Taylor & Francis
Taylor & Francis Group

Science and Technology of Advanced Materials: Methods

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tstm20

Effects of data bias on machine-learning–based material discovery using experimental property data

Masaya Kumagai, Yuki Ando, Atsumi Tanaka, Koji Tsuda, Yukari Katsura & Ken Kurosaki

To cite this article: Masaya Kumagai, Yuki Ando, Atsumi Tanaka, Koji Tsuda, Yukari Katsura & Ken Kurosaki (2022) Effects of data bias on machine-learning–based material discovery using experimental property data, Science and Technology of Advanced Materials: Methods, 2:1, 302-309, DOI: 10.1080/27660400.2022.2109447

To link to this article: https://doi.org/10.1080/27660400.2022.2109447

SCIENCE AND TECHNOLOGY OF ADVANCED MATERIALS: METHODS
2022, VOL. 2, NO. 1, 302–309
https://doi.org/10.1080/27660400.2022.2109447

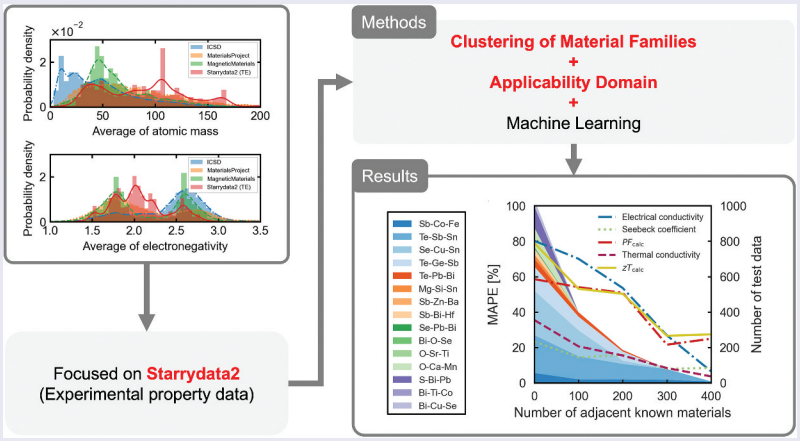
Effects of data bias on machine-learning–based material discovery using experimental property data

Masaya Kumagai[a,b,c], Yuki Ando[d], Atsumi Tanaka[e], Koji Tsuda[c,d,e], Yukari Katsura[c,d,e] and Ken Kurosaki[a,f]

[a]Institute for Integrated Radiation and Nuclear Science, Kyoto University, Sennan-gun, Osaka, Japan; [b]SAKURA Internet Research Center, SAKURA Internet Inc, Tokyo Tatemono Umeda Building 11F, 1-12-12, Umeda, Kita-ku, Osaka, 530-0001, Japan; [c]Center for Advanced Intelligence Project, RIKEN, Chuo-ku, Tokyo, Japan; [d]Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science (NIMS), Tsukuba, Japan; [e]Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba, Japan; [f]Research Institute of Nuclear Engineering, University of Fukui, Tsuruga, Fukui, Japan

ABSTRACT

Materials informatics (MI) research, which is the discovery of new materials through machine learning (ML) using large-scale material data, has attracted considerable attention in recent years. However, in general, the large-scale material data used in MI are biased owing to differences in the targeted material domains. Moreover, most studies on MI have not clearly demonstrated the influence of data bias on ML models. In this study, we clarify the influence of data bias on ML models by combining the concept of the applicability domain and clustering for large-scale experimental property data in the Starrydata2 material database previously developed by our group. The results show that data bias influences the error and reliability of the predictions made by the ML model. The predictions of the ML model within the applicability domain are highly reliable compared to those made outside the domain. This indicates that the material space that can be reliably discovered by the constructed ML model is limited. Nonetheless, we apply the ML model to a large dataset comprising various material classes and find that new materials similar to known materials can be proposed within a limited space. Thus, our findings demonstrate the importance of considering data bias when constructing and evaluating ML models in MI.



ARTICLE HISTORY
Received 10 March 2022
Revised 27 June 2022
Accepted 1 August 2022

KEYWORDS
Machine learning; material informatics; large-scale material data; data bias

1. Introduction

With the rapid development of computer technology and machine-learning algorithms in information science, research on material informatics (MI), which integrates materials science and information science, has attracted considerable attention. In particular, studies have actively focused on the discovery of new materials through machine learning (ML) using large material datasets, and applications in various material fields have been reported, such as magnetic refrigeration materials [1], energy materials [2,3], shape memory alloys [4], and superalloys [5]. In general, the holdout method or k-fold cross-validation method, in which the training data and validation data are randomly divided, is used to evaluate the performance of ML models. These validation methods evaluate the performance assuming that the unknown data are similar to the known data because the feature spaces of the validation data and the training data are similar. For example, in the case of predicting a user's

CONTACT Masaya Kumagai kumagai.masaya.3n@kyoto-u.ac.jp Institute for Integrated Radiation and Nuclear Science, Kyoto University, 2, Asashiro-Nishi, Kumatori, Sennan-gun, Osaka 590-0494, Japan

Supplemental data for this article can be accessed online at https://doi.org/10.1080/27660400.2022.2109447.

movie preferences, performance evaluation using these methods is effective because we can assume that most users have a feature space similar to that of many others. However, in the case of MI, unknown materials exhibiting properties that have never been seen before can be often discovered, i.e. properties that are not similar to those of known materials. However, extrapolative predictions are technically difficult. Therefore, the material space in which reliable interpolative predictions are possible must be clearly indicated.

