Descriptors for predicting the lattice constant of body centered cubic crystal

Keisuke Takahashi,^{1,2,a)} Lauren Takahashi,³ Jakub D. Baran,⁴ and Yuzuru Tanaka¹

¹Center for Materials research by Information Integration (CM²), Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science (NIMS), 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan

²Graduate School of Engineering, Hokkaido University, N-13, W-8, Sapporo 060-8628, Japan

³Freelance Researcher, Central Ward, Sapporo 064, Japan

⁴Department of Chemistry, University of Bath, Claverton Down, Bath BA2 7AY, United Kingdom

(Received 24 February 2017; accepted 11 May 2017; published online 24 May 2017)

The prediction of the lattice constant of binary body centered cubic crystals is performed in terms of first principle calculations and machine learning. In particular, 1541 binary body centered cubic crystals are calculated using density functional theory. Results from first principle calculations, corresponding information from periodic table, and mathematically tailored data are stored as a dataset. Data mining reveals seven descriptors which are key to determining the lattice constant where the contribution of descriptors is also discussed and visualized. Support vector regression (SVR) technique is implemented to train the data where the predicted lattice constants have the mean score of 83.6% accuracy via cross-validation and maximum error of 4% when compared to experimentally determined lattice constants. In addition, trained SVR is successful in predicting material combinations from a desired lattice constant. Thus, a set of descriptors for determining the lattice constant is identified and can be used as a base descriptor for lattice constants of further complex crystals. This would allow for the acceleration of the search for lattice constants of desired atomic compositions as well as the prediction of new materials based on a specified lattice constant. *Published by AIP Publishing.* [<http://dx.doi.org/10.1063/1.4984047>]

I. INTRODUCTION

With the rapid accumulation, growth, and availability of high quality material data, information science has the potential to accelerate the material discovery process and can possibly change how material science is treated.^{1–3} In principle, prediction of material properties and the design of new materials can be made achievable if trends or repeated sequences in material data, referred to as a material genome, are revealed.^{4–7} In material informatics, a material genome for a material is often represented by descriptors that determine the structural and material properties of a material. The source of such descriptors can range from basic information found within the periodic table to data acquired experimentally or computationally.^{6,7} If sets of descriptors representing structural and material properties of materials are identified, implementing data mining algorithms such as machine learning can be effective for predicting structural and material properties.^{8,9} However, while independent descriptors may participate in the prediction of multiple properties, a set of descriptors is mainly unique to the particular property. Thus, it is important to uncover sets of descriptors for particular properties in order to be able to predict all the properties of a given material.

The lattice constant is a fundamental part of the structural information related to crystal structures. By understanding the lattice constant, it can be expanded upon and used in order to

estimate the volume and density of a crystal, thereby allowing for a quicker establishment of the structural information of a material. This is quite important when designing new materials as specific structures are often demanded during the material design process. In general, the lattice constant is determined experimentally by using X-ray diffraction while first principle calculations are generally used to calculate the lattice constant in computational material science. If data science is able to determine the descriptor for the lattice constant, it would not only accelerate the prediction of the lattice constant but also allow the prediction of material combinations based on a desired lattice constant. Here, a material dataset is constructed using first principle calculations and descriptors for determining the lattice constant are explored. Once descriptors for the lattice constant are determined, machine learning is implemented in order to predict the lattice constant.

The workflow for predicting the lattice constant using data science is proposed as shown in Fig. 1. The lattice constant of a binary body centered cubic (BCC) crystal is investigated as a prototype case where the descriptors representing the lattice constant are explored. A dataset is prepared using first principle calculations and corresponding information from the periodic table. In particular, a binary system within BCC is considered where a total of 1541 element combinations with a 1:1 element ratio are calculated.¹⁰ It is important that the material dataset includes unrealistic materials such as the binary compound AgHe where He has a closed shell structure, as it is unlikely to form solid. Such unrealistic compounds give a great advantage when the machine learns the dataset as the machine can

^{a)}keisuke.takahashi@eng.hokudai.ac.jp

FIG. 1. Proposed workflow for predicting the lattice constant in data science where a dataset is constructed using first principle calculation and machine learning is implemented for prediction. Binary element AB in body centered cubic crystal is used where the ratio of AB is 1 to 1.

then classify realistic and unrealistic cases.⁶ Once the dataset is prepared, the search for descriptors determining the lattice constant is carried out and machine learning is implemented in order to learn the discovered descriptors and predict the lattice constant. Direct prediction of the lattice constant from descriptors would greatly increase the rate of material prediction, which can aid both experiments and first principle calculations.

