Deep learning model for finding new superconductors

Tomohiko Konno,[1,*] Hodaka Kurokawa,[2,†] Fuyuki Nabeshima,[2] Yuki Sakishita,[2] Ryo Ogawa,[2] Iwao Hosako,[1] and Atsutaka Maeda[2]

[1]*National Institute of Information and Communications Technology, 4-2-1 Nukui-Kitamachi, Koganei, Tokyo 184-8795, Japan*

[2]*The University of Tokyo, 7 Chome-3-1 Hongo, Bunkyo City, Tokyo 113-8654, Japan*

(Received 2 December 2019; accepted 3 December 2020; published 12 January 2021)

Exploration of new superconductors still relies on the experience and intuition of experts, and is largely a process of experimental trial and error. In one study, only 3% of the candidate materials showed superconductivity [Hosono *et al.*, Sci. Technol. Adv. Mater. 16, (2015)]. Here, we report a deep learning model for finding new superconductors. We introduced the method named "reading periodic table" that represented the periodic table in a way that allows deep learning to learn to read the periodic table and to learn the law of elements for the purpose of discovering novel superconductors which are outside the training data. It is recognized that it is difficult for deep learning to predict something outside the training data. Although we used only the chemical composition of materials as information, we obtained an R^2 value of 0.92 for predicting T_c for materials in a database of superconductors. We also introduced the method named "garbage-in" to create synthetic data of nonsuperconductors that do not exist. Nonsuperconductors are not reported, but the data must be required for deep learning to distinguish between superconductors and nonsuperconductors. We obtained three remarkable results. The deep learning can predict superconductivity for a material with a precision of 62%, which shows the usefulness of the model; it found the recently discovered superconductor $CaBi_2$ and another one $Hf_{0.5}Nb_{0.2}V_2Zr_{0.3}$, neither of which is in the superconductor database; and it found Fe-based high-temperature superconductors (discovered in 2008) from the training data before 2008. These results open the way for the discovery of new high-temperature superconductor families. The candidate materials list, data, and method are openly available on the Internet.

DOI: 10.1103/PhysRevB.103.014509

I. INTRODUCTION

Extensive research has been conducted on superconductors with a high superconducting transition temperature T_c because of their many promising applications, such as low-loss power cables, powerful electromagnets, and fast digital circuits. However, finding new superconductors is very difficult. In one study, it was reported [1] that only 3% of candidate materials showed superconductivity. Theoretical approaches have been proposed for predicting new superconducting materials. According to Bardeen-Cooper-Schrieffer (BCS) theory [2], which explains phononmediated superconductivity in many materials, high T_c is expected for compounds made of light elements. T_c values of over 200 K have been reported for sulfur hydride [3] and lanthanum hydride [4]. However, very high pressures (over 150 GPa) are required. Superconductivity with a rather high T_c has been observed for cuprates [5] and iron-based materials [6] at ambient pressure, where unconventional superconductivity beyond the BCS framework is realized. However, the strong electron correlations in these materials make it very difficult to conduct first-principles calculations [7–10] to calculate their electronic structures and predict their T_c values. Therefore, new approaches for finding superconductors are needed. Materials informatics, which applies the principles of informatics to materials science, has attracted much interest [11–14]. Among machine learning methods, deep learning has achieved great progress. Deep learning has been used to classify images [15], generate images [16], play Go [17], translate languages [18], perform natural language tasks [19], and make its own network architecture [20,21]. To predict the properties of materials using the conventional methods in materials informatics, researchers must design the input features of the materials; this is called feature engineering. It is very difficult for a human to design the appropriate features. A deep learning method can design and optimize features, giving it higher representation capabilities and potential compared to those of conventional methods. Many studies have been reported on drug discovery and organic chemistry by deep learning (mainly by graph neural networks [22,23]), and on molecules [24,25]. Our results show the possibility of application of deep learning to inorganic materials and condensed matter physics, as additional areas outside organic chemistry.