In addition, the large material datasets generally used in MI have data bias owing to differences in the material domain of interest. Figure 1 shows a histogram of the average atomic mass and electronegativity based on the chemical composition of the material datasets used in MI [1,6–8]. Even in the same material-science field, data bias exists based on the dataset used [9]. For example, the thermoelectric material (TE) dataset of Starrydata2 includes semiconductors containing heavy metals known to exhibit high-thermoelectric performance, whereas ICSD includes many insulators containing light elements, as can be observed from their respective density distributions. Moreover, the frequency counts of the elements in the chemical compositions (Figure S1 in Supplemental Materials) confirm that the distribution of the major elements in each data set differs. Therefore, performance evaluation of ML models in MI should consider the data bias of the dataset used and clearly indicate the material space evaluated. Nevertheless, MI studies often evaluate the performance of ML models in an ambiguous material space that shows only the names of the datasets used. Some studies [10,11] have made the material space explicit by artificially focusing on the research scope (e.g. material families); however, the diversity of the new materials that must be discovered is lost.

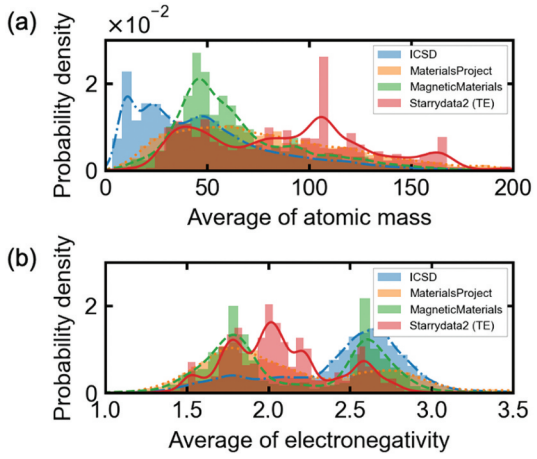


Figure 1. Histograms and density distributions of the averaged (a) atomic masses and (b) electronegativities based on the chemical compositions in the widely used datasets in MI.

We focus on the applicability domains (ADs) used in quantitative structure – activity relationship studies [12]. AD mechanically defines the material space to which the ML model can be applied based on its similarity with the known materials used for training. This implies defining a material space that can be interpolated with high-reliability. However, AD not only lacks human interpretability because it is defined mechanically, but also limits the scope of discovery for new materials. In this study, we define a wide AD using a large-scale dataset containing various material families, and provide human interpretability by clustering within the AD. The obtained interpretability is used to clarify the influence of data bias on the ML model and evaluate the possibility of discovering new materials within the AD. The contribution of this study is the demonstration of the importance of considering data bias in MI for constructing and evaluating ML models.

2. Methods

2.1 Data preparation and construction of the machine learning model

We used experimental property data from Starrydata2, a material database developed by our group [8,13], comprising 42,005 samples and 2,236,338 records extracted from 7,698 papers as on 27 September 2021 [14]. This is the largest dataset that includes the temperature dependence of the experimental properties in the field of thermoelectric materials, worldwide. Although there is a data bias toward the thermoelectric field (Figure 1), this dataset has a comprehensive collection of experimental data for various material families, performances, and temperature ranges, which reduces the selection bias compared to other experimental value databases [15,16]. All the data from Starrydata2 are available for free on GitHub [14]. This includes two types of data: The first is raw data extracted directly from the figures in a paper; the second includes data extracted only from the physical property values where the x-axis denotes the temperature, and each property value is interpolated by a 5-th order polynomial for every 50 K temperature point. In this study, the data in Starrydata2 as on 27 September 2021, interpolated by the 5-th order polynomial, were used for training and validation. In particular, only thermoelectric property data (294,616 records) from 6,594 papers on thermoelectric materials were used in this study.

The performance of thermoelectric materials was evaluated as a dimensionless figure of merit, $zT=S^2\sigma/\kappa$, where S is the Seebeck coefficient, σ is the electrical conductivity, κ is the thermal conductivity, and T is the temperature. The discovery of high-performance thermoelectric materials involves searching for

materials with a high-power factor ($PF = S^2\sigma$) and low κ. For several years, degenerate semiconductors based on heavy metals such as Bi, Pb, Te, and Sb have been the mainstream materials exhibiting high zT. Numerous other material systems, such as skutterudite and clathrate compounds, Zintl phase and Heusler compounds, and oxides and silicide compounds, have been reported as thermoelectric materials and are included in the dataset used in this study.

