Chemistry Letters, 2024, **53**, upae090
https://doi.org/10.1093/chemle/upae090
Advance access publication 30 May 2024

Letter

CSIML: a cost-sensitive and iterative machine-learning method for small and imbalanced materials data sets

Shengzhou Li[1,2], **Ayako Nakata**[1,2,*]

[1]Department of Computer Science, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

[2]Research Center for Materials Nanoarchitectonics, National Institute for Materials Science, 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan

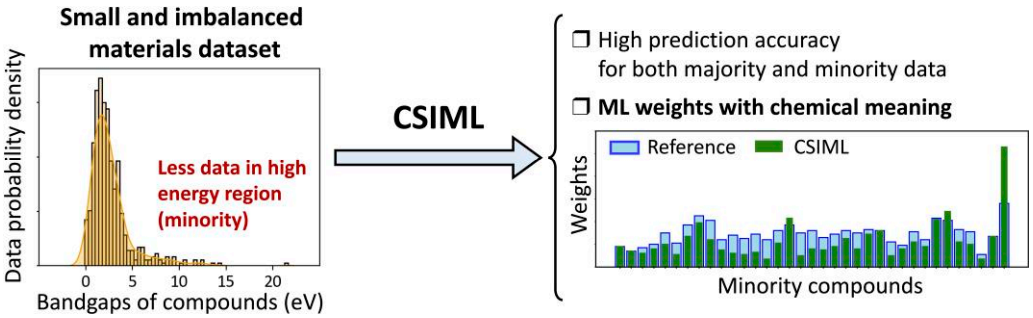
*Corresponding author: Research Center for Materials Nanoarchitectonics, National Institute for Materials Science, 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan. Email: NAKATA.Ayako@nims.go.jp

Abstract

Materials science research benefits from the powerful machine-learning (ML) surrogate models, but it is also limited by the implicit requirement for sufficiently big and balanced data distribution for ML. In this paper, we propose a model to obtain more credible results for small and imbalanced materials data sets as well as chemical knowledge. Taking 2 bandgaps imbalanced data sets as instances, we demonstrate the usability and performance of our model compared with common ML models with normal sampling and resampling methods.

Keywords: data-set bias, machine learning, materials data.

Graphical Abstract



We propose a model to obtain more credible results for small and imbalanced materials data sets as well as chemical knowledge. Taking 2 bandgaps imbalanced data sets as instances, we demonstrate the usability and performance of our model compared with common ML models with normal sampling and resampling methods.

Recently, data-driven machine-learning (ML) methods have been widely used in materials research to analyze and discover novel insights from materials experimental[1] and computational[2] data sets. While there are several large materials data sets such as Materials Project,[3] the Open Quantum Materials Database (OQMD),[4] and DICE,[5] researchers often make their own data sets based on their experiments/calculations or the collection of the data of interest, which are often imbalanced (or biased).[6] For instance, 95% of the compounds in OQMD are conductors with zero bandgap energies.[2c] The size and balance of the data set significantly affect the performance of ML models.[7] Fujinuma et al.[8] reported that big data is not always necessary, but data-set bias plays a particularly important role in ML in materials science. The precision of the standard ML methods would not be good for training data, and even totally bad for unknown test data when the materials data set is imbalanced,[9] while this problem is often invisible when using tens of thousands of data or quite complicated ML models (ex. Neural Networks).

Data resampling such as oversampling and undersampling[10] has been performed to improve ML performance for imbalanced data. In oversampling, the minority data are included multiple times to increase the weights of the minority data to be comparable to those of the majority data. Oppositely, in undersampling, the majority data are thinned out to reduce the weights of the majority data to be comparable to those of the minority data. Lu et al.[11] reported that a data-set bias in a small and imbalanced data set with 539 hybrid organic–inorganic perovskites (HOIPs) and 24

[Received on 10 April 2024; revised on 1 May 2024; accepted on 1 May 2024; corrected and typeset on 30 May 2024]

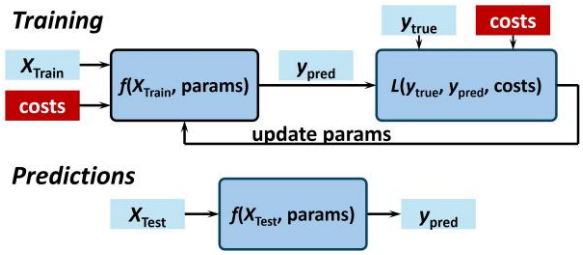


Fig. 1. Workflow of the CSIML method. *X* is the features and *y* is the target property. *f* is the ML model and params is the parameters for the model *f*. *L* is the loss function.

