Acceleration search method of higher T_c superconductors by machine learning algorithm

Kaname Matsumoto* and Tomoya Horide

Department of Materials Science and Engineering, Kyushu Institute of Technology, Tobata-ku, Kitakyushu 804-8550, Japan

E-mail: matsu@post.matsc.kyutech.ac.jp

Abstract

We propose a method to efficiently search for superconductors with higher critical temperature T_c by machine learning based on a superconductor database. The T_c prediction and the search for new superconductors are still difficult problems. With the progress of computer power and calculation algorithms, the possibility of finding new materials with higher T_c at high throughput is emerging. Using the obtained T_c prediction model, the scope is expanded to the search space of multielement materials which has never been searched, and candidates for superconductors with higher T_c and which can be synthesized are proposed.

Material development is highly dependent on the experience and intuition of developers, and its development requires a long time and high cost. With the diversification and complexity of materials themselves, it is also becoming difficult to understand their physical properties and functions. There is a need for new development methods that are more efficient and comprehensive ones, as well as provide hints for understanding of the physical properties and functions. Materials Informatics which utilizes first-principles calculations, evolving material databases, and artificial intelligence technology meets these needs.[1,2,3)] There are various theoretical studies on methods to predict the superconducting critical temperature T_c, and the McMillan's equation[4)] and the Allen-Dynes modified equation[5)] have been known. These were used in recent predictions of high T_c of hydrogen sulfide and LaH_{10} under ultra-high pressures[6,7)] and were great motivators to conduct the experiments.[8,9)] However, the prediction requires calculations of electron-phonon coupling parameter after determining the crystal structure, which is somewhat less versatile. As an early study to find the correlation between experimental values of T_c and material parameters, there was also the valence electron rule by Matthias.[10)] This is an empirical rule that T_c is maximal at a certain electron concentration e/a (e: total number of valence electrons, a: total number of atoms), and clear correlations between T_c and electron concentration in transition metals and their alloys have been reported.[11)] Similar studies include heuristic quantum structure diagrams, correlations between normal state properties and superconductivity and so on.[12, 13)] As a modern approach, the method of classifying T_c by acquiring the fingerprints of the Brillouin zone and the density of states, the T_c prediction method by using machine learning (ML) based on a superconductor database, etc., have been reported one after another.[14,15,16,17)] New attempts at such T_c prediction are just beginning, but with the evolution of databases and ML packages, there is a potential for significant development in the future. The purpose of this study is also to predict high T_c materials using ML and database. Especially we focused on the elemental composition rather than the crystal structure and examined the search space for ternary materials. Assuming $\alpha_a\beta_b\gamma_c$ ternary system consisting of a combination of 78 elements α, β, γ from hydrogen (H) in the first period to bismuth (Bi) in the sixth period of the periodic table excluding rare gas elements, the prediction of T_c for all possible compositions a, b, c ($a + b + c = 1$) by using ML has been carried out and the T_c distribution maps for the ternary substance group have been successfully constructed. By comparing with the known stable phases in the equilibrium diagram collected from the material database, it is also possible to identify the composition area of unknown ternary materials with higher T_c. Superconductivity often appears by adding or partially replacing elements in the basic matrix. Under these conditions, calculations on stable crystal structures and electronic states based on density functional theory with the supercell require enormous computational costs, making their execution difficult.[18)] On the other hand, use of superconductor database and ML can reveal the relationship between substance composition and T_c more easily. Based on the composition information of the higher T_c substances thus predicted, efficient experimental verification will be possible. In this study, we aimed to find out unknown higher T_c material candidates by comprehensively examining the region that has not been searched so far.