Because the data in Starrydata2 are extracted from previously published papers, numerous errors exist in the papers themselves and in the extraction process. Therefore, the data were strictly preprocessed using the following two procedures. These procedures are expected to provide reasonable values for T, S, σ, κ, and zT.

(1) Remove the records in which at least T, S, σ, κ, or zT is missing (63,405 records).
(2) Remove the records for which the mean absolute percentage error rate (MAPE) between zT_{calc} calculated from T, S, σ, and κ, and zT extracted directly from the paper is greater than 5% (35,393 records).

The input vectors used for ML included 26 feature values based on the chemical composition and the (measured) temperature T. The features were the calculated mean, variance, and difference (maximum value-minimum value within the major elements) of the 'group', 'period', 'atomic number', 'Mendeleev number', 'atomic radius', 'atomic weight', 'electro negativity', and 'VEC' of the containing elements, and 'number of containing elements' and 'number of major elements'. Here, the major elements were those with a ratio of 0.1 or greater in the chemical composition such that the overall ratio was unity [17]. It is interesting to note that as Starrydata2 records the temperature dependence of the experimental physical properties, the temperature can be added to the input vector. Features other than the temperature, are similar to the Magpie/Matminer [18,19] feature set. However, for the material in Starrydata2, several small amounts of additive elements are often introduced to modulate the carrier concentration or reduce lattice thermal conductivity. Moreover, there are complex chemical compositions consisting of up to 10 elements. The 'range', 'minimum', and 'maximum' attributes in the Magpie/Matminer feature set may overestimate the effect of these additive elements, and were therefore excluded from the input vector. In addition, the electronic structure attributes and ab initio calculated attributes were excluded from our feature set to avoid overcomplicating the feature space. Records that could not be converted into feature vectors were deleted (35,332 records). The Python library for material analysis, pymatgen [20], was used to create records of various features. The target variables were thermoelectric properties S, σ, κ, and zT_{calc}. The training and test data were split based on the publication year of the paper rather than the usual random split. Data published before 2020 were used as training data (21,775 records), whereas those from 2021 onward were used as test data (563 records). Thus, the problem was set as predicting materials in 2021 that were unknown before 2020. Because the materials up to 2020, which would normally have been included in 2021, are completely removed, the test data include only completely unknown chemical compositions. Note that the test data is used to evaluate the generalization performance, and is different from the validation data used to optimize the hyperparameters.

For the ML method, we used the random forest algorithm. This is an ensemble learning algorithm, which combines multiple decision trees to improve the generalization performance. In addition, it is robust to noise and overfitting, and has often been used in previous research in materials science [21,22]. The hyperparameters were optimized using the grid search provided by Scikit-learn [23]. The MAPE and root mean squared logarithmic error (RMSLE) were used to evaluate the prediction accuracy.

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{t_i - y_i}{y_i}\right|, \tag{1}$$

$$\mathrm{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(log(t_i+1) - log(y_i+1)\right)^2}, \tag{2}$$

where n is the number of test data points, t is the experimental value, and y is the predicted value. As target variables S and σ are physical properties with a wide range of variation in the order of magnitude, the absolute error and root mean square error, which are generally used, are not used in this study.

In order to propose new materials, the ML model constructed in this study was applied to the chemical compositions recorded in the Materials Project [7] to predict various thermoelectric properties. We used all the Materials Project data as on 18 October 2018 (83,989 records). Note that only the chemical compositions of stable materials (energy above hull > = 15 meV), including known metastable materials, and the materials included in the AD were used.

2.2 Clustering of material families

To clarify the material space of the dataset, soft clustering using variational Bayesian estimation of Gaussian mixture distribution was used. In materials science, the use of soft clustering is reasonable

because some materials belong to multiple material family clusters. For variational Bayesian estimation of the mixture Gaussian distribution, we used an algorithm provided by the Python ML library, Scikit-learn [23]. The variance-covariance matrix of the Gaussian distribution was used as the common variance-covariance matrix for each cluster (tied type) [24], and the number of clusters was set to 15. The input vector used for clustering is the vector described in Section 2.1. The number of clusters was arbitrarily determined as a value that could be identified by thermoelectric material experts. To clarify the characteristics of each cluster, the elements contained in the materials within the clusters were counted, and the top three elements with the highest number of occurrences were assigned labels.