II. READING PERIODIC TABLE

Here, we report a deep learning model for the exploration of new superconductors. Using deep learning to discover new

*Corresponding author: tomohiko@nict.go.jp
†Second Corresponding author: scottie0018@gmail.com

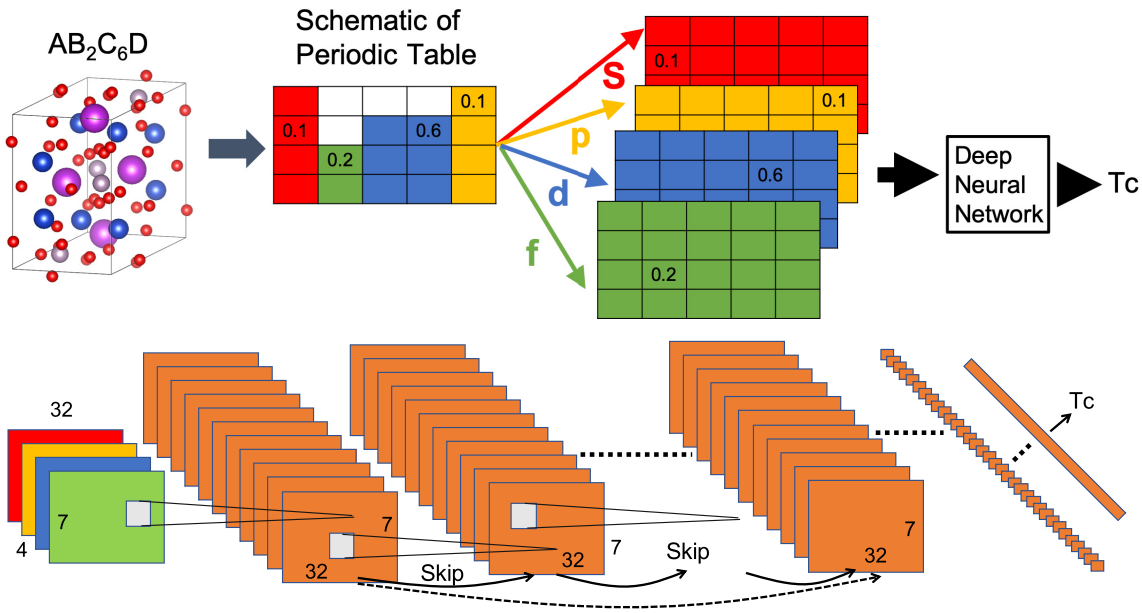

FIG. 1. Proposed method named reading the period table. (Top) The representation of a material by the method. The composition ratios of the material [26] are entered into the two-dimensional periodic table. We then divide the original table into four tables corresponding to s, p, d, and f blocks, which show the orbital characteristics of the valence electrons, to allow the deep learning model to learn the valence orbital blocks. The dimensions of the representation are $4 \times 32 \times 7$. The neural network learns the rules from the periodic table by convolutional layers. (Bottom) The representation by the method and neural network.

superconductor families from known ones is analogous to using deep learning to recognize dogs from training data containing only cats. This form of learning, called zero-shot learning, is very difficult. However, that the properties of elements can be learned by deep learning is shown by us, and they can be applied to materials. Our strategy is to suitably represent these properties, use this representation as training data, and have the deep learning model learn these properties. We made the deep learning model learn how to read the periodic table as human experts do. Although humans cannot recall tens of thousands of data points, computers can. For this purpose, we represented the periodic table in a way that allows a deep learning model to learn it, as illustrated in Fig. 1. The convolutional layers learn the relative positions of the elements on the table, because they use the same local weights to whole periodic table. This is the reason why we use convolutional layers. Full connection layers should be basically avoided, since over-fitting easily occurs, and they do not learn the relative relationship. This method, named reading periodic table, is our first contribution to deep learning. We considered inorganic crystal superconductors because the number of known organic superconductors is small. We used only the composition of materials because the applied superconductor database does not have sufficient spatial information. (See more detail in Supplemental Material [27].)

We used the deep learning model to predict the critical temperatures T_c of superconductors in the SuperCon data set [28], which has the T_c values of about 13 000 superconductors. We refer to the model trained with only SuperCon as the preliminary model. The train-test split was 0.05. A scatter plot of the predicted and actual T_c values is shown in Fig. 2. The R^2 value is 0.92, which is higher than that previously reported (0.88) for a random forest regression model [29], where materials were restricted to those with $T_c > 10$ K (half of all materials). In contrast, our preliminary model does not have any restrictions regarding T_c (see Supplemental Material [27]). The random forest requires many appropriate input features of the materials (e.g., atomic mass, band gap, atomic configuration, melting temperature) to be manually designed. Here, even without such feature engineering, we achieved much better results.

