Unveiling descriptors for predicting the bulk modulus of amorphous carbon

Keisuke Takahashi and Yuzuru Tanaka

Phys. Rev. B **95**, 054110 — Published 14 February 2017

DOI: 10.1103/PhysRevB.95.054110

Unveiling descriptors for predicting the bulk modulus of amorphous carbon

Keisuke Takahashi*
Center for Materials research by Information Integration (CMI^2),
National Institute for Materials Science (NIMS),
1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan and
Graduate School of Engineering, Hokkaido University, N-13, W-8, Sapporo 060-8628, Japan

Yuzuru Tanaka
Meme Media Laboratory, Hokkaido University, N-13, W-8, Sapporo 060-8628, Japan

(Dated: January 27, 2017)

Descriptors for bulk modulus of amorphous carbon are investigated through the implementation of data mining where data sets are prepared using first principle calculations. Data mining reveals that the number of bonds in each C atom and the density of amorphous carbon are found to be descriptors representing the bulk modulus. Support vector regression (SVR) within machine learning is implemented and descriptors are trained where trained SVR is able to predict the bulk modulus of amorphous carbon. An inverse problem, starting from bulk modulus towards structural information of amorphous carbon, is performed and structural information of amorphous carbon is successfully predicted from the desired bulk modulus. Thus, treating several physics factors in multidimensional space allows for the prediction of physical phenomena. In addition, the reported approach proposes that 'big data' can be generated from a small data set using maching learning if descriptors are well defined. This would greatly change how amorphous carbon would be treated and help accelerate further development of amorphous carbon materials.

I. INTRODUCTION

In materials science, the term amorphous is a term generally used when long range characteristics are not identified in a crystal. In general, the properties of amorphous solids strongly rest on how the atoms are placed and form bonds with surrounding atoms [1]. However, the structures of amorphous solids are strongly coupled with experimental techniques and conditions which result in various local structures. While several structural models have been proposed, amorphous carbon and amorphous silicon are commonly recognized amorphous solids where carbon and silicon atoms are randomly placed [2, 3]. Such amorphous materials are synthesized by implementing deposition techniques or mechanical alloying[4, 5]. Cases have been reported where amorphous carbon is produced when a defect is introduced in graphene or graphite [6, 7]. It has also been noted that an amorphous state is reported in metal when the metal is mechanically alloyed.[8, 9] Because the atoms are randomly located in amorphous solid, the range of possible applications of amorphous solids is vast, including applications in catalysts, optics, electronics, and structural materials.

Amorphous carbon is chosen as a case model where properties of amorphous carbon are strongly coupled with how the carbon atoms are placed [10]. The atomic configuration of amorphous carbon is random on an atomistic scale; however, one can consider that the properties of amorphous carbon can be determined by certain factors which can often be referred to as its descriptors. Descriptors are understood to be the core factors that ultimately decide the properties of a material. In other words, the materials properties could be reliant on several physical factors when such physical factors are treated within multi dimensional space [11]. If such descriptors of amorphous carbon can be determined, then in principle it would be possible to predict the properties of amorphous carbon.

With the rapid growth of first principle calculations and development of supercomputers, construction of a material dataset becomes achievable within a short period of time. With the aid of data science, trends and descriptors in a dataset can then be acquired. If descriptors of a material genome are determined, machine learning can be an effective algorithm used to predict a materials properties [12–17]. Descriptors of amorphous carbon are therefore explored by applying data mining techniques to an amorphous carbon dataset where dataset of the structures and corresponding properties of amorphous carbon are prepared using the first principle calculations. Once descriptors are determined, machine learning is then implemented to train the dataset in order to predict the properties of amorphous carbon from the discovered descriptors. In addition, inverse problem from properties to structure information of amorphous carbon is proposed by using trained machine. Thus, materials properties of amorphous carbon is investigated in term of data science and materials physics.