2.3 Applicability domain using the k-nearest neighbor method

There are various types of ADs, such as Bounding Box, Convex Hull, Clustering, and the k-nearest neighbor method, depending on how the domain is defined [12,25]. Bounding Box is the simplest and fastest method for defining an n-dimensional hyper-rectangle as an AD based on the maximum and minimum values of each descriptor. However, due to the simple box of the defined feature space, several false positive domains are generated when the data points are nonuniformly distributed. Convex Hull is a method that improves on the Bounding Box using the smallest convex domain of the feature space as the AD; however, false positives may occur in concave regions when non-convex domains exist in the feature space. In addition, this method is computationally expensive. AD using Clustering, such as k-means, is computationally less expensive, and the definition of the AD is not as crude as that of the Bounding Box. Therefore, this method is effective for defining the AD in cases where reasonable clustering can be achieved. Moreover, in terms of clearly defining the learned range by Clustering, it is similar to the leave-one-cluster-out (LOCO) cross-validation proposed by Bryce Meredig et al [26]. Therefore, it is also an effective tool for evaluating extrapolation. However, false positive domains may occur when the training data points are non-uniformly distributed and the cluster center of the gravity point is not in a data dense domain. AD with k-nearest neighbors is a method that defines regions using the average distance from the k nearest neighbors in all training data points as the threshold. This method is more rigorous than its other counterparts because the AD is clearly defined by the distance between each training data and the test data. In this study, we focused on the AD with the k-nearest neighbor method to rigorously discuss the effects of data bias. Clustering in Section 2.2 was not used to define AD but only to identify material families for the purpose of providing human interpretability.

Among the k-nearest neighbor methods, we used the approach proposed by Sahigara et al [27]. In general, a single threshold (e.g. the 95th percentile) is determined based on the average distance around k for all the data points. However, a single threshold may be determined to be biased toward data points in regions of high data point density (e.g. existing high-performance materials) and may not fit regions of low data point density (e.g. new materials). In the methods proposed by Sahigara et al., thresholds are set for each data point, and AD is determined individually. In addition, the value of the number of neighbors, k, is important for determining the AD using the k-neighborhood method. A low value of k leads to overly strict limitations, while a high value unnecessarily expands the AD. Therefore, k is determined by maximizing the percentage of validation data retained in the AD using the Monte Carlo method. In this study, we determined the optimal k by randomly dividing the AD 1,000 times such that 20% (4,355 records) of the training data (21,775 records) became the validation data, and obtained the percentage of data retained in the AD for each different k.

3. Results and discussion

Figure 2 shows a stacked face chart of the number of training data for each year of publication. The colors on the faces represent the results of clustering using variational Bayesian estimation of the Gaussian mixture distributions. Interestingly, despite the clustering being performed by unsupervised learning, existing thermoelectric material families, such as skutterudite and silicide compounds, Zintl and Heusler phases, clathrate compounds, and oxides, were identified as clusters. This indicates that the input vectors used for

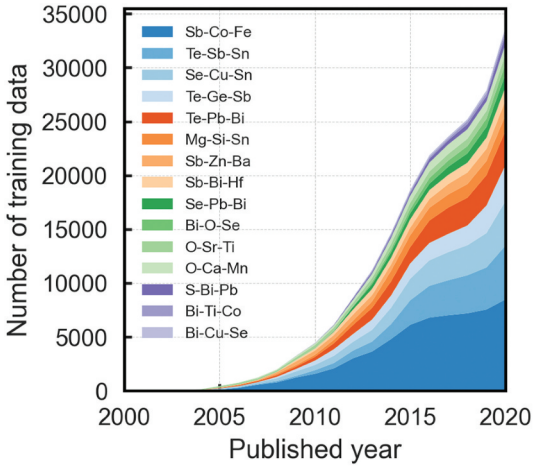


Figure 2. Dependence of the number of training data on the publication year of the paper. The number of training data is stacked with the results of material family clustering.

clustering and property predictions include the necessary features for distinguishing the representative material families in the thermoelectric materials field. The Starrydata2 dataset contains numerous Te-based and skutterudite compounds. Note that Figure 2 does not accurately represent the history of the material families in thermoelectric materials because it is a clustering result obtained using the preprocessed dataset.

The results of the Monte Carlo method are shown as box plots in Figure 3. For a low k, less data are retained in the AD, leading to a reduction in the data diversity. The size of the box (quartile range) decreases with an increase in k. Because the box is small and the AD can hold a large amount of data, a k value of nine or more is optimal. Finally, to avoid unnecessary expansion of the AD, we set $k = 9$.