III. THE PROBLEM IN USING DATA OF SUPERCONDUCTORS ONLY AND THE METHOD NAMED GARBAGE-IN FOR OVERCOMING IT

We used the preliminary model trained with SuperCon to predict the T_c values of 48 000 inorganic materials in the Crystallography Open Database (COD) to find new superconductors for experiments. However, for about 17 000 of the materials, the predicted T_c was >10 K, which is unreasonable. The failure to find new superconductors by this preliminary model seems to originate from the fact that the training data (SuperCon) included only 60 nonsuperconductors; the preliminary model was thus unable to learn nonsuperconductors. Data on nonsuperconductors are needed to differentiate superconductors from nonsuperconductors. However, no such data set is available. Hence, we created synthetic data on nonsuperconductors, supposing that the T_c values of the inorganic materials in COD that are not in SuperCon are 0 K under the assumption that most of these materials do not become superconductors with finite T_c. We used the synthetic data and SuperCon as the training data. We refer to this data generation method as garbage-in, which is our second contribution to deep learning.

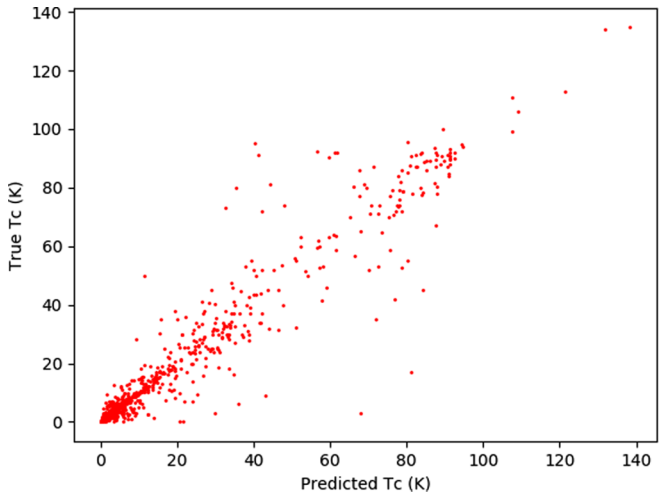

FIG. 2. Scatter plot of predicted and true (SuperCon) T_c values.

As demonstrated by the above results for the preliminary model, scores of tests using only superconductor data, SuperCon, are not good for evaluating models. Usually, density functional theory is applied for evaluation in materials informatics; however, density functional theory cannot be used to evaluate models, because it is very difficult to calculate T_c for strongly correlated systems. A database of nonsuperconductors is thus necessary.

IV. THE PREDICTION OF SUPERCONDUCTIVITY

We applied a list of materials reported by Ref. [1] to evaluate the models. The list has about 400 materials found since 2010; importantly, it includes 330 nonsuperconductors. To temporally separate the materials on the list from the training data, we used only the data added to SuperCon or COD before 2010 as training data. The temporal separation test scheme is better than a random split of training and test data. The training and test data may end up being very similar after a random split. The temporal separation is the same situation when we use deep learning model to find new materials. We investigated outliers in T_c predictions (see Fig. 2) and found that the under- and overestimated materials are cuprates, which have high T_c that are sensitive to small changes in the ratio of elements. The surprise was that our deep learning model was sufficiently capable to find the mistake in the database. Some outliers are due to wrongly recorded T_c values in SuperCon (database of superconductors). Mistakes in data are common. The R^2 is sensitive to such outliers. To compare the capability of a model with expert predictions, we evaluated whether the model could predict superconductivity for the given materials. Hence, we will use precision, recall, and f1 for evaluation.

Randomly selecting a material from the list with $T_c > 0$ K yields a precision of 32%. This is considered the baseline because all the materials on the list were expected to be superconductors before the experiments. For the model predicting T_c with respect to whether it would be higher than 0 K, the results had a precision of 62%, an accuracy of 76%, a recall of 67%, and an f1 score of 63%. This precision is about two times higher than the baseline (32%), which is about 10 sigma above it. The AUC was 0.78. Another interesting threshold is 10 K because only a limited number of superconductors have $T_c > 10$ K. The deep learning method predicted materials as being above this T_c threshold with a precision of 75%, which is about seven times higher than the baseline random precision (10%). The accuracy, recall, and f1 score were 95%, 76%, and 75%, respectively. The AUC was 0.94. In contrast, the preliminary model, trained with SuperCon only, predicted that all the materials would be superconductors, even though the training data were up to the year 2018 (i.e., not temporally separated). A previous study [29] used a random forest method. We also performed random forest binary classification with garbage-in and deep learning binary classification, which classify materials in terms of whether the T_c is beyond 0 K or not. The AUC were 0.78 and 0.96, respectively. The results, summarized in Table I, demonstrate that our deep learning model has good capability to predict superconductivity and clearly outperformed the previous method of random forest. (See Supplemental Material [27]).