II. WORKFLOW

The workflow for predicting the structural information from desired material properties is proposed as shown in Figure 1. A set of randomly generated amorphous carbon

* keisuke.takahashi@eng.hokudai.ac.jp

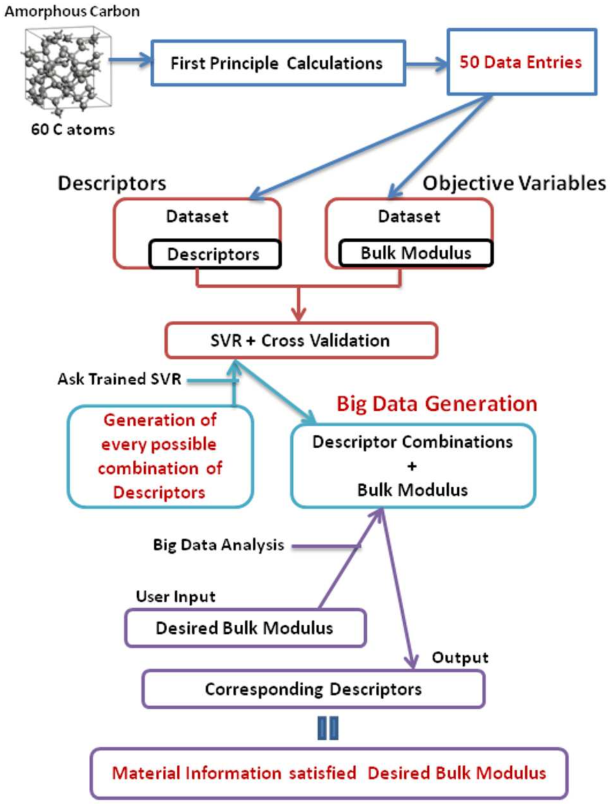

FIG. 1: Proposed workflow model for predicting the structure of amorphous carbon from properties using first principle calculation and data science.

is prepared and calculated using first principle calculations. Material properties including structural information are extracted from the calculated results and stored as a dataset. Note that the dataset in this step is a rather small dataset with 50 entries. The dataset is then classified into two groups: descriptors and objective variables. Descriptors are combinations of variables responsible for determining the objective variables. Support vector regression (SVR) within machine learning is implemented in order to train the descriptors and corresponding objective variables in the dataset. Various descriptors are trained and cross–validation is applied where descriptors with high scores are sought after. Descriptors with a high score from cross validation and the corresponding objective variables in the dataset are then trained using SVR. Once SVR is trained, all possible combinations of descriptors are generated where the number of combinations can vary between a few thousand to a few million as the number of possible combinations grows exponentially when the number of involved descriptors increase. All of the generated combinations of descriptors are then given to the trained SVR which returns corresponding objective variables; it is this step where big data is generated. Desired objective variables are searched for within the generated 'big data' where a list of corresponding descriptors which satisfy the desired objective variable is returned. The corresponding descriptors are composed of the material information that satisfies the objective variables, making it possible to solve the inverse problem of deriving structural information from material properties. Thus, if descriptors can be extracted from a small dataset, machine learning can be used to essentially generate 'Big Data' from a small dataset. The time generally taken when using first principles calculations to generate big data is, therefore, dramatically reduced with the aid of data science.

III. COMPUTATIONAL METHOD AND DATASET PREPARATION

Grid based projector augmented wave (GPAW) method is implemented for first principle calculations [18]. Exchange correlation of Perdew–Burke–Ernzerhof (PBE) is applied with 4x4x4 special k points of the Brillouin zone sampling [19, 20]. 50 amorphous carbon structures consisting of 60 C atoms are constructed in a cubic unit cell where each of the 60 C atoms are randomly placed in a cubic unit cell in order to obtain various structures. A cubic cell of 7.5 Å x 7.5 Å x 7.5 Å is designed as a base unit cell and 60 C atoms are randomly generated into a constructed cubic cell. Each 50 constructed amorphous carbon is relaxed and lattice optimization is performed by shrinking and expanding the lattice where the lowest energy lattice constant is taken. Once the lowest energy lattice is determined, another relaxation is performed to optimize the atomic configuration of C atoms.

The following properties are extracted from each optimized amorphous carbon: lattice constant, density, bulk modulus, total energy, and number of bonds in each C atom. Note that density is calculated on the basis of the lowest energy lattice constant with 60 C atoms and atomic mass of carbon. Bulk modulus is calculated by 5 % of shrinking and expanding the cubic lattice as shown in Figure 1. The number of bonds in each C atom is counted by scanning every C atom where a bond length with neighboring atoms is defined if it is smaller than 1.75 Å as carbon allotrope has a large C–C distance in comparison to graphite and diamond. Collected properties are listed in Supporting Information [21].

IV. RESULT AND DISCUSSION

A. Descriptors Search

Descriptors for determining the bulk modulus are explored in terms of data mining. Prediction of the bulk modulus of amorphous carbon is performed by implementing machine learning and constructed dataset. In particular, suitable descriptors are explored by combining maching learning and cross–validation algorithms where the support vector regression (SVR) algorithm within scikit–learn is implemented for the machine learning process [22].