ML applies statistical analysis methods to a large amount of data, and extracts useful rules and classifications existing among them. For example, when T_c of a superconductor is taken as function F, in addition to the known F values, the features and composition of the accompanying constituent elements are selected as the descriptors x_1, x_2, ..., x_n in a dataset from the search space of a large amount of material data. Then, the computer learns the prediction model of $F = f(x_1, x_2, x_3, \cdots, x_n)$ by the ML procedure.[2] The obtained prediction model can be applied to the combination of other elements and compositions in the vast search space to calculate T_c, and to select the substance groups with high T_c. We used ML to investigate the relationship between known substances and T_c based on the "SuperCon" database,[19] which has been collected by NIMS for many years. Focusing on combinations of ternary systems and below, we obtained the information of both chemical formula and measured T_c values for about 2000 substances including AlB_2, Chevrel, A15 (Cr_3Si), spinel, NaCl (B1), skutterudite type superconductors. The initial datasets were made by adding the T_c values of the elemental superconductors to this information. When the same compound name and composition existed in multiple numbers in the database, the average value was used as T_c. Since we focused on ternary materials, cuprate and Fe-based superconductors[20] were omitted from the datasets; therefore, the substance showing the highest T_c in the datasets is MgB_2 (average T_c of MgB_2 listed in SuperCon is 38.6 K).[21] This mainly corresponds to the limitation to the T_c prediction of superconductors based on the phonon mechanism. In addition to constituent element name and element ratio, "atomic number", "atomic weight", "valence electron number", "period", "group", "van der Waals radius", "covalent bond radius", "Pauling electronegativity", "electron affinity", "first ionization energy", "melting temperature", and "*s*, *p*, *d*, *f* orbital electron numbers" for the target ternary substances were obtained from the existing databases such as Materials Project, Pymatgen and Mendeleev[22,23,24] and these were made the element parameter group x_i. Euclidean norm $\|f\|$ for the composition ratio f_i ($f_1 + f_2 + f_3 = 1$) of constituent elements was used as the stoichiometric attribute.[25] For element parameter groups, mean value x_{mean}, mean deviation x_{avd}, and standard deviation x_{std} were calculated for each composition and made into 53 descriptors in total, and these were combined with T_c values to make 1221 datasets. Equations (1) to (4) used to create the descriptors are given as [26]

$$\| f \| = \sqrt{\sum f_i^2}\ , \tag{1}$$

$$x_{mean} = \sum f_i x_i\ , \tag{2}$$

$$x_{avd} = \sum f_i \left| x_i - x_{mean} \right|, \tag{3}$$

$$x_{std} = \sqrt{\sum f_i \left(x_i - x_{mean} \right)^2}\ . \tag{4}$$

ML procedure for prediction of T_c was performed by dividing the datasets into training data and test data at a ratio of 80:20 at random. We tried ML methods[27] including "regression", "random forest", "support vector machine", etc., at the preliminary examination stage, but in this research, we adopted the "random forest" regression method with the highest prediction accuracy. The cross-validation method was adopted to reduce the bias of predicted values, so the learning process was repeated 10 times to obtain the final values as the average values. The R^2 determination coefficient used to evaluate the prediction accuracy of the learning model is given as

$$R^2 = 1 - \frac{\sum \left(y_i^{\exp} - y_i^{pred} \right)^2}{\sum \left(y_i^{\exp} - \overline{y^{\exp}} \right)^2}\ , \tag{5}$$

where, $y_i^{\exp}$ is the measured value, y_i^{pred} is the predicted value, and $\overline{y^{\exp}}$ is the average value of the measured values. R^2 represents the correlation between the measured values and the predicted values, and the closer to 1.0, the higher the accuracy.