II. COMPUTATIONAL METHOD

The dataset is prepared on the basis of the grid based projector augmented wave (GPAW) algorithm within density functional theory calculations.¹¹ The exchange correlation of Perdew–Burke–Ernzerhof (PBE) and spin polarization are applied.¹² Periodic boundary conditions are applied for all directions where $4 \times 4 \times 4$ special k points of the Brillouin zone sampling are applied.¹³ Binary compounds in a BCC crystal are considered where a $(1 \times 1 \times 1)$ cell is constructed and the ratio of elements is set to 1:1. Lattice optimization is performed in order to find the ground state lattice constant for each material.¹⁰ A total of 1541 binary materials in a BCC crystal of element AB as shown in Fig. 1 are calculated where A represents Ag, Al, As, Au, Co, Cr, Cu, Fe, Ga, Li, Mg, Na, Ni, Os, Pd, Pt, Rh, Ru, Si, Ti, V, W, and Zn and B represents the elements with atomic numbers 1–42, 44–57, and 72–83. Calculated results such as the lattice constant, formation energy, bulk modulus, magnetic moment, and bond distance as well as corresponding information from the periodic table such as the atomic radius, electronegativity, and electron affinities are stored in a dataset.

III. DATA MINING METHOD

Machine learning within data mining is implemented in order to explore the descriptors from a constructed material dataset. In particular, support vector regression (SVR) within scikit-learn is implemented for training the dataset where the radial basis functional (RBF) kernel is applied.¹⁴ The hyper

parameters of the SVR are set to the following: $C = 1.0$, $\epsilon = 0.1$, and $\text{kernel} = \text{rbf}$. Cross-validation is implemented alongside SVR in order to evaluate the accuracy of the trained SVR. Within cross-validation, the dataset is split to 10% test data set and 90% train data set. An average score of 10 random test and train data sets are evaluated along median score, highest score, and standard deviation. Various descriptors are tested and descriptors with the highest average score in cross-validation are determined.

IV. RESULT AND DISCUSSION

A. Descriptors search

A material dataset is constructed on the basis of first principle calculations where the calculated material properties are stored in a dataset. In addition to the computational results, the following corresponding information is also acquired from the periodic table: atomic number, atomic radius, electronegativity, electron affinity, atomic orbital, and number of valence electrons. Note that the atomic orbitals are converted to numerical values where s, p, d, and f orbitals are represented as 0, 1, 2, and 3, respectively. Furthermore, the above data are tailored by mathematically manipulating the data such as combining/subtracting/applying the power function, where the results are then defined as new variables. Including this type of tailored data is known for improving the machine learning process.⁶

Once the potential descriptor candidates are organized and stored in a material dataset, descriptors for determining the lattice constant are then explored using machine learning and the constructed dataset. In particular, SVR is implemented in order to train the constructed dataset where the objected variable is set to the lattice constant and the descriptors are set to various explanatory variables. Initial investigations of the data suggested that linear modeling do not accurately process the data and thus are not appropriate for this particular set of descriptors. Rather, nonlinear modeling is required due to the appearance of trends and groups in the descriptor variables, leading to the selection of SVR and the RBF kernel. Various combinations of explanatory variables from the constructed data set are considered as descriptors during the SVR process. Cross-validation is then implemented in order to evaluate the accuracy of the trained machine where the dataset is split into 90% trained data and 10% test data, and the score is calculated based on taking the average of 10 random trained and test datasets. The sets of descriptors with high scores are explored.

Various sets of descriptors for the lattice constant and the corresponding scores are collected in Table I. Descriptors of the atomic number of elements in AB in the binary BCC crystal and atomic density return a mean score of 80.1%. This indicates that atomic numbers and the density of the BCC crystal have a large contribution towards predicting the lattice constant. Adding the atomic orbital group (s, p, d, f) for each element greatly improves the mean score from 80.1% to 82.9%. Mathematically tailored descriptors are then added for further improvement of the score. In particular, the difference in electronegativity between elements A and B is added into a descriptor set which increases the mean score to 83.5%. Further data mining reveals that adding the tailored descriptor “Atomic Orbital B + Difference in Electronegativity between

TABLE I. Selected descriptors for support vector regression and the corresponding mean score, median score, highest score, and standard deviation (SD) of the mean score from cross-validation. A and B are the elements of AB in a binary body centered cubic crystal.

