 J. Zhao, F. Huggins, Z. Feng, F. Lu, N. Shah and G. Huffman, *Journal of Catalysis*, 1993, **143**, 499–509.

25 A. Gaur, B. Shrivastava and S. Joshi, 2009, **190**, 012084.

26 J. Demšar, T. Curk, A. Erjavec, v. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik and B. Zupan, *Journal of Machine Learning Research*, 2013, **14**, 2349–2353.

27 P. Hoffman, G. Grinstein, K. Marx, I. Grosse and E. Stanley, Visualization'97., Proceedings, 1997, pp. 437–441.

28 G. Leban, B. Zupan, G. Vidmar and I. Bratko, *Data Mining and Knowledge Discovery*, 2006, **13**, 119–136.

29 *SPring-8 Experimental Data Repositry System Portal*, `https://sp8dr.spring8.or.jp/portal/display`, (Accessed on 09/24/2018).

30 D. M. Hawkins, *Journal of Chemical Information and Computer Sciences*, 2004, **44**, 1–12.