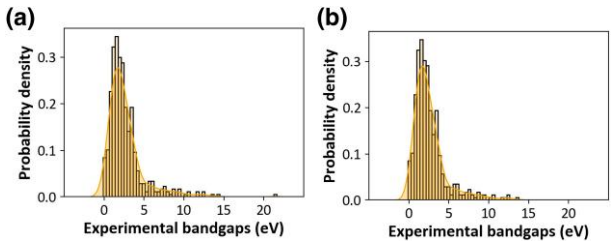


Fig. 2. Bandgap data distributions of a) dataset-S[12] and b) dataset-L.[13]

non-HOIPs caused a bad classification performance. They tried a lot of resampling and generation methods, including oversampling and the synthetic minority oversampling technique (SMOTE) method, to solve this data-set bias. It is more difficult to tackle prediction tasks than classification tasks for imbalanced materials data sets.

In this work, we propose a cost-sensitive and iterative ML method (CSIML) to build an accurate predictive ML model for small and imbalanced materials data sets. The workflow of the CSIML method is shown in Fig. 1 and an algorithm description is provided in Supplementary Algorithm S1. The code is available in GitHub repository (https://github.com/zhonger/CSIML). As in Fig. 1, first, we train an ML model f only based on the majority training set $\{X_M, y_M\}$, where X_M and y_M are the features and bandgaps of the majority training set. This first model naturally will not perform well for the minority set. Then, one of the minority instances i with $\{X_i, y_i\}$ is added into the training set and the ML model is retrained for the added training set by optimizing the costs (i.e., weights) of the instances in the model based on the loss function L. L can be also used to determine the order in which instances are to be added, such as selecting the instance with the smallest prediction errors in the current model to add to the next training set (ascending prediction error order). Thus, the minority instances are added into the training set one by one iteratively, as well as retraining the ML model with the progress. As a result, minority training instances play a more significant role in the ML model than in the conventional ML model. This will also improve predictions for unknown minority instances. Furthermore, because the weights of minority data are optimized one by one, this method enables not only the improvement and balance of the performance of the majority and minority data set, but also extraction of some knowledge from the iterative training process.

Two imbalanced data sets of materials bandgaps are used in the present study: a small data set with 472 bandgaps (dataset-S)[12] and a large data set with 3,895 bandgaps (dataset-L),[13] which is originally from OQMD. The data-set distributions of them are shown in Fig. 2. Both data sets have peaks at approximately 2 eV and most materials in the data set have small bandgaps that are close to 0 eV. From the distributions, it is assumed to be possible to build a highly accurate predictive ML model for the bandgaps in the range of [0, 5] eV, whereas the prediction performance for larger bandgaps (ex. >10 eV) would not be good enough because of the lower data quantity. It is almost impossible to predict for Ne, whose bandgap is very large (21.48 eV). Therefore, 4 materials whose bandgaps are bigger than 15 eV are removed from dataset-S when training ML models, resulting in 468 bandgaps in dataset-S. Based on these, the threshold for splitting the majority and minority sets into 2 bandgap data sets is set to 5 eV (more details can be found in Supplementary Fig. S1). With this threshold, the data whose bandgap is smaller/larger than 5 eV belongs to the majority/minority set and the ratios of the majority and minority data are 381:39 and 3,349:155 for dataset-S and dataset-L, respectively (more details appear in Supplementary Table S1).