Descriptors	Mean (%)	Median (%)	Highest (%)	SD (%)
Atomic number A, atomic number B, density	80.1	81.0	85.2	3.8
Atomic number A, atomic number B, density, atomic orbital A, atomic orbital B	82.9	82.3	88.0	3.3
Atomic number A, atomic number B, density, atomic orbital A, atomic orbital B, difference in electronegativity between A and B	83.5	84.1	89.2	4.1
Atomic number A, atomic number B, density, atomic orbital A, atomic orbital B, difference in electronegativity between A and B, atomic orbital B + difference in electronegativity between A and B	83.6	84.2	90.0	4.5

A and B” improves the mean score to 83.6% with a standard deviation of 4.5%; meanwhile, the median and highest scores are 84.2% and 90.0%, respectively.

The result of the cross-validation with the seven descriptors “Atomic Number A,” “Atomic Number B,” “Density,” “Atomic Orbital A,” “Atomic Orbital B,” “Difference in Electronegativity between A and B,” and “Atomic Orbital B + Difference in electronegativity between A and B” is shown in Fig. 2. Figure 2(a) shows that the predicted lattice constant matches the true lattice constant where a linear relation is observed. In addition, the percentage of error of the predicted lattice constants in the trained data and test data is also collected into histograms in Figs. 2(b) and 2(c), respectively. Within the trained data, the majority of error is located at -3% . Similarly, the percentage of error of the predicted lattice constant using the test data set is concentrated at $+3\%$ as shown in Fig. 2(c). The error range for predicting the lattice constant using both the trained and test data sets is found to be small when compared to the true lattice constant. Thus, cross-validation with the SVR technique reveals that the following descriptors can be considered to determine the lattice constant of binary body centered cubic AB: “Atomic Number A,” “Atomic Number B,” “Density,” “Atomic Orbital A,” “Atomic Orbital B,” “Difference in Electronegativity between A and B,” and “Atomic Orbital B + Difference in Electronegativity between A and B.”

The lattice constants of experimentally synthesized binary BCC materials are predicted in order to test the accuracy of the trained machine. In particular, the lattice constants of FeAl, FeTi, CoTi, MgAg, and ScAl are predicted as shown in Table II, where the lattice constant of such materials are experimentally measured.^{15–19} Note that the lattice constants of FeAl, FeTi, CoTi, MgAg, and ScAl are removed from the dataset in order to evaluate the accuracy of the trained SVR. The lattice constant of those materials are predicted by trained SVR using seven descriptors: “Atomic Number A,” “Atomic Number B,” “Density,” “Atomic Orbital A,” “Atomic Orbital B,” “Difference in Electronegativity between A and B,” and “Atomic Orbital B + Difference in Electronegativity between A and B.” The predicted lattice constants and experimental lattice constants are collected in Table II. Table I indicates that the predicted lattice constant using the trained machine has less than 4% error when compared to the experimental and the calculated lattice constant. This error is considered to be as good as the results produced using first principle calculations and shows that machine learning has the potential to accelerate the

search for lattice constants since the predictions obtained from machine learning require much less time than first principle calculations. Thus, the discovered seven descriptors are able to predict the lattice constant of a binary BCC crystal with high accuracy. One can consider that those seven descriptors could also be base descriptors for determining the lattice constant of further complex crystals.

B. Inverse problem

The trained SVR is able to predict the lattice constant of binary BCC crystals with a good accuracy using the previously uncovered seven descriptors. Given this, one can consider that the inverse problem can also be investigated using trained SVR and prediction of material combinations is therefore performed using a desired lattice constant. In particular, the workflow proposed in Fig. 1 is taken for the inverse problem where every possible combination of descriptor variables are given to the trained SVR which returns the corresponding lattice constants, thereby rapidly expanding upon the data available for evaluation. Once the data are expanded into a much larger dataset, the desired lattice constant can then be investigated and the corresponding descriptor variables can then be returned.

Every possible combination of the seven descriptors (a total of 154 100 variations) is given to the trained SVR which then returns the corresponding lattice constant. Note that 1.00 g/cm^3 – 15.00 g/cm^3 (100 cut) is used for density. The generated 154 100 possible combinations with corresponding lattice constants are then stored and treated as a new, expanded dataset. Material prediction is performed by searching for the desired lattice constant within this new dataset and the corresponding material combination and density that produce the desired lattice constant are then returned. In particular, the desired lattice constants $3.00 \text{ \AA}$ and $3.50 \text{ \AA}$ are sought for within the new dataset, and materials that possess such lattice constants are collected in Table III. Table III shows that five materials possessing a lattice constant of $3.00 \text{ \AA}$ are predicted while 24 materials possessing a lattice constant of $3.50 \text{ \AA}$ are predicted.

