Applied Physics Letters ARTICLE scitation.org/journal/apl

Structure prediction in high-entropy alloys with machine learning

Cite as: Appl. Phys. Lett. **118**, 231904 (2021); doi: 10.1063/5.0051307
Submitted: 24 March 2021 · Accepted: 30 May 2021 ·
Published Online: 10 June 2021

D. Q. Zhao,[1] S. P. Pan,[1,2] Y. Zhang,[3] P. K. Liaw,[4] and J. W. Qiao[1,2,a)]

AFFILIATIONS

[1]College of Materials Science and Engineering, Taiyuan University of Technology, Taiyuan 030024, China
[2]Key Lab of Interface Science and Engineering in Advanced Materials, Ministry of Education, Taiyuan University of Technology, Taiyuan 030024, China
[3]State Key Laboratory for Advanced Metals and Materials, University of Science and Technology Beijing, Beijing 100083, China
[4]Department of Materials Science and Engineering, The University of Tennessee, Knoxville, Tennessee 37996-2200, USA

Note: This paper is part of the APL Special Collection on Metastable High Entropy Alloys.
[a)]**Author to whom correspondence should be addressed:** qiaojunwei@gmail.com

ABSTRACT

High-entropy alloy is an alloy design concept without a principal component. This concept not only refers to the complexity of alloy compositions but also means that when the high-entropy alloy transits from a high-energy state to low-energy state, there will be more intermediate metastable states. Corresponding to different states are the changes in the degree and manner of order in the microstructure. In this study, we used machine learning to combine elemental characteristics with long-term ordering and established 87% of prediction accuracy. This data-driven method can correlate elemental characteristics and metastable states and accelerate the discovery of potential compositions.

High-entropy alloys (HEAs), also known as complex concentrated alloys, are becoming a research frontier in the metallic-material field due to their unique characteristics and unprecedented application potentials.[1,2] The birth of HEAs is accompanied by the discussion of unusual phase stability. The concept was hypothesized to emphasize the stabilizing effect of high configuration entropy on the solid solution (SS) phase. With the in-depth research in the past ten years, more and more examples have proved that this effect is not as significant as people once thought.[3] Although attention has shifted from searching for single-phase solid solutions to microstructure design,[4,5] the abnormal short-range or long-range order distribution is still the focus.[6,7] Atoms in solid solutions may exhibit short-range order or cluster, which can be regarded as the initial characteristics of condensed matter. The degree of order may be further expanded to form long-range order or concentration fluctuations.[6–9] As a transition state from the micro- to macro-condition, this initial state will affect many aspects of HEAs. When considering the vastness of the HEA space, only a few hundred individual alloys have been investigated so far, and meanwhile, their discovery has been largely done by a costly and time-consuming methodology. It is necessary to develop accelerated strategies for the composition exploration.

Considering the importance of high entropy alloy structure, people have studied them from many different aspects, such as phase-diagram calculations[10] and *ab initio* methods.[11] However, these methods have some inherent limitations, such as the need for accurate databases. Although the thermodynamic phase diagrams and interatomic potentials have been gradually improved, they still cannot meet the needs of people. In the actual experimental process, it is difficult to achieve the complete equilibrium conditions. Hence, we put the research object on the metastable alloy dataset. Moreover, scientists are keen to explore heuristic models from a data-centric method.[12–14] People use domain knowledge and experience to extract the characteristics related to the existence of solid solutions, such as atomic radii and electronegativities. With the continuous exploration of alloy compositions, the features originally proposed, based on small datasets, need to be reviewed. Furthermore, though all the parameters proposed in the past have some physical groundings, a simple linear combination of these parameters cannot provide sufficiently robust predictions for the likely phase constitution.[15] On the one hand, several research activities have been devoted to using professional knowledge to find more features with a physical background.[16–18] On the other hand, more accurate phenomenological models or algorithms are pressingly needed to establish to guide alloy designs.[19–21] The former will

determine the upper limit of the prediction accuracy, while the latter will approach this upper limit.