Considering the balance of the majority and minority sets, we use the cross-validation (CV) method, as shown in Fig. 3, inspired by the leave-one-cluster-out cross-validation methods.[14] The present CV method is suitable for investigating imbalanced materials data sets as they keep the imbalance of the bandgap data sets according to the data-splitting results of dataset-S and dataset-L with the 10-fold CV shown in Supplementary Table S1. We use 136 features proposed by Zhuo et al.,[13] which are constructed from 34 elemental structure and property parameters according to composition elements with 4 kinds of calculation: maximum, minimum, average deviation, and mean (Supplementary Table S2). A support vector regression (SVR) ML model with python scikit-learn library[15] is used throughout the study. We optimize hyperparameters based on the simplified CV method (CV') shown in Supplementary Fig. S2 and Supplementary Table S3. From the hyperparameter optimization results in Supplementary Fig. S3, the hyperparameters in the SVR model (penalty parameter C, radial basis kernel function parameter γ, and acceptable error tolerance δ) are set to 10, 0.01, and 0.2, respectively.

The performance of the CSIML model is verified by comparing it with those of the oversampling and undersampling methods. Here, we use the imbalanced-learn library[16] for resampling. CV without any resampling methods is also compared as a standard. Several metrics including the root mean square error (RMSE), variance, mean absolute percentage error (MAPE), and R^2 for the test sets are used to evaluate the performance, which in the test set are summarized in Table 1. These metrics are defined in Supplementary Table S4 and the performance for the training and validation sets are provided in Supplementary Table S5.

For CV without resampling, the test RMSE of dataset-L is 0.438 eV, which is comparable to that reported by Zhuo et al. (0.45 eV).[13] The test RMSE of dataset-L is smaller than that of dataset-S (0.982 eV). By comparing the imbalance (ratio) between the majority and minority training sets for 2 data sets in Supplementary Table S1, it is found that dataset-L is more skewed or imbalanced than dataset-S. The minority test RMSE is much worse than the majority test RMSE for dataset-L, despite the good performance of the ML model for most materials, showing small total RMSE.

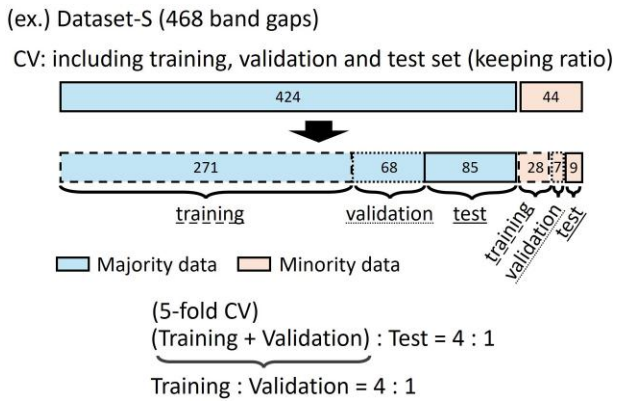


Fig. 3. CV method. Here, 5-fold CV is taken as an instance. The ratio is between the majority and minority training instance numbers. The iteration for CV is defined as the category of the combination of training, validation, and test set, so the iteration for normal k-fold CV is k.[4]

Table 1. The statistical results for 3 methods with 625 iterations for dataset-S and dataset-L in the test set: (a) w/o resampling, (b) oversampling, (c) undersampling, (d) CSIML.

Dataset-S				
Metrics	(a)	(b)	(c)	(d)
Variance	0.491	0.593	0.770	0.645
RMSE	0.982	1.078	1.734	1.261
Maj RMSE	0.680	1.079	1.725	1.271
Min RMSE	3.843	1.072	1.823	1.168
MAPE	0.388	0.452	0.629	0.476
R^2	0.768	0.737	0.580	0.691
Dataset-L				
Metrics	(a)	(b)	(c)	(d)
Variance	0.311	0.325	0.485	0.346
RMSE	0.438	0.578	0.927	0.687
Maj RMSE	0.383	0.580	0.944	0.694
Min RMSE	1.591	0.524	0.570	0.535
MAPE	0.519	0.589	0.779	0.626
R^2	0.809	0.748	0.596	0.701

Maj RMSE and Min RMSE are the RMSE for the majority and minority, respectively. The units of variance and RMSE are eV. (C = 10, random_seed = 10.)

This means that a more imbalanced data set would not lead to poorer performance for the minority test set if the data set was big enough. In other words, the influence of data-set bias is easily ignored if we only focus on the total test RMSE. Hence, the consideration about data-set bias is crucial in an ML model with imbalanced materials data sets.