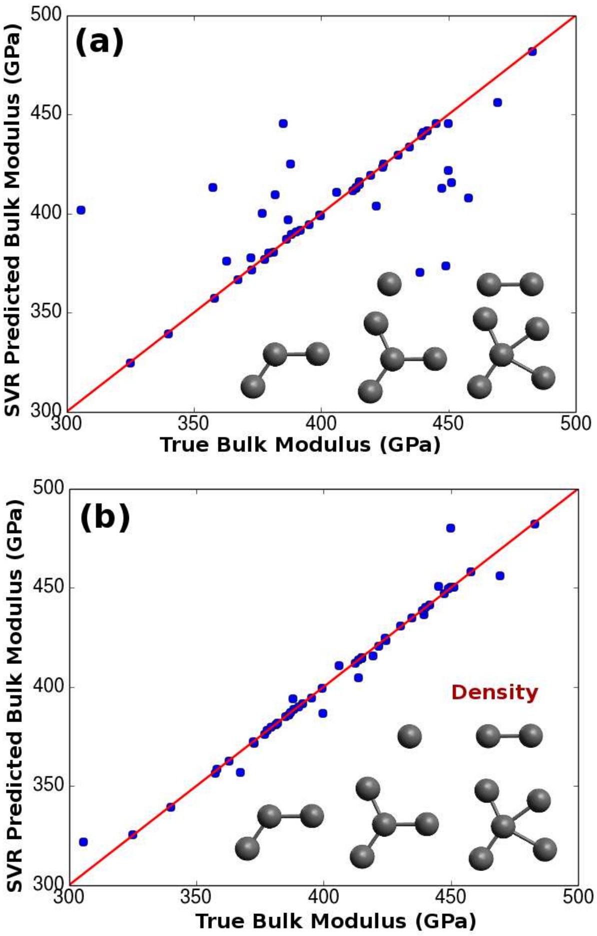

FIG. 2: Predicted bulk modulus against true bulk modulus with descriptors: (a) the number of bonds in each C atom, and (b) the number of bonds in each C atom with density. Structure models of bond type in amorphous carbon is also shown.

The number of bonds in each C atom is chosen as a descriptor for predicting the bulk modulus of amorphous carbon. In particular, five types of bonds are defined as the following: 0 to 4 bonds in each C atom where the cut off bond distance is set to 1.75 Å. 50 amorphous carbon samples composed of 60 C atoms are trained with five descriptors using support vector regression with corresponding bulk modulus. Prediction of bulk modulus is performed by using a trained data set where the results are shown in Figure 2 (a). Figure 2 (a) shows the mismatch between the predicted and true bulk modulus. One can consider that another descriptor could be contributing to the bulk modulus. Therefore, another descriptor is explored where density is found to be another key descriptor for bulk modulus. By adding density as a descriptor,the mismatch between predicted and true bulk modulus is greatly improved as seen in Figure 2 (b). The number of bonds for each C atom and density within amorphous carbon are therefore found to be descriptors for the prediction of the bulk modulus of amorphous carbon.

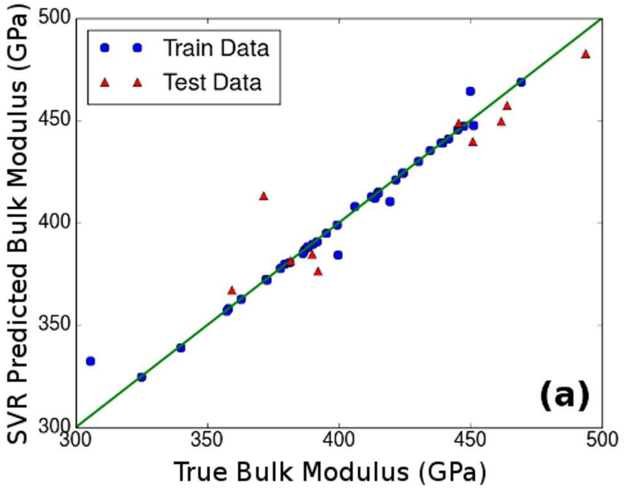

FIG. 3: Cross-validation of bulk modulus with randomly sorted data set of 80% of trained data and 20 % of test data where followign descriptors are used: the number of bonds in each C atom with density.