These results demonstrate that material prediction based on a desired lattice constant can be carried out if the proper descriptors are identified. However, it must be noted that this approach has limitations. For cases involving a complex material property or structure, the number of possible descriptor variables would most likely dramatically increase. Investigating the inverse problem for these cases may require a longer period of time to process or may max out the computer resource

FIG. 2. (a) True and predicted lattice constants in SVR where the following descriptors are used: “Atomic Number A,” “Atomic Number B,” “Density,” “Atomic Orbital A,” “Atomic Orbital B,” “Difference in Electronegativity between A and B,” and “Atomic Orbital B + Difference in Electronegativity between A and B.” Error (%) histograms of (b) trained datasets and (c) test datasets.

limit. Thus, care must be given when attempting to address the inverse problem in this manner.

C. Meaning behind the descriptors

The roles played by the following seven descriptors are investigated: “Atomic Number A,” “Atomic Number B,”

TABLE II. Predicted, calculated, and experimental lattice constants in Å and corresponding error are indicated.

Material	Predicted	Calculated	Error %	Experiment	Error %
FeAl	2.92	2.90	0.7	2.91 ¹⁵	0.3
FeTi	3.08	2.96	3.9	2.98 ¹⁶	3.3
CoTi	3.07	2.99	2.6	3.00 ¹⁷	2.3
MgAg	3.41	3.36	1.5	3.31 ¹⁸	2.9
ScAl	3.47	3.37	2.9	3.39 ¹⁹	2.3

“Density,” “Atomic Orbital A,” “Atomic Orbital B,” “Difference in Electronegativity between A and B,” and “Atomic Orbital B + Difference in Electronegativity between A and B.” The dataset constructed by using first principle calculations is used for visualizing the descriptors in order to reveal the relation between the descriptors. Table II shows that “Atomic Number A” and “Atomic Number B” play a key role in determining the lattice constant as the mean score of cross-validation is 80.1%. This leads to a further investigation into the relation between the lattice constant and the atomic elements. Figure 3 is a visualization of “Atomic Number A” and “Atomic Number B” plotted with colorization according to the corresponding lattice constant. Patterns in the lattice constants appear in Fig. 3 where short and large lattice constants alternate as “Atomic Number B” increases. In particular, the

TABLE III. Prediction of material combinations and corresponding densities (g/cm^3) based on a desired lattice constant (Å).

Desired lattice constant	Material	Density
3.00	AlMn	5.10
3.00	CoN	4.82
3.00	CoV	7.36
3.00	CoFe	10.89
3.00	CoZn	9.48
3.50	AgNb	3.12
3.50	AlGa	1.98
3.50	AsCa	5.66
3.50	AsSe	6.23
3.50	AuP	4.11
3.50	AuZr	9.06
3.50	FeSb	5.52
3.50	GaLi	1.28
3.50	GaIn	7.22
3.50	GaBi	11.46
3.50	NaTe	6.65
3.50	NaAu	5.10
3.50	NiTe	6.37
3.50	OsSb	5.38
3.50	OsTe	3.54
3.50	PdS	4.95
3.50	PdMo	7.64
3.50	PdSn	4.53
3.50	PtAs	8.21
3.50	PtZr	6.23
3.50	PtRh	9.06
3.50	RhMo	2.13
3.50	TiNb	3.26
3.50	WSi	4.82

FIG. 3. A graph representing “Atomic Number A” plotted against “Atomic Number B.” The color bar indicates the corresponding lattice constant (Å) of the AB material.

lattice constant increases as “Atomic Number B” increases until it reaches a noble gas, where it then decreases and begins to increase again until it reaches another noble gas. These results show that the atomic element has a large impact on the development of the lattice constant of the material.