The correlation between predicted T_c and measured T_c by the random forest regression model is shown in Fig. 1. As a result of cross-validation, the R^2 determination coefficients were high, 0.98 for training data and 0.92 for test data. It should be noted that the random forest regression method is powerful for the present datasets and its prediction accuracy is much higher than other regression methods. The reason why R^2 was so high is that the datasets were limited to the ternary group containing MgB_2, which are phonon mechanism superconductors, and high T_c groups such as cuprate and Fe-based superconductors were excluded. The same tendency can be seen in the report of T_c prediction by Stanev[15)] using the SuperCon database as well as present study. The combination of all the elements of the α-β-γ ternary system using 78 elements was ${}_{78}C_3$=76076 in total. In this research, the composition space was divided into 231 points for each combination, and T_c was calculated for each composition using this prediction model. The T_c distribution in the Mg-B-Ti system thus obtained is shown in Fig. 2 as an example. The equilibrium phase diagram at absolute zero including stable and unstable phases, created from Materials Project,[28)] is also shown in the figure. The high T_c region was concentrated near the Mg:B=1:2 composition, and its maximum value was 38 K, which correctly reflects the experimental results. The T_c distribution and the phase diagram in the Fe-Te-Se system found after the discovery of the Fe-based superconductor are shown in Fig. 3. A comparison of the two figures shows that the high T_c region is present on the tie line of FeTe and FeSe. The predicted maximum value of T_c was 14 K, which is similar to the relationship of T_c of $FeTe_{1-x}Se_x$ with x dependence reported by Mizuguchi and Takano.[29)] The datasets do not contain any data on so-called "Fe-based superconductors", but it was possible to accurately predict the T_c distribution of the Fe-Te-Se system. This means that the prediction model works well because the Fe-Te-Se system is a superconductor like the material group of present datasets. The same high prediction accuracy was obtained also in the T_c distribution of the Fe-Te-S system in which Se is substituted with S. However, the situation was different for the Mg-B-Ti system. If the prediction was performed by intentionally excluding MgB_2 and its series (over 150 datasets) from the whole datasets, the prediction accuracy was extremely reduced, and the maximum T_c became 20 K or less. That is, it seems that the present model cannot predict beyond the maximum value of T_c in the datasets.

Although the prediction model has certain limitations, it is very effective for material search of phonon mechanism superconductors, especially material search below the T_c of MgB_2. Therefore, α was fixed to a specific element in the α-β-γ system, and the maximum T_c when various elements of β and γ were replaced was comprehensively calculated and compared. The number of combinations of β and γ elements was ${}_{77}C_2$=2926 in total. As in the previous methods, the maximum T_c value in each system was determined after dividing the composition into 231 points and generating the T_c distribution. The distribution maps of maximum T_c when α was fixed to Fe or B (boron), while β and γ were variously changed to perform exhaustive prediction are shown in Fig. 4. In Fe-β-γ, there was a high T_c region when it contained a light element whose atomic number was smaller than that of Ne, in contrast to this, T_c was lower in systems with larger atomic numbers. On the other hand, in the B-β-γ system containing the light element B, periodic fluctuations of T_c was observed, and high T_c regions also appeared in three places near atomic numbers 10, 20, and 38. The periodic variation of T_c seems to be similar to the behavior in Matthias's valence law. The prediction of high T_c in a system containing light elements is appropriate from the fact that the datasets almost consist of phonon mechanism superconductors. The T_c distribution map for the Ca-B-C system at absolute zero is shown in Fig. 5 as an example of a candidate substance with high T_c. This system is based on the B-C binary system including B, C, $B_{13}C_2$ as stable phases and B_9C, B_4C, BC, BC_5, BC_7 as unstable phases, but these phases do not match regions where high T_c is expected. However, the addition of a third element such as Ca broadens the high T_c region. For example, if a compound can be successfully synthesized on a tie line of CaB_6 and $B_{13}C_2$ or in the vicinity, a high T_c substance exceeding 30 K is expected. Such a T_c distribution map can be seen not only with the addition of Ca but also with the addition of other elements. In boron-carbide system, 23 K has already been reported in the YPd_2B_2C quaternary system,[30)] and recently, there has also been a report of $T_c = 36$ to 55 K in the B-doped Q carbon thin film,[31)] so, this system is expected as a target substance of experimental verification.

In summary, ML was used to establish a T_c prediction model for ternary materials based on the SuperCon database. The R^2 determination coefficient showed high accuracy of 0.92, and the T_c distribution maps of Mg-B-Ti system and Fe-Te-Se system predicted using this model showed a good correspondence with the experimental results, despite the lack of crystal structure information. Even though the data of Fe-based superconductors were not included in the datasets, the fact that T_c could be predicted correctly indicates the effectiveness of the prediction model. We established an algorithm to predict the maximum T_c of various α-β-γ systems using this model and revealed that the T_c of systems containing light elements is high and that the behavior of T_c periodically changes as the atomic number increases. We also suggested the Ca-B-C system as a candidate substance of high T_c. In order to search for higher T_c, it is necessary to develop the present prediction method into a quaternary or a five-element system including cuprate and Fe-based superconductors.