In recent years, as HEAs have received more and more attention, the generation of experimental data has been dramatically accelerated, making the value mining of data more complicated and difficult. Machine learning can provide a solution that shows good applicability in the classification, regression, and other tasks related to high-dimensional data. Machine learning aims to extract knowledge and gain insights from massive databases. It learns from previous calculations to produce reliable and repeatable decisions and results, which has played an influential role in many fields.[22–26] In the present work, we reviewed some of the previously proposed features with a large data set, then employed a self-organizing algorithm (SOM) to comprehensively analyze these features, and further utilized MATLAB[27] to build some common ML algorithms [such as a k-nearest neighbor model (KNN), support vector machine model (SVM), and artificial neural network model (ANN)] to find the most suitable ML model for predicting more HEAs. The current database consists of 482 alloy components, each of which has at least four components to comply with the concept of high entropy. We select the as-cast structure from different research. The instability of the as-cast structure itself may lead to accidental differences. We hope to find inevitability from accidental through big data.

Compared to binary alloys, the complex alloy compositions will blur the difference between the solute and solvent. Accordingly, the concept of solid solutions and intermetallic compounds (IMs) will be further expanded. Moreover, the order degree and order mode of atom arrangements make the solution's behavior in HEAs more complex. As a macroscopic reaction, the phase can reveal the complex mechanism of a high-entropy solid solution. Miracle and Senkov proposed that alloys can be divided into three categories according to their degree of long-range order.[28] If atoms randomly occupy lattice positions, it is a solid solution (SS). If it is a complete long-range order, it is an intermetallic compound (IM). If both are present, it is (IM + SS).

Feature engineering is the most important part and most creative part of the data-centric method. Different phenomenological models make various assumptions and approximations. Thus, the formation of phases in multicomponent systems may be related to many factors. The selected features in the present work include the valence electron concentration (VEC),[17] atomic-size mismatch (δ),[18] configuration entropy (S_{mix}), electronegativity difference ($\Delta\chi$),[29] and melting point of mixing (T_{mix}). Parameters from Miedema include the enthalpy of mixing (H_{mix}), enthalpy of competing phases (H_{max}), precipitation temperature of competing phases (T_{com}), and phase stability parameter (Φ).[19] The feature calculation method can be found in the supplementary material. In order to display a group of data intuitively in a high-dimensional space, the parallel coordinate diagram is selected. A point in an n-dimensional space is represented as a polyline with vertices on the parallel axes. The position of the vertex on the ith axis corresponds to the ith coordinate of the point, and the ordinate is a dimensionless value. The goal is to make different features comparable: scale the data proportionally without changing the data distribution so that they fall into similar intervals.

It can be seen from Fig. 1 that a solid solution will not be formed when the lattice distortion and electronegativity difference are too large. There is a large overlap area between the solid solution and intermetallic phase on the coordinate axis. At the same time, the simple linear division of a single feature cannot meet the needs of alloy design. The solid solution is mainly concentrated on the upper and lower ends on the VEC axis, which can largely distinguish the lattice structures. Different from the original hypothesis, high entropy will not inhibit the emergence of IM. On the contrary, with the increase in entropy, the probability of the occurrence of IMs increased. Senkov *et al.* obtained similar results using the calculated phase-diagram method.[10] They rapidly screened more than 130 000 systems and found that with the increase in the configuration entropy, the possibility of forming IM pairs was increased. Schön *et al.* employed cluster variation methods to study the order–disorder transformation in body-centered-cubic (BCC) systems and then extended the model from the symmetric to the asymmetrical condition. They found that the entropy effect in HEAs played a relatively minor role in stabilizing simple and disordered lattice structures.[30] It can be seen from the T_{mix} axis that with the increase in the melting point, the proportion of a SS phase is enhanced, which may be related to the entropy enhancement at high temperatures or the nondense lattice structure of the BCC phase of refractory elements.