Figure 4 indicates the relationship between the experimental and predicted values of the various thermoelectric properties using the test data. The distribution around the diagonal line indicates that the prediction accuracy is higher. The darker color of the plots indicates that the number of adjacent known materials in the AD is larger. The number of adjacent known materials has been explained as the reliability of prediction in previous studies [27]. The prediction error for the test data within the AD is less than or equal to that for the test data outside the AD (Table S1

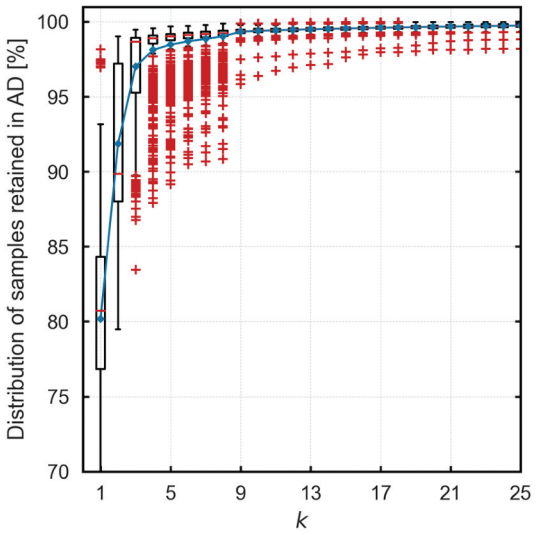


Figure 3. Box-And-Whisker plot of the test samples (%) retained within the AD for different k-values during k-optimization.

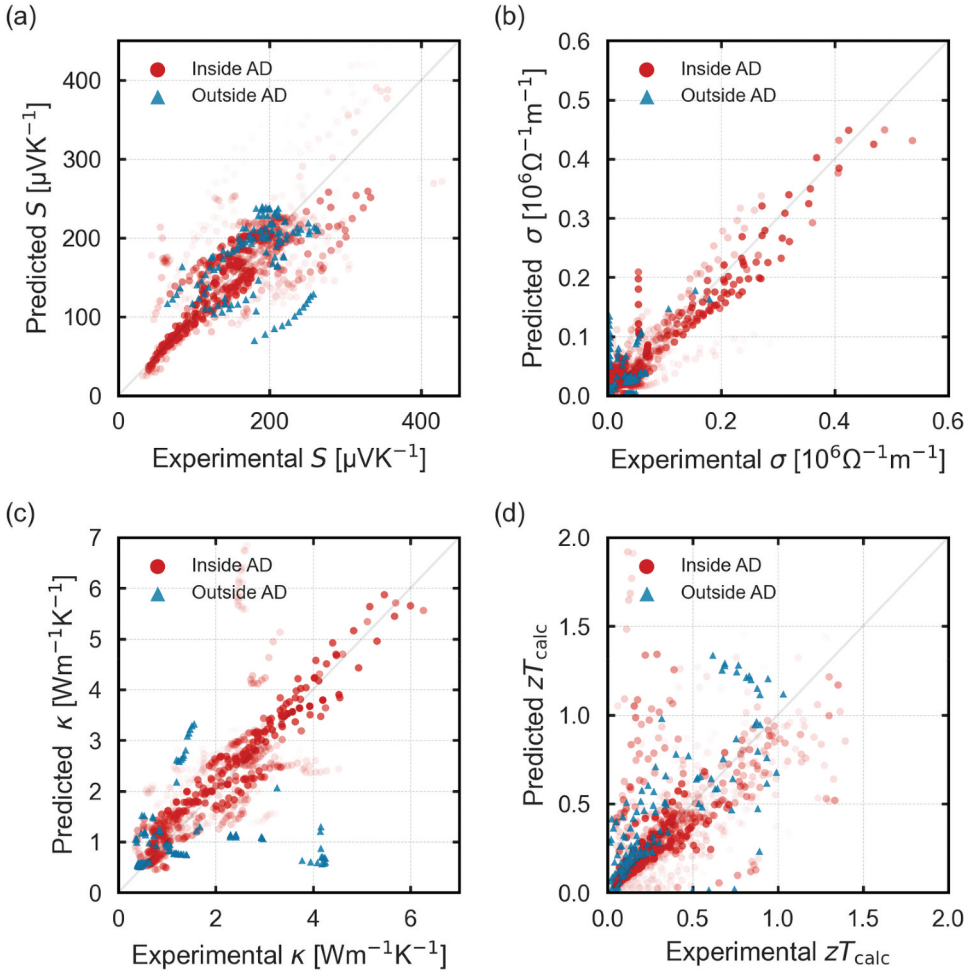


Figure 4. Actual vs. predicted TE properties for the test data. The darker the plot color, the more are the number of neighboring known materials in the AD.

in Supplemental Materials). This is in good agreement with the fact that the data outside the AD ('Outside AD' in Figure 4) are distributed farther from the diagonal; the data within the AD ('Inside AD' in Figure 4) are distributed on the diagonal as the number of adjacent known materials increases. This indicates that the defined AD is appropriate and the constructed ML model correctly predicts within the AD. In Figure 4(c), data points can be observed away from the diagonal, which are the half-Heusler alloys $ZrCoSb_{0.9-x}Sn_{0.1}Te_x$ [28]. The reason for the large prediction error is that there are several other half-Heusler alloys that show high-thermal conductivity, and the number of training data is less.