References

1) A. Agrawal and A. Choudhary, APL Mater. **4**, 053208 (2016).

2) R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, Npj Comput. Mater. **3**, 54 (2017)

3) L. Ward and C. Wolverton, Curr. Opin. Solid State Mat. Sci. **21**, 167 (2017).

4) W. McMillan, Phys. Rev. **167**, 331 (1968).

5) P. Allen and R. Dynes, Phys. Rev. B **12**, 905 (1975).

6) Y. Li, J. Hao, H. Liu, Y. Li, and Y. Ma, J. Chem. Phys. **140**, 174712 (2014).

7) H. Liu, I. Naumov, R. Hoffmann, N. Ashcroft, and R. Hemley, PNAS **114,** 6990 (2017).

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

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

10) B. Matthias, Phys. Rev. **97**, 74 (1955).

11) J. Hulm and R. Blaugher, Phys. Rev. **123**, 1569 (1961).

12) P. Villars and J. Phillips, Phys. Rev. B **37**, 2345 (1988).

13) J. Hirsch, Phys. Rev. B **55**, 9007 (1997).

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

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

16) B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I.Foster, B. Gibbons,d, J. Hattrick-Simpers, A. Mehtaf, and L. Ward, Mol. Syst. Des.

3, 819 (2018).

17) K. Hamidieh, Comput. Mater. Sci. **154**, 346 (2018).

18) A. Erba, J. Baima, I. Bush, R. Orlando, and R. Dovesi, J. Chem. Theory Comput. **13**, 5019 (2017).

19) Superconducting Material Database (SuperCon), National Institute of Materials Science, available at: http://supercon.nims.go.jp/index_en.html

20) M. Norman, Rep. Prog. Phys. **79**, 074502 (2016).

21) C. Buzea and T. Yamashita, Supercon. Sci. Technol. **14**, R115 (2001).

22) A. Jain, S. Ong, G. Hautier, W. Chen, W. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, K. Ceder, K. Persson, APL Mater. **1**, 011002 (2013).

23) S. Ong, W. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. Chevrier, K.Persson, G. Ceder, Comput. Mater. Sci. **68**, 314 (2013).

24) L. M. Mentel, mendeleev -A Python resource for properties of chemical elements, ions and isotopes. 2014–. available at: https://bitbucket.org/lukaszmentel/mendeleev.

25) L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, Npj Comput. Mater. **2**, 16028 (2016).

26) A. Furmanchuk, J. Saal, J. Doak, G. Olson, A. Choudhary, and A. Agrawal, J. Comput. Chem. **39**, 191 (2018).

27) F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, and J. Vanderplas, J. Mach. Learn. Res. **12**,2825 (2011).

28) S. Ong, L. Wang, B. Kang, and G. Ceder, Chem. Mater. **20**, 1798 (2008).

29) Y. Mizuguchi and Y. Takano, J. Phys. Soc. Jpn. **79**, 102001 (2010).

30) R. Cava, H. Takagi, B. Batlogg, H. Zandbergen, J. Krajewski, W. Peck Jr, and J. Lee, Nature **367**, 146 (1994).

31) A. Bhaumik, R. Sachan, S. Gupta, and J. Narayan, ACS nano **11**, 11915 (2017).

Figure captions

Fig. 1 Correlation between predicted T_c and measured T_c by the random forest regression model. The 1221 datasets were divided into training data and test data at a ratio of 80:20. According to cross validation, the R^2 determination coefficient is 0.92.

Fig. 2 (a) The predicted T_c distribution map of Mg-B-Ti system. (b) The equilibrium phase diagram of Mg-B-Ti system at absolute zero, where black points are stable phases and blue ones are unstable phases. The high T_c region exists only near the Mg: B = 1: 2 composition, and the maximum T_c is 38 K.

Fig. 3 (a) The predicted T_c distribution map of Fe-Te-Se system. (b) The equilibrium phase diagram of Fe-Te-Se system at absolute zero, where black points are stable phases and blue ones are unstable phases. The high T_c region is present on a tie line of FeTe and FeSe, which corresponds to the relationship of T_c of $FeTe_{1-x}Se_x$ with x dependence reported by Mizuguchi and Takano.[29)]

Fig. 4 Predicted maximum T_c distribution map of α-β-γ system when α was fixed to Fe or B, while β and γ were variously changed. (a) Fe-β-γ system. (b) B-β-γ system. The number of composition combinations in each system is ${}_{77}C_2$=2926.