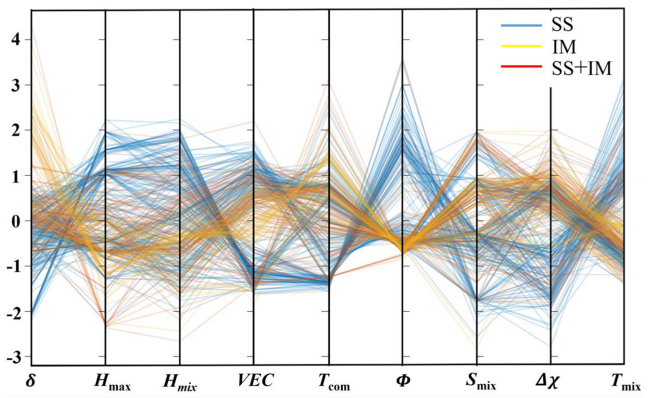

FIG. 1. Parallel coordinate map of HEA composition features.

Compared with other features, Φ can provide a higher degree of discrimination between SS and IM. King *et al.* suggested that Φ can be used to denote the ability of the phase separation.[19] The Φ factor is the ratio of the maximum two elements separated enthalpy of formation to the enthalpy in solid solutions. The enthalpy comes from the Miedema rule related to the boundary electron density and the chemical potential of the metal's electronic charge. The Miedema rule is one of the best known heuristic models proposed by Miedema *et al.* in the mid-1970s,[31] describing the energy effects during alloying. This model makes it possible to predict the formation enthalpies of binary-alloy systems. Although the Miedema rule as a semi-empirical criterion lacks quantum physical meaning in the long-term practice process, it still reflects the thermodynamic effects of binary alloys well.

Different features will play a particular role in classification, but no feature is decisive. The range of different features and the classification effect in this range are diverse for different input elements. What is needed is not a linear combination rule but a mixed decision-making judgment mechanism. In many works, the atomic-size mismatch is considered to be the most critical factor. The atomic size is an

abstract concept in quantum mechanics, and different phenomenological models will explain it from various aspects. For example, the atomic-size mismatch may affect the stacking mode in the steel-ball model,[32] produce the excess entropy,[33] and change the liquid viscosity.[34] At the same time, the atomic size is an important supplement to the Miedema rule. The original Miedema rule does not consider the influence of different atomic sizes on the boundary electron concentration in a complex environment.

To further understand the data screening, the SOM method is employed to visualize and extract the information. SOM is a type of ANN, which has many advantages over the conventional feature-extraction methods, such as empirical orthogonal functions (EOFs) or principal component analysis (PCA). PCA can reduce the dimensionality of variables but obscure the original physical meaning when mapped to an unknown space, and the visibility is poor. SOM uses a neighborhood function to preserve the topological properties of the input space and shows more ability in visualization by creating a low-dimensional view of the high-dimensional data. The SOM algorithm established in MATLAB is used to obtain the two-dimensional weight plane in the current work. The number of output neurons is vital to detect the deviation of the data. After trying different grid sizes, 10×10 units are selected, large enough to represent features and small enough to be interpreted.

SOM is designed to mimic the reaction process of the human brain when processing different information. For example, visual and auditory signals cause responses in different parts of the brain, and similar external information is continuously mapped in the corresponding areas. As shown in Fig. 2(a), the alloy data are applied to train SOM. Neurons will compete according to the Euclidean distance of the input vector and the weight of neurons, and the best matching unit (BMU), which has a minimum distance, is the winner. Both the BMU and its neighborhood neurons tend to move toward the winner. The weights of neurons will be updated during the move. In this way, similar vectors will be clustered by neurons with similar weights.

Figure 2(b) plots a SOM layer showing neurons as gray–blue patches and their direct neighbor relations with red lines. The neighbor patches are colored from black to yellow to show how close each neuron's weight vector is to its neighbors. The darker color corresponds to the larger distance.