Figure 5 shows the dependence of (a) MAPE and (b) RMSLE on the number of adjacent known materials in the AD, and a stacked plane graph (second axis) of the number of test data (unknown materials). The test data used for evaluation is from 2021. All the chemical compositions in the test data are not included in the training data. The number of unique chemical compositions is 176. The chemical compositions with the highest number of adjacencies are mostly derived from studies on skutterudite and Te-base (such as SnTe [29] and PbTe [30]), which have the highest number of recordings in Figure 2. The chemical compositions with fewer adjacencies are relatively new material groups such as Mg_3Sb_2-Mg_3Bi_2 alloy [31] and $Zn_2Cu_3In_3Te_8$ [32]. As shown in Figure 5(a), the MAPE decreases with the increase in the number of known neighboring materials, but the diversity of the material families also decreases. This indicates that reliable predictions can be made for material families with a large amount of training data; however, there is a risk of unreliable predictions for material families with a small amount of training data. This also implies that the existence of clusters with high-data density may increase the apparent prediction accuracy when the general performance is evaluated, where the training and test data are randomly divided. In the case of ML models proposing new materials from diverse material families, the prediction accuracy and reliability are likely to be low. In addition, we can confirm that the error in σ is larger than those of the other properties. This is clearly shown in Figure 5(b), which depicts the log scale error. Because the order of magnitude of σ significantly varies depending on the material, the prediction error is considered to be larger than that of the other properties. PF and zT, which contain σ, also have larger errors.

Figure 6 shows the heat map of S in descending order at each temperature (300–900 K) applied to the ML model constructed for the chemical compositions of the Materials Project data included in the AD. The chemical compositions contained in the training data are masked in gray. The predicted materials at 300 K and 900 K have fewer adjacencies than those predicted at the other temperatures. This is due to the small number of data recorded below 300 K and above 900 K in Starrydata2. The reason for the large number of data from 400 K to 800 K can be explained in terms of the temperature range of common measurement equipment and material stability. In addition, most materials predicted to have high-S are materials with a bandgap. This result agrees well with the characteristics of materials showing high-S. $Li_2Ti_3VO_8$ and $Mg_{11}Ti_{25}O_{60}$ are considered to be new materials that are not derivatives of the known materials in the training data, although the prediction errors are large due to the small number of adjacencies. $InAgSe_2$ is not included in the training data; however, other chalcopyrite material families such as $CuInSe_2$ are included and therefore, there is a higher number of adjacencies. The Seebeck coefficient of $Ag_{1+x}InSe_2$ measured by Pengfei Qiu et al. in 2017 [33] is in approximate agreement with the predictions. The heat map shows only the Seebeck coefficient, which has the least prediction error; however, it can be created for the other properties as well. Although there are certain limitations in the

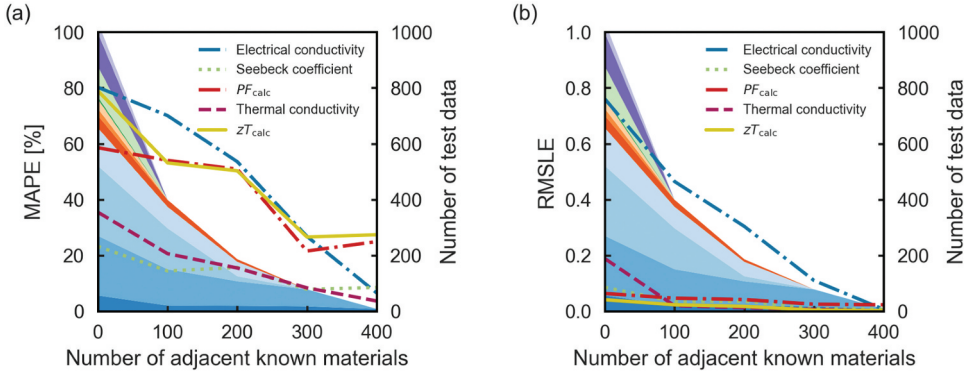


Figure 5. Dependence of (a) MAPE (line chart) and (b) RMSLE (line chart) on the number of adjacent known materials in the AD and the number of test data (unknown materials) on the stacked surface graph (second axis). The colors of the faces are identical to those of the material system clusters in Figure 2.