Fig. 5 (a) The predicted T_c distribution map of Ca-B-C system. (b) The equilibrium phase diagram of Ca-B-C system at absolute zero, where black points are stable phases and blue ones are unstable phases. The predicted high T_c region is on a tie line of CaB_6 and $B_{13}C_2$ or in the vicinity and the maximum T_c is 36 K.

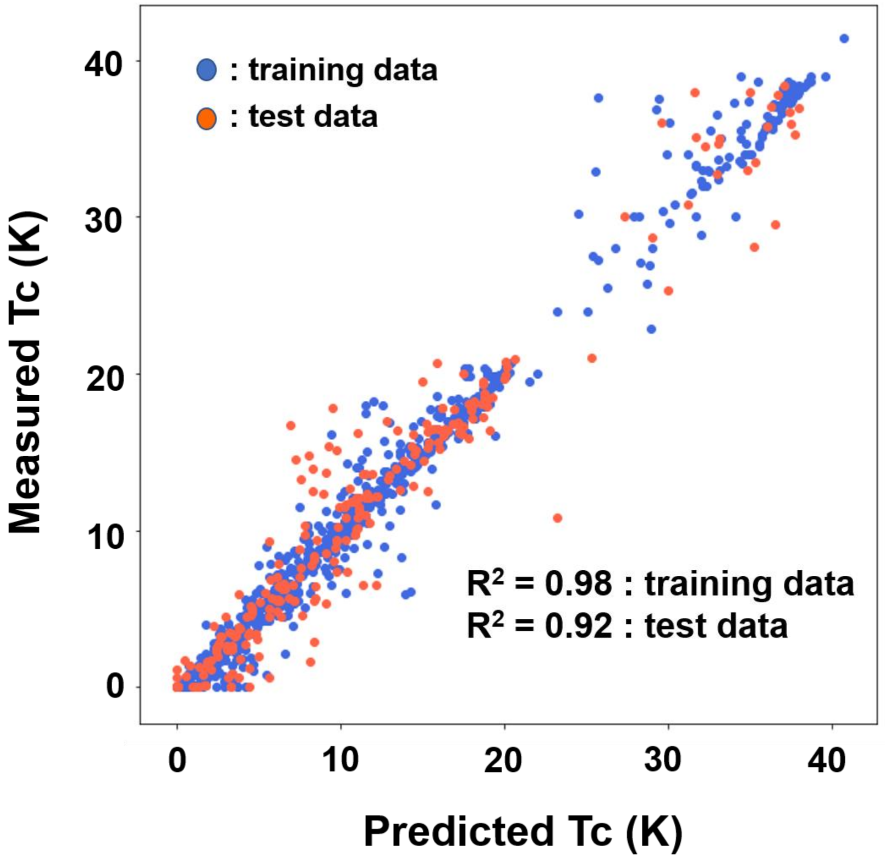

Fig. 1   K. Matsumoto

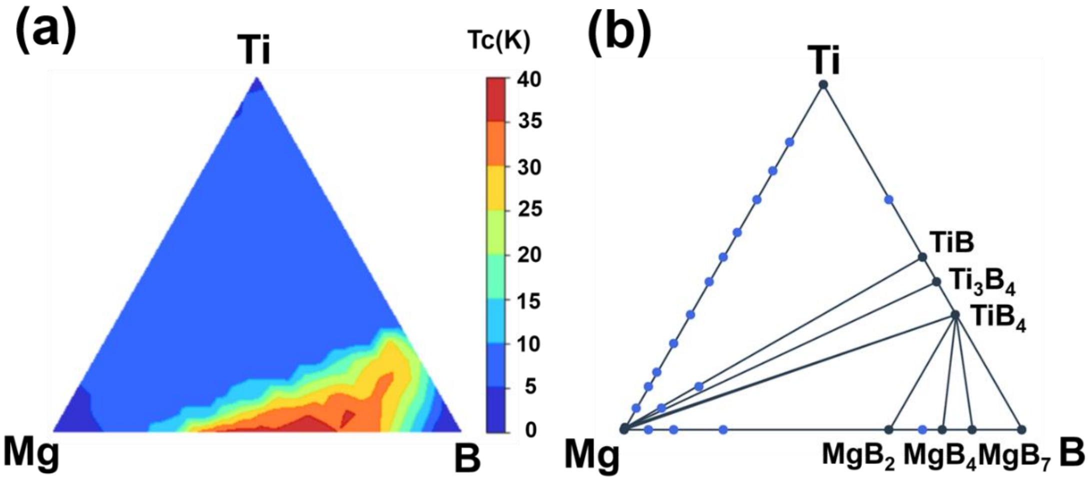