Figure 2(c) shows the number of input vectors that neurons classify. The samples in the figure are divided into two categories: upper right and lower left. IMs are mainly concentrated in the lower-left corner, and the solid solution is primarily concentrated in the upper right corner. From the current alloy development model, most of the alloy composition adjustment is based on a major alloy composition, and some neurons are highly enriched.

Each subplot shows the weights from the features input to the layer's neurons. When the two-weight planes are similar, the two factors have similar effects on the classification results. It can be seen from the present results that different features are independent of each other and have various effects on the classification. Compared with the selection of features through Pearson coefficients, the weight map can provide a more intuitive view of the classification's impact.

It should be emphasized that different feature combinations will have different effects on the weight map, and the weight map obtained by the same feature combination is not unique according to the variation in hyperparameters, but there will be a general trend. The concept of long-range order is unknown in advance to the algorithm itself, and

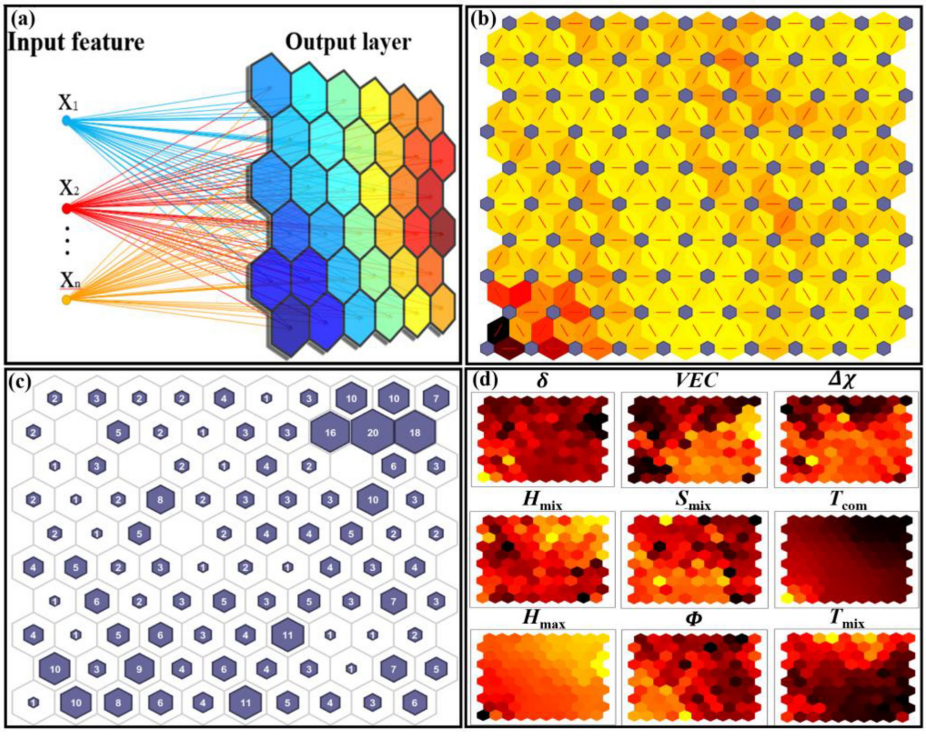

FIG. 2. Feature analysis using self-organizing map. (a) Principle of self-organization map; (b) neighboring weight distances; (c) a Hits map that neurons classify; and (d) weight planes.

the purpose is to reveal the inherent properties and laws of the data through the learning of unlabeled samples. The impact of different features on the classification results can be compared more intuitively through the weight map. A transparent gradient in the distribution represents a high contribution to the classification. It can be seen from the figure that the H_{com} and T_{com} factors gradually change from the upper right to lower left. The phase competing factors represented by these two features have a significant and global influence on the classification of alloys concerning the long-range ordering degree. Huang *et al.* found a global gradient from left to right in the H_{mix} weight plane.[21] This trend is supposed to be the correlation between H_{max} and H_{mix}. Both terms are essentially related to the bond energy of the first adjacent pair of atoms. H_{mix} may be a proxy for the H_{max} system. This kind of agent can only vaguely distinguish between SS and IM.