Cross–validation is performed to confirm further accuracy of trained support vector regression for the bulk modulus. 50 samples in a data set are randomly sorted where 80% is set to trained data and 20 % is set to test data. Cross–validation results are shown in Figure 3 where average score of ten random test and train data set is 83% accuracy, median score is 83%, standard deviation is 3% and the highest score of 89% is achieved for the prediction of the bulk modulus. Thus, the revealed 6 descriptors can be considered as global descriptors for predicting the bulk modulus of amorphous carbon.

B. Physical Meaning

The physics behind the chosen descriptors rests on several factors. One factor is the electronic structure of the number of bonds for each C atom. In general, sp and sp2 states of carbon would be stable in a two dimensional form as seen in graphite or graphene. However, one can consider that the sp and sp2 states in three dimensional space of amorphous carbon would be in a metastable state. Figure 4 shows the projector density of state (PDOS) of the sp2 state of C atom in amorphous carbon and in graphene. One can see that there are large peaks of s-electrons in the anti-bonding state of the C atom in sp2 of amorphous carbon as shown in Figure 4 (a). Meanwhile, there are less peaks in the anti-bonding state of the C atom in sp2 of graphene as shown Figure 4 (b). Thus, the sp2 state in amorphous carbon can be unstable compared to the sp2 state in graphene. This metastable sp2 state can be considered to be a key factor for determining the bulk modulus and density of amorphous carbon as such metastable sp2 states can be reactive and could be responsible for how C-C bonds would form in amorphous carbon. The relation between 'density and bulk modulus' or 'the number of bonds and bulk modulus can be imaginable, however, as high-dimensionally combining several physics factors (such as density and the number of bonds) allows for the prediction of a more precise bulk modulus. Thus, with the aid of data science, several physics factors can interact within a high dimensional space, resulting in the prediction of physical properties.

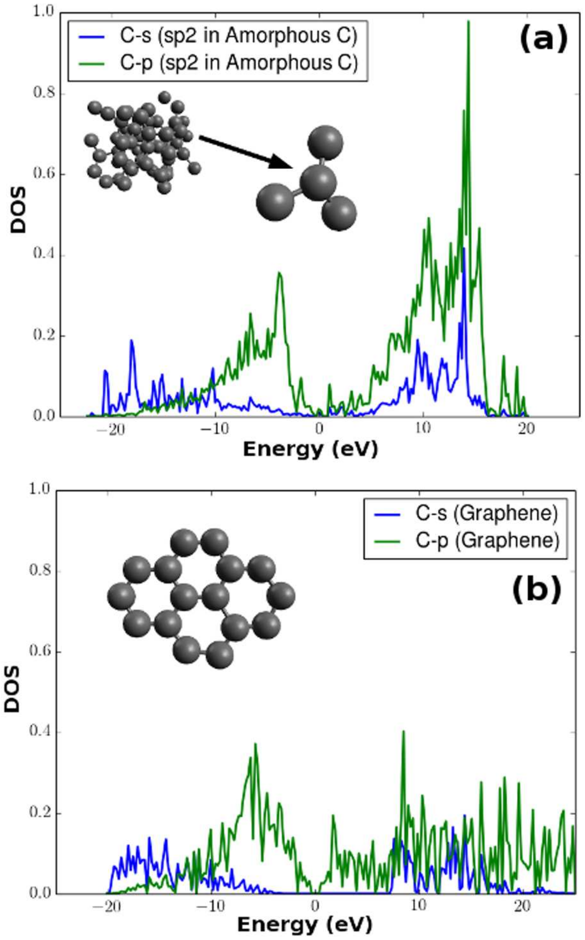

FIG. 4: Projector densty of state of C atom in (a) sp2 in amorphous carbon and (b) sp2 in graphene. Structures of sp2 in amorphous carbon and graphene is also shown.

V. INVERSE PROBLEM

The inverse problem is performed in order to predict the number of bonds and density of amorphous carbon from a given bulk modulus. The inverse problem is solved on the basis of the workflow shown in Figure 1 where every possible combination of descriptors is given to the trained SVR which then returns the corresponding bulk modulus, essentially generating 'Big data'. Note that generated 'Big Data' refers to a large number of descriptors variables with corresponding bulk modulus. Once big data is generated, a desired bulk modulus is searched for within the big data and corresponding combinations of descriptors are uncovered. Here, six descriptors– the number of bonds in each C atom and density– are determined for predicting the bulk modulus of amorphous carbon. However, it is challenging to extract six descriptors from a single variable, which in this case is the bulk modulus. Based on Figure 1, every possible combination of variables within the six descriptors are generated and given to the trained SVR where the corresponding bulk modulus is returned. In particular, the following variables are generated: 0 bonds,0-11; 1 bond,5-18; 2 bonds,10-25; 3 bonds,13-26; 4 bonds,8-21; and density,2.65-2.85(20 cut) where the total number of atoms is set to be 60 atoms in the number of bonds cases. A total of 7,250,100 possible combinations are generated with corresponding bulk modulus from the trained support vector regression. The generated 7,250,100 datasets containing the number of bonds, density, and corresponding bulk modulus are treated and analyzed as 'big data'.