Among the results of CV without/with resampling, CV with oversampling seems the best. The balanced materials data set after oversampling is obviously better than the CV without resampling. The data-set size after oversampling (twice that of the majority set) is larger than that after undersampling (twice that of the minority set). ML models can benefit from larger data-set size so that oversampling performs better than undersampling naturally. Based on these results, it is suggested that resampling methods are indeed helpful for minority bandgaps prediction, although they may have a negative impact on predicting some majority bandgaps.

For CSIML with an ascending prediction error order, the minority RMSE is obviously reduced from 3.843 eV of CV without resampling to 1.168 eV for dataset-S, although the

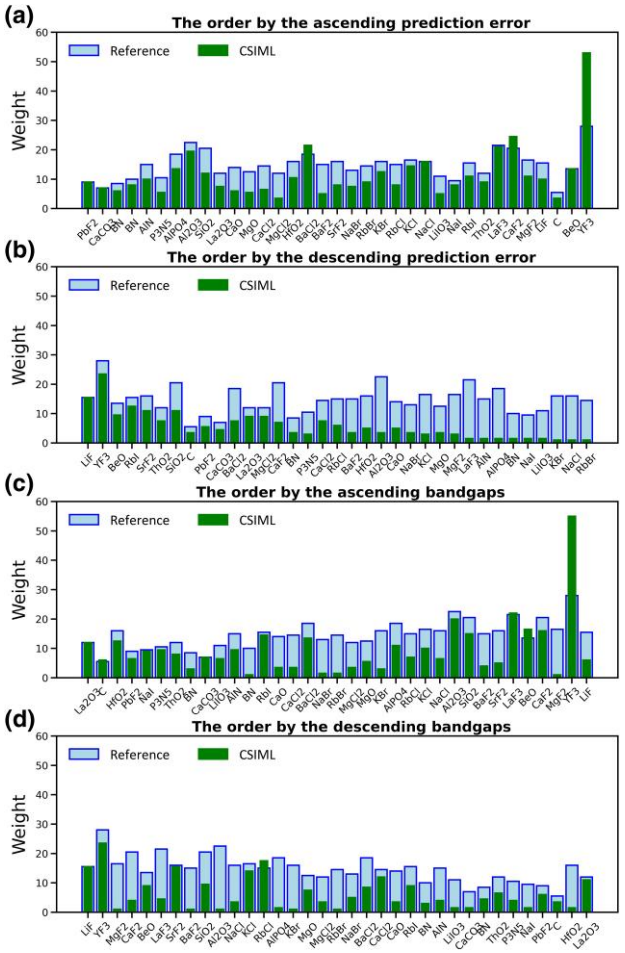


Fig. 4. The weights of the minority training instances in 10-fold CV with different training orders: a) by the ascending prediction error, b) by the descending prediction error, c) by the ascending bandgaps, and d) by the descending bandgaps.

majority RMSE increases. In dataset-L, the minority obtains an even bigger boost (1.056 eV) to the little loss (0.311 eV) in the majority set. CSIML performs well similarly with different data distributions and various thresholds, as shown in Supplementary Fig. S4 and Supplementary Table S6. CSIML and CV with oversampling have comparable performances in both data sets but, as discussed below, CSIML has the potential for discovering some physical/chemical knowledge from the training process.

In the oversampling and undersampling methods, the weight for each instance is determined randomly while taking the balance between the instances. This means that the weights in these methods do not reflect the physical/chemical aspects of the materials in the minority set. On the other hand, the weights in CSIML are optimized one by one for each instance; therefore, there is a possibility that the weights contain some physical/chemical meaning.

Figure 4 shows the weights of the training minority instance in CSIML for one of the iterations in CV. Since we determine the weights iteratively, the order of the trained instances affect the weights. We examined 4 kinds of training orders: ascending (Fig. 4a) and descending (Fig. 4b) orders for the prediction errors, ascending (Fig. 4c) and descending (Fig. 4d) orders for bandgap energies. To demonstrate the effect of the training

order clearly, the training minority materials in the x-labels in Fig. 4 are aligned according to the training orders. In Fig. 4, the weights for which each minority instance is trained together only with the majority data and no other minority data are also provided as the reference values for comparison. The difference from the reference weight reflects the effect of the minority data trained previously.