Fig. 2 K. Matsumoto

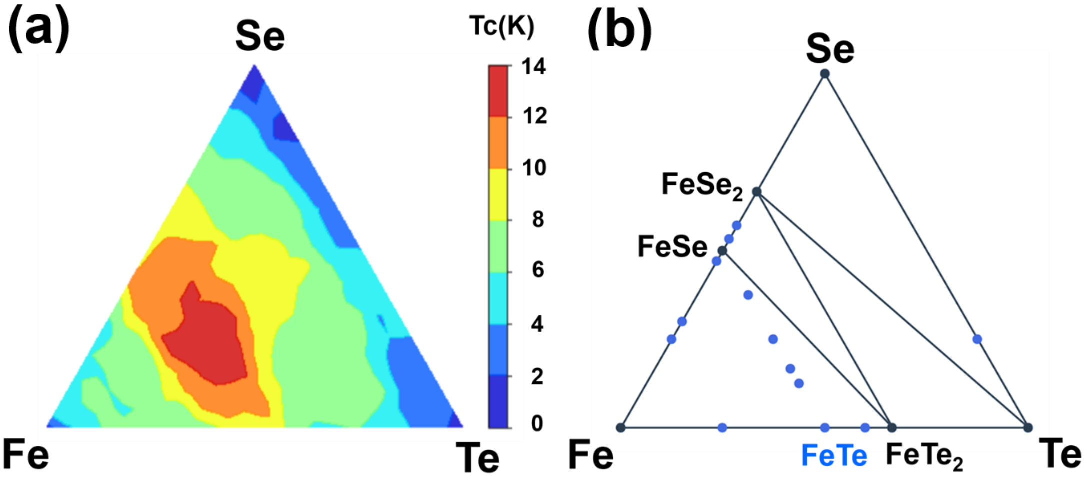

Fig. 3   K. Matsumoto

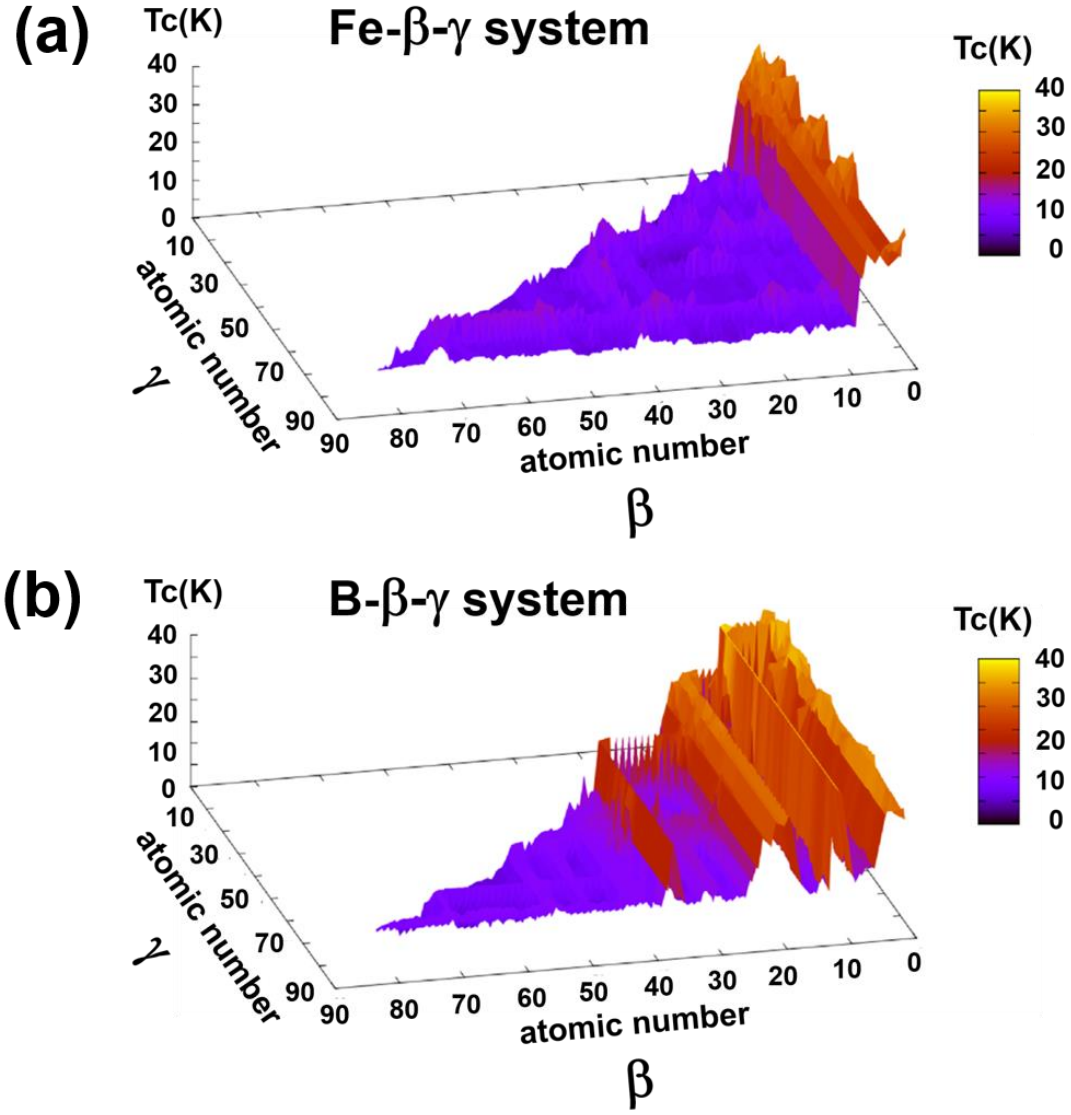

Fig. 4 K. Matsumoto

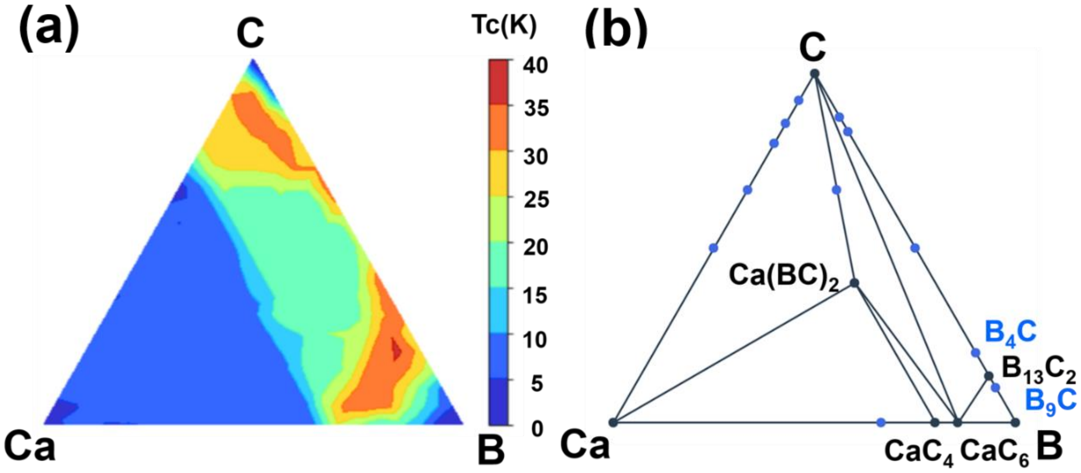

Fig. 5 K. Matsumoto