The contribution of density, atomic orbital, and electronegativity are also investigated as such descriptors increase the score as shown in Table II. “Density,” “Atomic Orbital B,” and “Atomic Orbital B + Difference in Electronegativity between A and B” are visualized as shown in Fig. 4. Figure 4 indicates that “Atomic Orbital B” is classified by “Density” and “Atomic Orbital B + Difference in Electronegativity between A and B.” Figure 4 also shows that “Density” increases as “Atomic Orbital B” changes in the order of s(0), p(1), and d(2) while the f(3) orbital element is located between the p and d orbitals. In addition, the mathematically tailored descriptor “Atomic Orbital B + Difference in Electronegativity between A and B” has a large impact on “Atomic Orbital B” where “Atomic Orbital B + Difference in Electronegativity between A and B” increases as “Atomic Orbital B” changes from s(0) to f(3). “Density,” “Atomic Orbital B,” and “Atomic Orbital B + Difference in Electronegativity between A and B”

FIG. 4. “Density” (g/cm^3) and “Atomic Orbital B + Difference in Electronegativity between A and B.” The color bar indicates the corresponding atomic orbital B where s, p, d, and f orbitals are represented as 0, 1, 2, and 3, respectively.

contribute towards the classification of the dataset as shown in Fig. 4, thereby improving the accuracy of the prediction of the lattice constant as shown in Table I.

These results demonstrate that trends can be unveiled depending on how descriptors are visualized and treated within multi-dimensional space. One can consider that the prediction of material properties can be made achievable in principle if the appropriate descriptors are determined. Additionally, this set of descriptors can act as the base descriptors for more complex structures, which can help reduce the time spent on searching for descriptors.

V. CONCLUSION

The lattice constant of binary BCC materials is explored using first principle calculations and machine learning. In particular, 1541 binary BCC crystals are constructed and calculated using density functional theory within first principle calculations. The calculated results, corresponding information from the periodic table, and mathematically tailored data are stored as a dataset. SVR within machine learning is implemented for training the data set where seven descriptors are discovered for predicting the lattice constant of binary BCC materials. Cross-validation of the discovered seven descriptors and SVR shows accuracy with a mean score of 83.6%, and the predicted lattice constants have a less than 4% margin of error compared to the lattice constants in experiment and computation. In addition, the inverse problem is performed using trained SVR where prediction of material combinations using a desired lattice constant is achieved and the contribution of the seven descriptors are also discussed. Thus, SVR using the seven discovered descriptors can contribute towards a more rapid search for lattice constants while the descriptors themselves can act as base descriptors for lattice constants in more complex crystals.

ACKNOWLEDGMENTS

Part of this work is funded by “Materials research by Information Integration” Initiative (MI²I) project of the Support Program for Starting Up Innovation Hub from Japan Science and Technology Agency (JST). In addition, funding from EPSRC Grant No. EP/K016288/1 is acknowledged. Computational work is supported in part by Hokkaido University academic cloud, Information Initiative Center, Hokkaido University, Sapporo, Japan.

¹G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, *Chem. Mater.* **22**, 3762 (2010).

²G. Hautier, C. Fischer, V. Ehlacher, A. Jain, and G. Ceder, *Inorg. Chem.* **50**, 656 (2010).

³O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, and S. Curtarolo, *Chem. Mater.* **27**, 735 (2015).

⁴A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder *et al.*, *APL Mater.* **1**, 011002 (2013).

⁵A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka, *Phys. Rev. B* **89**, 054303 (2014).

⁶K. Takahashi and Y. Tanaka, *Dalton Trans.* **45**, 10497 (2016).

⁷K. Takahashi and Y. Tanaka, *Phys. Rev. B* **95**, 014101 (2017).

⁸L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, *Phys. Rev. Lett.* **114**, 105503 (2015).

⁹A. Jain, Y. Shin, and K. A. Persson, *Nat. Rev. Mater.* **1**, 15004 (2016).

¹⁰K. Takahashi and Y. Tanaka, *Comput. Mater. Sci.* **112**, 364 (2016).

- ¹¹J. J. Mortensen, L. B. Hansen, and K. W. Jacobsen, *Phys. Rev. B* **71**, 035109 (2005).
- ¹²J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).
- ¹³H. J. Monkhorst and J. D. Pack, *Phys. Rev. B* **13**, 5188 (1976).
- ¹⁴F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, *J. Mach. Learn. Res.* **12**, 2825 (2011).
- ¹⁵S. A. Makhlof, T. Nakamura, and M. Shiga, *J. Magn. Magn. Mater.* **135**, 257 (1994).
- ¹⁶P. Thompson, J. Reilly, and J. Hastings, *J. Appl. Crystallogr.* **22**, 256 (1989).
- ¹⁷T. Takasugi and O. Izumi, *Phys. Status Solidi A* **102**, 697 (1987).
- ¹⁸S. Sharma and A. Weiss, *J. Less-Common Met.* **104**, L5 (1984).
- ¹⁹J. Schuster and J. Bauer, *J. Less-Common Met.* **109**, 345 (1985).