Senkov and Miracle proposed a criterion based on both enthalpy and entropy terms of the competing phases and tested it in 45 HEA datasets.[35] The model has achieved good results in predicting the phase-formation ability. This study expands the verification set and obtains similar results in a more extensive range compared with them. Phase competition is an essential factor that cannot be ignored for the phase-formation problem.[3,35] Though some studies have been done,[16,28] an effective model is still lacking. Troparevsky *et al.* calculated the formation enthalpy of binary compounds by a high-throughput density-functional-theory (DFT)-based method and proposed a criterion for predicting single-phase solid solutions.[16] However, the criterion proposed by them has great limitations: the predicted results can only correspond to the alloy in the ideal equilibrium state. Therefore, the conclusion is limited, ignoring many alloy compositions with a stable phase. Instead, we want to build a more optimized model applied to larger systems and make the distinction more accurate.

Different feature combinations are applied to various algorithms and adjust the hyperparameters to achieve the maximum accuracy. The data projected to the high-dimensional space are easier to linearly separate, but the current alloy data are limited. Therefore, it is necessary to select an appropriate number of features to achieve the best prediction effect and prevent overfitting. According to the principle that there is no free lunch in the world,[36] different algorithms have various effects on solving different problems. We have tried some commonly used algorithms to find the optimal solution to the problem. The most accurate is the ANN algorithm, which can reach 87% in the case of three classifications. Similar results have been obtained in previous articles: the artificial neural network has higher classification ability than other algorithms.[20] The factors that affect the stability of phase may be hierarchical, and the structure of the neural network can deal with such hierarchical data more easily.

The present work utilized ANN to build a model to identify and classify data. An ANN is based on a series of connected units or nodes called artificial neurons, which simulate the function of biological neurons. Neural networks are trained by processing examples. Each example contains a known "input" and "outcome" to form a probability-weighted association between the two. These associations are stored in the data structure of the network. The ANN consists of three parts: the input layer, hidden layer, and output layer. We found that when the hidden layer is less than three, the expected test accuracy is not up to the expected effect. When the hidden layer is increased to five, increasing the number has no obvious effect on the result or slightly decreases. Therefore, we set the number of hidden layers to five. The tansig [tansig(n)$= \frac{2}{1+e^{-2n}} - 1$] function as the activation function updates the weights and deviations according to the Levenberg–Marquardt optimization[37] method, which is used to solve nonlinear least squares problems. The features of all samples are normalized, and we adopt 0.01 as the learning rate. For the number of neurons in the hidden layer, our choices are from 5 to 50 to ensure the range-wide enough. To remedy the problem of overfitting, we apply tenfold cross-validation. The accuracy mentioned in this article refers to as global average accuracy.

People tend to be more interested in an alloy with a certain composition because it shows better properties. We further screened the database to exclude some similar components. The same proportion of FCC and BCC alloys were selected, and the repeated sampling method was used to compensate the imbalance of different phase numbers in the database. Figure 3 shows the accuracy of different features combinations in different algorithms. By introducing some amorphous data, the prediction effect is improved. Although the increase is only 2%, it is still considerable. Metallic glasses, high entropy alloys, and conventional crystalline alloys can be regarded as three macro structures with different short-range ordering degrees. In practical applications, one hopes to better distinguish between SS and SS + IM. The accuracy of our classification is up to 90%. This trend shows that the features added greatly improve the classification ability. The phenomenological model is not proposed in the present work since the phase formation factors may be of multi-level and multi-condition. The Occam razor (entities should not be multiplied unnecessarily) may not be suitable for this question, and further descriptions require a larger dataset.