TABLE I. Scores for predictions of superconductivity for materials reported by Ref. [1]. Reg and Cls are abbreviations for regression and classification, respectively.

	Accuracy	Precision	Recall	f1
Baseline (0 K)		32%	–	–
Our DL model Reg (0 K)	76%	62%	67%	63%
Our DL model Cls (0 K)	78%	72%	50%	59%
Random Forest Cls (0 K)	73%	71%	27%	39%
Baseline (10 K)		10%	–	–
Our DL model Reg (10 K)	95%	75%	76%	75%
Our DL model Cls (10 K)	95%	76%	77%	77%
Random Forest Cls (10 K)	92%	88%	26%	40%

V. THE DISCOVERY OF TWO SUPERCONDUCTORS $CaBi_2$ AND $Hf_{0.5}Nb_{0.2}V_2Zr_{0.3}$

Next, we used the model to predict the T_c values of the materials in COD. The number of materials predicted to be superconductors was different every time we trained the models from scratch, which is expected with deep learning. We made a search target list for the experiment. After we removed cuprates and Fe-based superconductors (FeSCs) from the list, we obtained 900 materials predicted to be superconductors with $T_c > 0$ K, 280 materials with $T_c > 4$ K, and 70 materials with $T_c > 10$ K, which is more reasonable compared to the results obtained using the preliminary model. These materials are candidates for new superconductors. Although the prediction results on materials reported by Hosono *et al.* show that the model is useful, experiments (currently under way) are required to validate the method. The list included $CaBi_2$, which was recently found to be a superconductor [30] and another superconductor, $Hf_{0.5}Nb_{0.2}V_2Zr_{0.3}$ [31]. The two superconductors are not listed in SuperCon. We had not known these were superconductors beforehand. It can be concluded that the deep learning model found actual superconductors. We have made the list openly available.

Another interesting prediction regards BeB_2. The material MgB_2 is a famous superconductor with $T_c = 40$ K, and the

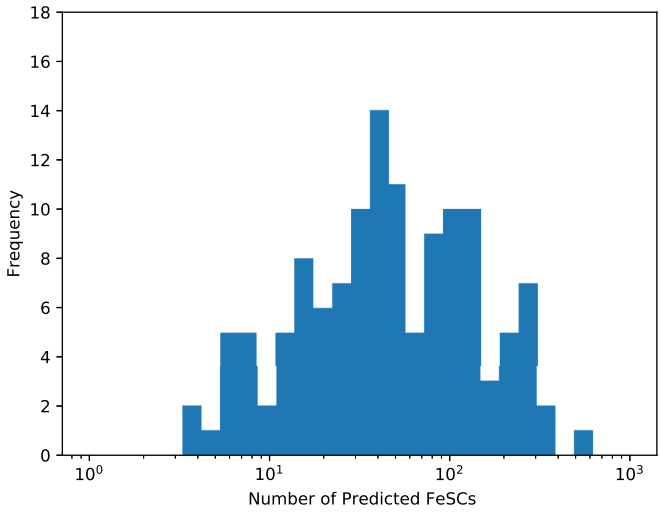

FIG. 3. Histogram of the number of predicted FeSCs with $T_c > 0$ K (log scale).

element Be is just above the element Mg in the periodic table. In the used database, the number of two-element materials that include B is more than 300. Although BeB_2 is not a superconductor, it is not a coincidence that deep learning predicted BeB_2 as such. This is evidence that the deep learning model reads the periodic table to predict superconductors in a similar way as human do.