From Fig. 4, we find that most of the weights in CSIML are smaller than the reference weights, which means that the trained instances improve the prediction for the subsequent instances. As expected, the weights in CSIML change when we use a different training order. In many cases, it is found that the weight of a material is smaller than its reference weight when another material with similar properties is already trained. For example, the weights of a target material tend to be smaller when materials containing the same elements or with the same compositional pattern (e.g., alkali metal with halogen) as the target compound are already trained.

There are some exceptions; for example, the weight of Al_2O_3 in CSIML is much smaller than the reference weight in Fig. 4b,d, while they are less different in Fig. 4a,c, and several compounds with the Al or O element are already trained in Fig. 4a,c. It is found that the weight of Al_2O_3 is small when SiO_2 is trained before Al_2O_3, as shown in Fig. 4b,d. This means that, for Al_2O_3, the training of SiO_2 is more effective than the training of the materials consisting of the same elements (Al and O), although the properties of the constituent elements in the materials are used as the features in the present ML. This suggests that there is some correlation between SiO_2 and Al_2O_3, which is consistent with the fact that SiO_2 and Al_2O_3 form various silica–alumina composite materials such as zeolite[17] and multicomponent glass systems.[18] Furthermore, the principal component analysis method also finds similarity in the features of SiO_2 and Al_2O_3, discussed in the Supplementary Information (Supplementary Fig. S5).

It is also found that YF_3 has the largest weight for all of the 4 training orders. When YF_3 is trained last in the training order, even after the training of LaF_3, which has the same compositional pattern as YF_3, the weight of YF_3 is rather increased. This indicates that YF_3 has significant difference from other materials in the minority set. Thus, we can expect that the weights in CSIML reflect physical/chemical insights of materials.

In conclusion, we have investigated how to realize, pay more attention to, and tackle data-set bias in materials data sets. We have proposed a method, CSIML, to overcome the data-set bias by optimizing the weight values of minority data one by one in the ML model. For 2 kinds of bandgap data sets, by considering data-set bias, the ML models with CSIML and the resampling methods obtain more reliable performance than the conventional ML model using all bandgap data without considering the bias. CSIML has a better balance between majority and minority sets. Moreover, there exists some physical/chemical knowledge of materials in the weights of minority instances in the training process. Although CSIML is designed for small and imbalanced materials data sets, it will be also possible to use it for large and imbalanced materials data sets after some modifications, such as example dividing large and imbalanced materials data sets into smaller pieces first, applying the CSIML method for them separately, and combining them ultimately into one ML model. It will be also possible to use CSIML for so-called online learning, i.e., training the divided data iteratively to make an ML model for a large data set efficiently. In the future, CSIML will be used to explore materials with desired material properties. Knowledge-guided exploration with CSIML will be more reliable and trustworthy, to help in understanding physical/chemical backgrounds in the prediction.

Acknowledgments

Some of the calculations in this study were performed on the Numerical Materials Simulator at NIMS.

Supplementary data

Supplementary material is available at *Chemistry Letters* online.

Funding

This study was supported by JSPS Grant-in-Aid for Transformative Research Areas (A) "Hyper-Ordered Structures Science" (Grant Nos. JP20H05883 and JP20H05878) and PRESTO, Japan Science and Technology Agency (JST) (Grant No. JPMJPR20T4).

Conflict of interest statement. None declared.