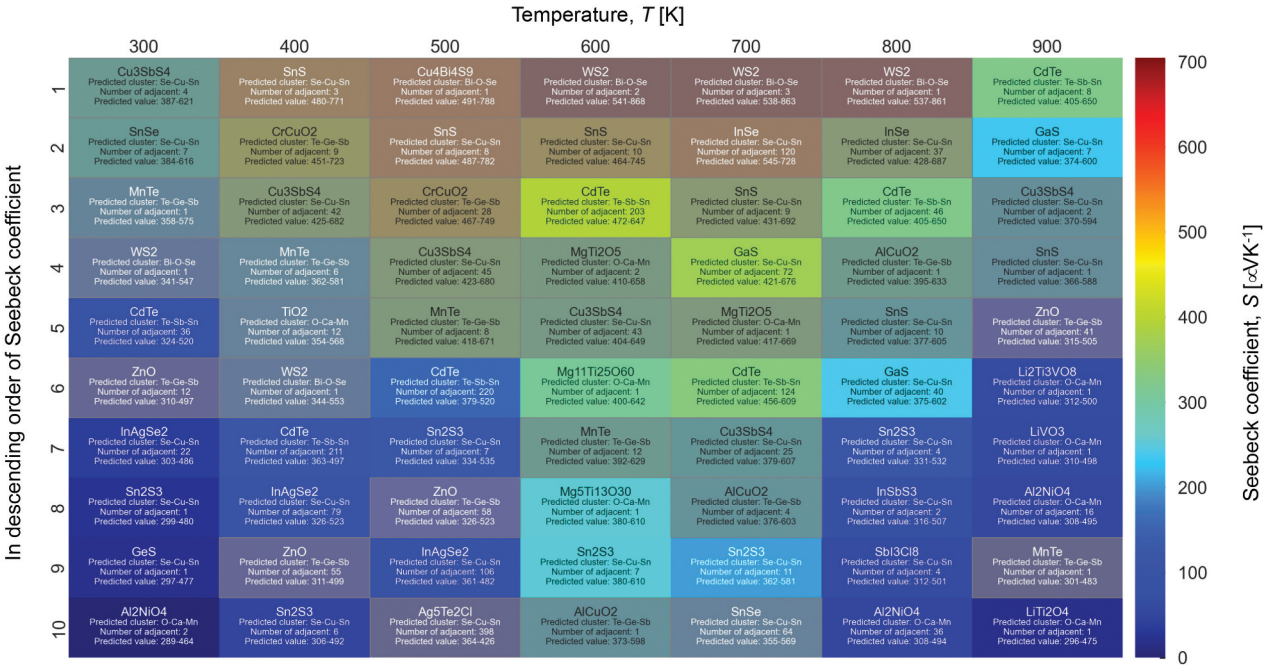


Figure 6. Top-10 predicted *S* values at each temperature (300 K–900 K) for the chemical composition of the Materials Project data included in the AD. The chemical compositions contained in the training data are masked in gray. At the bottom of each block, the cluster to which the chemical composition belongs, the number of adjacent known materials, and the range of the predicted *S* calculated from the MAPE are shown.

temperature range and material families when applying the AD, it is demonstrated that various new materials can be proposed using a large dataset containing a variety of material families.

4. Conclusions

In this study, the influence of data bias on the ML model was clarified using a combination of the concept of AD and clustering for a large-scale experimental physical property dataset recorded in Starrydata2, a web system originally developed by our group. We confirmed that our ML model could make reliable predictions for unknown materials similar to Te-based compounds and skutterudite, included in many of the currently available material datasets. The prediction accuracy significantly decreased outside the defined AD, and the error decreased within the AD as the number of neighboring known materials increased. These results suggest that the existence of data bias influences the error and prediction reliability of ML models. However, despite the limitations of the AD, it was possible to propose various new materials using a large dataset containing a variety of material families. The combined analysis of AD and clustering in this study effectively clarified the influence of bias. The results of this study not only show the importance of constructing and evaluating ML models considering the data bias but also the importance of creating diverse and large-scale material data with less bias. The python codes implemented in this study are available on https://github.com/kumagallium/matCL-knnAD.

Disclosure statement

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

Funding

This work was supported by JSPS KAKENHI [grant number JP20K22466].

ORCID

Yukari Katsura http://orcid.org/0000-0002-8905-2995
Ken Kurosaki http://orcid.org/0000-0002-3015-3206

References

[1] Baptista de Castro P, Terashima K, Yamamoto TD, et al. Machine-learning-guided discovery of the gigantic magnetocaloric effect in HoB_2 near the hydrogen liquefaction temperature. Npg Asia Mater. 2020;12(1):1–7.
[2] Pilania G, Balachandran PV, Kim C, et al. Finding new perovskite halides via machine learning. Front Mater. 2016;3:19.
[3] Oliynyk AO, Mar A. Discovery of intermetallic compounds from traditional to machine-learning approaches. Acc Chem Res. 2017;51:59–68.
[4] Xue D, Balachandran PV, Hogden J, et al. Accelerated search for materials with targeted properties by adaptive design. Nat Commun. 2016;7(1):1–9. DOI:10.1038/ncomms11241.