VI. THE DISCOVERY OF FE-BASED SUPERCONDUCTORS (FeSCs)

To test the capability of our deep learning model of finding new types of superconductor, we investigated whether we could find high-T_c FeSCs by using the model trained with data before 2008, the year FeSCs were discovered. We removed two materials, LaFePO and LaFePFO, from the training data because their discovery in 2006 led to the discovery of high-T_c FeSCs. We used the 1399 FeSCs known as of 2018 in SuperCon as the test data. A total of about 130 training and test runs were used. Although the models were made stochastically, we found some FeSCs that were predicted to have finite T_c. A histogram of the number of predicted FeSCs with $T_c > 0$ K is shown in Fig. 3. We obtained the same results for high-T_c cuprates (see Supplemental Material [27]). When we used shallow 10-layer networks that had as good R^2, precision, etc., as the current large model, FeSCs were not found. This is not strange, because most iron compounds show magnetism, which is incompatible with superconductivity, and there are few superconductors including iron except for FeSCs. Indeed, few researchers had anticipated that FeSCs could have high Tc values. It is recognized that larger models have better generalization performances. The fact that the larger model found FeSCs can be explained by a larger model having an improved search capability for new superconductors. We must mention that random forest models could not find FeSCs. These results suggest that FeSCs and cuprate superconductors might have been found by our deep learning model.

The code and the data are available on the Internet [32].

VII. DISCUSSION

If we had searched for FeSCs following the prediction, we would have discovered FeSCs. However, the predicted T_c of the FeSCs was rather low in our attempt to *discover* FeSCs. FeSCs might thus have been a low-priority target depending on how the model prediction was used. This problem will be considered in future research. We will incorporate crystal structure information to enhance the capability of the model of finding new high-T_c superconductor families. Nevertheless, the present model is still useful as an auxiliary tool. Furthermore, the present method could be applied to other problems where crystal structure is difficult to obtain.

Even though our method does not require feature engineering, unlike conventional methods in materials informatics, it achieved much better results. Our deep learning method may replace existing methods, just as other deep learning methods have done in computer vision, natural language processing, and reinforcement learning. Deep learning requires failure data (e.g., nonsuperconductors) for accurate prediction. As many data sets in materials search are a random train-test split, we must prepare temporally separated train-test data sets for the field to progress. Because our method does not use specific properties of superconductors and uses only chemical formulas, the method can be applied to other problems with, in particular, inorganic materials. We demonstrate band gap estimation by our method in Ref. [33]. We demonstrated the usefulness of our method and deep learning to inorganic materials and condensed matter physics as areas outside organic chemistry, the studies of which have been much reported yet. Our results open the way to the discovery of new high-T_c superconductor families, which must open up new physics.

VIII. SUMMARY OF INTRODUCED METHODS AND THE RESULTS

The summary is given for readers.

A. Summary of introduced methods

1. A deep learning model for finding new superconductors.
2. Reading periodic table: the method that allows deep learning to learn to read the periodic table in order to learn the laws of elements.
3. Garbage-in: the method to create synthetic data on nonsuperconductors.
4. Model evaluation scheme that uses temporally separate training and test data.

B. Summary of results

0. (Good R^2 value for estimating T_c by using data of superconductors only.)
1. The deep learning method predicted superconductivity for a material with a precision of 62%.
2. The deep learning method had better capabilities than random forest.
3. The deep learning method discovered superconductors $CaBi_2$ and $Hf_{0.5}Nb_{0.2}V_2Zr_{0.3}$.

4. The deep learning method found Fe-based high-temperature superconductors (discovered in 2008) from the training data before 2008.

Author contributions. Tomohiko Konno conceived and supervised the research. Tomohiko Konno, Hodaka Kurokawa, Yuki Sakishita, and Fuyuki Nabeshima had the primary roles. Tomohiko Konno, Hodaka Kurokawa, and Fuyuki Nabeshima discussed the direction and interpretation of the analysis, and were the main writers of the manuscript. Tomohiko Konno made the deep learning model and the two methods (reading periodic table and garbage-in), and specified how to evaluate a model using the materials reported by Hosono *et al.* and the temporal separation. Hodaka Kurokawa checked the materials in the candidate materials list and found the two superconductors, and investigated the corresponding original papers to determine the indefinite values in the materials reported by Hosono *et al.* Yuki Sakishita performed random forest analysis. Iwao Hosako brought together the experimenters and machine learning experts. All authors approved the final version of the manuscript for submission.

Data availability. All the data used, SuperCon [28], COD [34–36], and the materials reported by Ref. [1] are openly available. The materials reported by Hosono *et al.* have undetermined variables, such as x in $H_{2-x}O_{1+x}$. We investigated related papers and input the values for such variables. We then made a list of materials for the evaluation of models. This list will be openly available under the condition written for the community in Ref. [32]. See Supplemental Material [27] also for data handling.