References

1a. S. Li, H. Zhang, D. Dai, X. Wei, Y. Guo, *J. Alloys Compd.* **2019**, *782*, 110. https://doi.org/10.1016/j.jallcom.2018.12.136
1b. H. Zhang, Y. Zhang, D. Dai, M. Cao, W. Shen, *Mater. Des.* **2016**, *92*, 371. https://doi.org/10.1016/j.matdes.2015.12.081
1c. Z. Pei, J. Yin, J. A. Hawk, D. E. Alman, M. C. Gao, *NPJ Comput. Mater.* **2020**, *6*, 50. https://doi.org/10.1038/s41524-020-0308-7
2a. P. R. Kaundinya, K. Choudhary, S. R. Kalidindi, *Phys. Rev. Mater.* **2021**, *5*, 063802. https://doi.org/10.1103/PhysRevMaterials.5.063802
2b. J. R. Moreno, J. Flick, A. Georges, *Phys. Rev. Mater.* **2021**, *5*, 083802. https://doi.org/10.1103/PhysRevMaterials.5.083802
2c. B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski, T. Y.-J. Han, *NPJ Comput. Mater.* **2019**, *5*, 108. https://doi.org/10.1038/s41524-019-0248-2
2d. P. Borlido, J. Schmidt, A. W. Huran, F. Tran, M. A. L. Marques, *NPJ Comput. Mater.* **2020**, *6*, 96. https://doi.org/10.1038/s41524-020-00360-0
2e. V. Gladkikh, D. Y. Kim, A. Hajibabaei, A. Jana, C. W. Myung, K. S. Kim, *J. Phys. Chem. C.* **2020**, *124*, 8905. https://doi.org/10.1021/acs.jpcc.9b11768
3. A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL. Mater.* **2013**, *1*, 01102. https://doi.org/10.1063/1.4812323
4. S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, *NPJ Comput. Mater.* **2015**, *1*, 15010. https://doi.org/10.1038/npjcompumats.2015.10
5. DICE Homepage [accessed 2024 Feb 28]. https://dice.nims.go.jp.
6. G. Pilania, *Comp. Mater. Sci.* **2021**, *193*, 110360. https://doi.org/10.1016/j.commatsci.2021.110360
7. Y. Zhang, C. Ling, *NPJ Comput. Mater.* **2018**, *4*, 25. https://doi.org/10.1038/s41524-018-0081-z
8. N. Fujinuma, B. DeCost, J. Hattrick-Simpers, S. E. Lofland, *Commun. Mater.* **2022**, *3*, 59. https://doi.org/10.1038/s43246-022-00283-x
9. B. Krawczyk, *Prog. Artif. Intell.* **2016**, *5*, 221. https://doi.org/10.1007/s13748-016-0094-0
10a. J. G. Avelino, G. D. Cavalcanti, R. M. Cruz, *Artif. Intell. Rev.* **2024**, *57*, 82. https://doi.org/10.1007/s10462-024-10724-3
10b. V. Werner de Vargas, J. A. Schneider Aranda, R. dos Santos Costa, P. R. da Silva Pereira, J. L. Victória Barbosa, *Knowl. Inf. Syst.* **2023**, *65*, 31 https://doi.org/10.1007/s10115-022-01772-8

11. T. Lu, H. Li, M. Li, S. Wang, W. Lu, *J. Phys. Chem. Lett.* **2022**, *13*, 3032. https://doi.org/10.1021/acs.jpclett.2c00603
12. P. Borlido, T. Aull, A. W. Huran, F. Tran, M. A. L. Marques, *J. Chem. Theory Comput.* **2019**, *5*, 5069. https://doi.org/10.1021/acs.jctc.9b00322
13. Y. Zhuo, A. Mansouri Tehrani, J. Brgoch. *J. Phys. Chem. Lett.* **2018**, *9*, 1668. https://doi.org/10.1021/acs.jpclett.8b00124

14a. K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, K.-R. Müller, *J. Chem. Theory Comput.* **2013**, *9*, 3404. https://doi.org/10.1021/ct400195d

14b. R. Pollice, G. dos Passos Gomes, M. Aldeghi, J. Hickman Riley, M. Krenn, C. Lavigne, M. Lindner-D' Addario, A. Nigam, C. T. Ser, Z. Yao, A. Aspuru-Guzik, *Acc. Chem. Res.* **2021**, *54*, 849 https://doi.org/10.1021/acs.accounts.0c00785

14c. B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, A. Mehta, L. Ward, *Mol. Syst. Des. Eng.* **2018**, *3*, 819. https://doi.org/10.1039/C8ME00012C

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, E. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf
16. G. Lematre, F. Nogueira, C. K. Aridas, *J. Mach. Learn. Res.* **2017**, *18*, 1. https://jmlr.org/papers/volume18/16-365/16-365.pdf
17. G. Busca, *Catal. Today.* **2020**, *357*, 621. https://doi.org/10.1016/j.cattod.2019.05.011
18. J. E. Shelby, *Introduction to Glass Science and Technology*, 3rd edn. Royal Society of Chemistry, UK, **2021**.