The present method has the following limitations. Data-based algorithms have significant requirements on the quantity and quality of data.[38] The compositions of HEAs are still limited, and many of them are developed based on several alloys with superior properties,

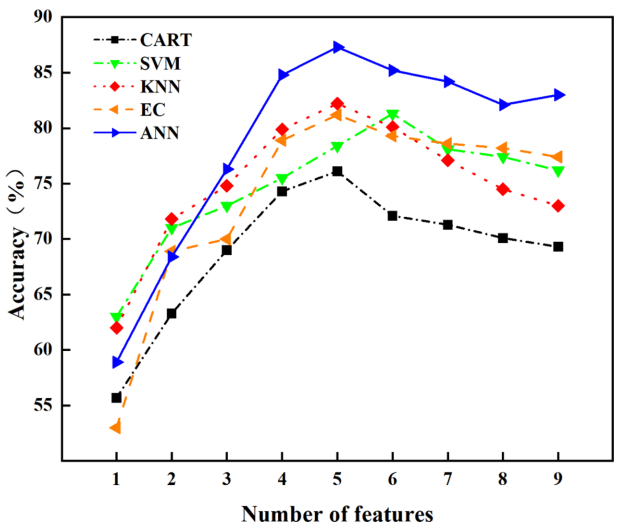

FIG. 3. The accuracy varies with the combination of features in different algorithms. A regression-tree model (CART), a support vector machine model (SVM), a k-nearest neighbor model (KNN), an ensemble classification model (EC), and a feed-forward artificial neural network model (ANN).

which will have a particular anchoring effect on the data. At that time, a public database that worked together could overcome this limitation. The data of those failed experiments can provide exceptionally useful information.[39] The algorithm can play a tremendous advantage in an extensive database, but the quality of the database is also crucial. Although the introduction of the binary and ternary alloy data can expand the database, it will obscure the inherent effects of complex alloys. Binary and ternary alloy information can be introduced, but more effective ways should be adopted, such as transfer learning.

See the supplementary material for the calculation of the parameters to predict the phase formation rules on high-entropy alloys.

The authors would like to acknowledge the financial support of the National Natural Science Foundation of China (No. 52071229) and the Natural Science Foundation of Shanxi Province, China (Nos. 201901D111105 and 201901D111114). P.K.L. very much appreciates the supports from (1) the National Science Foundation (Nos. DMR-1611180 and 1809640) with program directors, Dr. J. Yang, Dr. G. Shiflet, and Dr. D. Farkas, and (2) the U.S. Army Research Office (Nos. W911NF-13-1-0438 and W911NF-19-2-0049) with program managers, Dr. M. P. Bakas, Dr. S. N. Mathaudhu, and Dr. D. M. Stepp.

DATA AVAILABILITY

The data that support the findings of this study are available within the article and its supplementary material.