[1] H. Hosono, K. Tanabe, E. Takayama-Muromachi, H. Kageyama, S. Yamanaka, H. Kumakura, M. Nohara, H. Hiramatsu, and S. Fujitsu, Sci. Technol. Adv. Mater. **16**, 033503 (2015).

[2] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, Phys. Rev. **106**, 162 (1957).

[3] A. Drozdov, M. Eremets, I. Troyan, V. Ksenofontov, and S. Shylin, Nature (London) **525**, 73 (2015).

[4] M. Somayazulu, M. Ahart, A. K. Mishra, Z. M. Geballe, M. Baldini, Y. Meng, V. V. Struzhkin, and R. J. Hemley, Phys. Rev. Lett. **122**, 027001 (2019).

[5] J. G. Bednorz and K. A. Müller, Z. Phys. B **64**, 189 (1986).

[6] Y. Kamihara, T. Watanabe, M. Hirano, and H. Hosono, J. Am. Chem. Soc. **130**, 3296 (2008).

[7] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).

[8] 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).

[9] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, npj Comput. Mater. **1**, 15010 (2015).

[10] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli *et al.*, Comput. Mater. Sci. **58**, 227 (2012).

[11] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, Nature (London) **559**, 547 (2018).

[12] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, npj Comput. Mater. **3**, 54 (2017).

[13] L. Himanen, M. O. Jäger, E. V. Morooka, F. F. Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, Comput. Phys. Commun. **247**, 106949 (2020).

[14] L. Himanen, A. Geurts, A. S. Foster, and P. Rinke, Adv. Sci. **6**, 1900808 (2019).

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, 2012), pp. 1097–1105.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, in *Advances in Neural Information Processing Systems* (MIT Press, Cambridge, 2014), pp. 2672–2680.

[17] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, Nature (London) **529**, 484 (2016).

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, in *NIPS* (MIT Press, Cambridge, 2017).

[19] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, in *Proceedings of NAACL* (MIT Press, Cambridge, 2018).

[20] H. Pham, M. Y. Guan, B. Zoph, Q. V. Le, and J. Dean, arXiv:1802.3268.

[21] H. Liu, K. Simonyan, and Y. Yang, arXiv:1806.09055 (2018).

[22] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, and M. Sun, arXiv:1812.8434.

[23] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, arXiv:1901.0596.

[24] D. C. Elton, Z. Boukouvalas, M. D. Fuge, and P. W. Chung, arXiv:1903.4388 .

[25] K. Schütt, P.-J. Kindermans, H. E. S. Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller, in *Advances in neural information processing systems* (MIT Press, Cambridge, 2017) pp. 991 1001.

[26] K. Momma and F. Izumi, J. Appl. Crystallogr. **44**, 1272 (2011).

[27] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevB.103.014509 for the methods of reading periodic table and garbage-in, the data, including the availability and its handling, and more details.

[28] National Institute of Materials Science.

[29] V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, and I. Takeuchi, npj Comput. Mater. **4**, 29 (2018).

[30] M. J. Winiarski, B. Wiendlocha, S. Gołąb, S. K. Kushwaha, P. Wiśniewski, D. Kaczorowski, J. D. Thompson, R. J. Cava, and T. Klimczuk, Phys. Chem. Chem. Phys. **18**, 21737 (2016).

[31] Y. Dao-Le, X. Yun-Hui, Z. Zhi-Tao, Z. Li, L. Ji-Zhou, L. Zhu-Qi, Y. Chun-Tang, W. Shan-Ling, and H. Min, Acta Phys. Sin. **32** (1983).

[32] The candidate materials list, the data, the model, and the methods are openly available online at [https://github.com/tomo835g/Deep-Learning-to-find-Superconductors].

[33] T. Konno, J. Phys. Soc. Jpn. **89**, 124006 (2020).
[34] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quiros, N. R. Serebryanaya, P. Moeck, R. T. Downs, and A. Le Bail, Nucleic Acids Res. **40**, D420 (2011).
[35] S. Grazulis, D. Chateigner, R. T. Downs, A. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, and A. Le Bail, J. Appl. Crystallogr. **42**, 726 (2009).
[36] R. T. Downs and M. Hall-Wallace, Am. Mineral. **88**, 247 (2003).