[5] Conduit BD, Jones NG, Stone HJ, et al. Design of a nickel-base superalloy using a neural network. Mater Design. 2017;131:358–365.
[6] Hellenbrandt M. The inorganic crystal structure database (Icsd)—present and future. Crystallogr Rev. 2004;10(1):17–22.
[7] Jain A, Ong SP, Hautier G, et al. Commentary: the Materials Project: a materials genome approach to accelerating materials innovation. APL Mater. 2013;1(1):011002. DOI:10.1063/1.4812323.
[8] Katsura Y, Kumagai M, Kodani T, et al. Data-Driven analysis of electron relaxation times in PbTe-type thermoelectric materials. Sci Technol Adv Mater. 2019;20 (1):511–520. DOI:10.1080/14686996.2019.1603885.
[9] Barnard AS. Best practice leads to the best materials informatics. Matter. 2020;3(1):22–23.
[10] Pilania G, Mannodi-Kanakkithodi A, Uberuaga BP, et al. Machine learning bandgaps of double perovskites. Sci Rep. 2016;6(1):1. DOI:10.1038/srep19375.
[11] Dey P, Bible J, Datta S, et al. Informatics-Aided bandgap engineering for solar materials. Comput Mater Sci. 2014;83:185–195.
[12] Kar S, Roy K, Leszczynski J. Applicability domain: a step toward confident predictions and decidability for QSAR modeling. In: Nicolotti O, editor. Computational toxicology. Methods in molecular biology. New York (NY): Springer; 2018. p. 141–169.
[13] [cited 2022 Feb 21]. Available from: https://www.starrydata2.org/
[14] [cited 2022 Feb 21]. Available from: https://github.com/starrydata/starrydata_datasets/tree/master/datasets
[15] Gaultois MW, Sparks TD, Borg CKH, et al. Data-Driven review of thermoelectric materials: performance and resource considerations. Chem Mater. 2013;25(15):2911–2920. DOI:10.1021/cm400893e.
[16] Shi L, Zhang S, Arshad A, et al. Thermo-physical properties prediction of carbon-based magnetic nanofluids based on an artificial neural network. Renewable Sustainable Energy Rev. 2021;149:111341.
[17] Na GS, Jang S, Chang H. Predicting thermoelectric properties from chemical formula with explicitly identifying dopant effects. Npj Comput Mater. 2021;7(1):1.
[18] Ward L, Agrawal A, Choudhary A, et al. A general-purpose machine learning framework for predicting properties of inorganic materials. Npj Comput Mater. 2016;2(1):16028. DOI:10.1038/npjcompumats.2016.28.
[19] Ward L, Dunn A, Faghaninia A, et al. Matminer: an open source toolkit for materials data mining. Comput Mater Sci. 2018;152:60–69.
[20] Ong SP, Davidson Richards W, Jain A, et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. Comput Mater Sci. 2013;68:314–319.
[21] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–2830.
[22] Furmanchuk A, Saal JE, Doak JW, et al. Prediction of Seebeck coefficient for compounds without restriction to fixed stoichiometry: A machine learning approach. J Comput Chem. 2017;39:191–202.
[23] Revi V, Kasodariya S, Talapatra A, et al. Alankar a machine learning elastic constants of multi-component alloys. Comput Mater Sci. 2021;198:110671.
[24] Shimizu N, Kaneko H. Direct inverse analysis based on Gaussian mixture regression for multiple objective variables in material design. Mater Design. 2020;196:109168.
[25] Stanforth RW, Kolossov E, Mirkin B. A measure of domain of applicability for QSAR modelling based on intelligent K-means clustering. QSAR Combi Sci. 2007;26(7):837–844.
[26] Meredig B, Antono E, Church C, et al. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. Mol Syst Des Eng. 2018;3 (5):819–825. DOI:10.1039/C8ME00012C.
[27] Sahigara F, Ballabio D, Todeschini R, et al. Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. J Cheminform. 2013;5(1):1–9. DOI:10.1186/1758-2946-5-27.
[28] Kumar A, Chaturvedi KM, Bano S, et al. Enhanced thermoelectric performance of p-type $ZrCoSb_{0.9}Sn_{0.1}$ via Tellurium doping. Mater Chem Phys. 2021;258:123915.
[29] Lyu T, Yang Q, Meng F, et al. Enhancing thermoelectric performance of $Sn_{1-x}Sb_{2x/3}Te$ via synergistic charge balanced compensation doping. J Chem Eng. 2021;404:126925.
[30] Parashchuk T, Horichok I, Kosonowski A, et al. Insight into the transport properties and enhanced thermoelectric performance of n-type $Pb_{1-x}Sb_xTe$. J Alloys Compd. 2021;860:158355.
[31] Chen Y, Wang C, Ma Z, et al. Improved thermoelectric performance of n-type Mg_3Sb_2–Mg_3Bi_2 alloy with Co element doping. Curr Appl Phys. 2021;21:25–30.
[32] Zhang Q, Xi L, Zhang J, et al. Influence of Ag substitution on thermoelectric properties of the quaternary diamond-like compound $Zn_2Cu_3In_3Te_8$. J Materiomics. 2021;7(2):236–243. DOI:10.1016/j.jmat.2020.09.005.
[33] Qiu P, Qin Y, Zhang Q, et al. Intrinsically high thermoelectric performance in $AgInSe_2$ n-Type diamond-like compounds. Adv Sci. 2017;5 (3):1700727.