The number of bonds in each C atom and density that satisfied the bulk modulus of 350 GPa and 400 GPa are explored within the generated 'big data'. Note that the number of C atoms which have no bonds is set to 0 in order to narrow the results. The predicted structures (the number of bond and density) of amorphous carbon corresponding to desired bulk modulus of 350 GPa and 400 GPa are collected in Table I. One can see that density is predicted to be 2.82 g/cm^3 and 2.83 g/cm^3 for the bulk modulus of 350 GPa; however, there are 4 different bond states that could achieve 350GPa. This implies that different bond states with the same density exist that can achieve the specified bulk modulus. Similarly, the desired bulk modulus is set to 400GPa and the corresponding structural information of amorphous carbon is extracted from the big data. Compared to the bulk modulus of 350 GPa, 32 possible cases of amorphous carbon satisfying the bulk modulus of 400 GPa are discovered as seen in Table I. Thus, the trained SVR is able to create amorphous carbon 'big data' by asking every possible combination of descriptors and desired bulk modulus can be sought for in 'big data' which would return structural information of amorphous carbon.

Table I can potentially connect with experimental data. In particular, the ratio of sp^2 and sp^3 states of amorphous carbon can be estimated by using electron energy loss spectroscopy and Raman spectroscopy[2, 6]. In that sense, one can consider that the bulk modulus of amorphous carbon can be determined by searching the constructed 'big data' from the ratio of sp^2 and sp^3 states and density of amorphous carbon which can be

TABLE I: Predicted structural information (the number of bonds in each C atom and density in g/cm^3 of amorphous carbon with corresponding desired bulk modulus(DBM) in GPa.

DBM	Zero	One	Two	Three	Four	Density
350	0	7	22	13	18	2.83
350	0	12	11	20	17	2.83
350	0	14	10	17	19	2.82
350	0	14	10	20	16	2.83
400	0	5	13	23	19	2.83
400	0	5	17	25	13	2.79
400	0	5	19	17	19	2.77
400	0	7	11	24	18	2.83
400	0	7	17	22	14	2.80
400	0	7	18	15	20	2.78
400	0	7	18	21	14	2.80
400	0	8	16	21	15	2.80
400	0	8	24	16	12	2.81
400	0	9	16	21	14	2.80
400	0	9	22	19	10	2.81
400	0	9	23	18	10	2.81
400	0	10	19	22	9	2.81
400	0	11	21	14	14	2.83
400	0	11	23	15	11	2.82
400	0	12	10	21	17	2.74
400	0	12	14	23	11	2.81
400	0	12	15	22	11	2.81
400	0	12	17	22	9	2.82
400	0	12	18	22	8	2.82
400	0	12	19	13	16	2.84
400	0	13	11	18	18	2.74
400	0	13	12	15	20	2.77
400	0	13	14	20	13	2.80
400	0	13	14	24	9	2.84
400	0	13	16	22	9	2.83
400	0	13	17	14	16	2.84
400	0	13	19	18	10	2.83
400	0	13	20	16	11	2.84
400	0	14	10	23	13	2.78
400	0	14	18	17	11	2.85
400	0	15	21	16	8	2.85

acquired in experiment. This would potentially lead to link the processing and material properties of amorphous carbon. For instance, if experimental conditions such as temperature and pressure can be descriptors for predicting the ratio of sp^2 and sp^3 and density of amorphous carbon, machine learning can essentially create 'big data' of experimental conditions with corresponding ratios of sp^2 and sp^3 states and density using the same proposed approach as Figure 1. If Table *I* and 'big data' of experimental conditions with the corresponding ratios of sp^2 and sp^3 states and density are created, prediction of experimental conditions for synthesizing amorphous carbon upon the request of a desired bulk modulus can thus achievable in principle.

The advantage of implementing data science is the ability to generate 'big data' from descriptors found in a small dataset. This approach greatly reduces required computational time as the construction of such 'big data' using first principles calculations would result in a long period of computational time while a trained machine can generate 'big data' within a short period of time. In this sense, one can consider that if the experimental conditions and corresponding structural information (the number of bonds in each C atom and density) are linked, it would then be possible to predict experimental conditions based upon the desired bulk modulus. Thus, structural information of amorphous carbon can be directly predicted from desired properties if descriptors are well defined. Chosen descriptors for amorphous carbon could also be base descriptors for determining the structures of two and three dimensional allotropes of carbon [23, 24]. In general, structures of amorphous carbon are complex as various local structures can be considered. The discovered descriptors, though, can be the basis for predicting the bulk modulus of amorphous carbon as the SVR trained using 50 unique structures of amorphous carbon. One can therefore consider that increasing the dataset by adding further unique amorphous carbon structures would, in principle, allow for covering further key structure features of amorphous carbon.

VI. CONCLUSION

In conclusion, Descriptors for bulk modulus of amorphous carbon are investigated using machine learning where a data set is prepared by implementing first principle calculations. Although amorphous carbon is a noncrystalline material, the number of bonds in each C atom and the density of amorphous carbon is found to be descriptors for determining the bulk modulus of amorphous carbon. Support vector regression within machine learning is implemented and chosen descriptors are trained where cross–validation confirms a high accuracy for predicting the bulk modulus of amorphous carbon. Trained support vector regression is used to solve the inverse problem where prediction of structural information of amorphous carbon from desired bulk modulus is successful. Thus, properties of amorphous carbon can be predicted through machine learning if key descriptors are discovered, which would have a great impact on how amorphous carbon is treated. In addition, the approach proposes that treating several physics factors in multidimensional space allows for the prediction of physical phenomena and the inverse problem from physical properties to material information.

VII. ACKNOWLEDGEMENT

Part of this work is funded by Materials research by Information Integration Initiative (MI^{2}I) project of the Support Program for Starting Up Innovation Hub from Japan Science and Technology Agency (JST). Computational work is supported in part by Hokkaido university academic cloud,information initiative center, Hokkaido University, Sapporo, Japan.

[1] G. Galli, R. M. Martin, R. Car, and M. Parrinello, Phys. Rev. Lett. **62**, 555 (1989).

[2] A. C. Ferrari and J. Robertson, Phys. Rev. B **61**, 14095 (2000).

[3] D. E. Carlson and C. R. Wronski, Appl. Phys. Lett. **28**, 671 (1976).

[4] R. Chittick, J. Alexander, and H. Sterling, J. Electrochem. Soc. **116**, 77 (1969).

[5] C. Davis, G. Amaratunga, and K. Knowles, Phys. Rev. Lett. **80**, 3280 (1998).

[6] J. Huang, Acta materialia **47**, 1801 (1999).

[7] J. Kotakoski, A. Krasheninnikov, U. Kaiser, and J. Meyer, Physical Review Letters **106**, 105505 (2011).

[8] KLEMENT W. , WILLENS R. H., and DUWEZ POL, Nature **187**, 869 (1960), 10.1038/187869b0.

[9] J.-J. Kim, Y. Choi, S. Suresh, and A. Argon, Science **295**, 654 (2002).

[10] J. Robertson, Mater. Sci. Eng. R-Rep. **37**, 129 (2002).

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

[12] K. Rajan, Mater. Today **8**, 38 (2005).

[13] A. 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).

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

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

[16] K. Takahashi and Y. Tanaka, Dalton Trans. **45**, 10497 (2016).

[17] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, Phys. Rev. Lett. **117**, 135502 (2016).

[18] J. J. Mortensen, L. B. Hansen, and K. W. Jacobsen, Phys. Rev. B **71**, 035109 (2005).

[19] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[20] H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).

[21] See Supplemental Material at URL will be inserted by publisher for dataset used in this article..

[22] F. 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).

[23] S. Zhang, Q. Wang, X. Chen, and P. Jena, Proc. Nat. Acad. Sci **110**, 18809 (2013).

[24] S. Zhang, J. Zhou, Q. Wang, X. Chen, Y. Kawazoe, and P. Jena, Proceedings of the National Academy of Sciences **112**, 2372 (2015).

[25] A. Ferrari, J. Robertson, M. Beghi, C. Bottani, R. Ferulano, and R. Pastorelli, Appl. Phys. Lett. **75**, 1893 (1999).