REFERENCES

[1]B. Cantor, I. T. H. Chang, P. Knight, and A. J. B. Vincent, Mater. Sci. Eng. A **375–377**, 213 (2004).
[2]Y. Zhang, T. T. Zuo, Z. Tang, M. C. Gao, K. A. Dahmen, P. K. Liaw, and Z. P. Lu, Prog. Mater. Sci. **61**, 1–93 (2014).
[3]F. Otto, Y. Yang, H. Bei, and E. P. George, Acta Mater. **61**, 2628 (2013).
[4]L. J. Santodonato, Y. Zhang, M. Feygenson, C. M. Parish, M. C. Gao, R. J. K. Weber, J. C. Neuefeind, Z. Tang, and P. K. Liaw, Nat. Commun. **6**, 5964 (2015).
[5]Z. Li, K. Pradeep, Y. Deng, D. Raabe, and C. C. Tasan, Nature **534**, 227 (2016).
[6]Q. Ding, Y. Zhang, X. Chen, X. Fu, D. Chen, S. Chen, L. Gu, F. Wei, H. Bei, Y. Gao, M. Wen, J. Li, Z. Zhang, T. Zhu, R. O. Ritchie, and Q. Yu, Nature **574**, 223 (2019).
[7]M. S. Lucas, G. B. Wilks, L. Mauger, J. A. Muñoz, O. N. Senkov, E. Michel, J. Horwath, S. L. Semiatin, M. B. Stone, D. L. Abernathy, and E. Karapetrova, Appl. Phys. Lett. **100**, 251907 (2012).
[8]F. X. Zhang, S. Zhao, K. Jin, H. Xue, G. Velisa, H. Bei, R. Huang, J. Y. P. Ko, D. C. Pagan, J. C. Neuefeind, W. J. Weber, and Y. Zhang, Phys. Rev. Lett. **118**, 205501 (2017).
[9]P. Singh, A. V. Smirnov, and D. D. Johnson, Phys. Rev. B **91**, 224204 (2015).
[10]O. N. Senkov, J. D. Miller, D. B. Miracle, and C. Woodward, Nat. Commun. **6**, 6529 (2015).
[11]Y. Lederer, C. Toher, K. S. Vecchio, and S. Curtarolo, Acta Mater. **159**, 364 (2018).
[12]C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, Nat. Mater. **5**, 641 (2006).
[13]S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder, Phys. Rev. Lett. **91**, 135503 (2003).
[14]Y. Li and W. Guo, Phys. Rev. Mater. **3**, 095005 (2019).
[15]F. Tancret, I. Toda-Caraballo, E. Menou, and P. E. J. R. Díaz-Del, Mater. Des. **115**, 486 (2017).
[16]M. C. Troparevsky, J. R. Morris, P. R. C. Kent, A. R. Lupini, and G. M. Stocks, Phys. Rev. X **5**, 011041 (2015).
[17]S. Guo, C. Ng, J. Lu, and C. T. Liu, J. Appl. Phys. **109**, 103505 (2011).
[18]X. Yang and Y. Zhang, Mater. Chem. Phys. **132**, 233 (2012).
[19]D. J. M. King, S. C. Middleburgh, A. G. McGregor, and M. B. Cortie, Acta Mater. **104**, 172 (2016).
[20]Y. Zhang, C. Wen, C. Wang, S. Antonov, D. Xue, Y. Bai, and Y. Su, Acta Mater. **185**, 528 (2020).
[21]W. Huang, P. Martin, and H. Zhuang, Acta Mater. **169**, 225 (2019).
[22]G. Kim, H. Diao, C. Lee, A. T. Samaei, T. Phan, M. Jong, K. An, D. Ma, P. K. Liaw, and W. Chen, Acta Mater. **181**, 124 (2019).
[23]C. Lee, G. Kim, Y. Chou, B. L. Musicó, M. C. Gao, K. An, G. Song, Y. C. Chou, V. Keppens, W. Chen, and P. K. Liaw, Sci. Adv. **6**, eaaz4748 (2020).
[24]J. M. Rickman, G. Balasubramanian, C. J. Marvel, H. M. Chan, and M. T. Burton, J. Appl. Phys. **128**, 221101 (2020).
[25]J. Qi, A. M. Cheung, and S. J. Poon, Sci. Rep. **9**, 15501 (2019).
[26]B. Steingrimsson, X. Fan, A. Kulkarni, M. C. Gao, and P. K. Liaw, arXiv:2012.07583 (2020).
[27]S. Sivanandam and S. Deepa, *Introduction to Neural Networks Using MATLAB 6.0* (Tata McGraw-Hill Education, 2006).
[28]D. B. Miracle and O. N. Senkov, Acta Mater. **122**, 448 (2017).
[29]R. Martin, T. Zeng, and H. Roald, J. Am. Chem. Soc. **141**, 342 (2019).
[30]C. G. Schön, T. Duong, Y. Wang, and R. Arryave, Acta Mater. **148**, 263 (2018).
[31]D. Pettifor, Phys. Solid State **40**, 43 (1987).
[32]Y. J. Zhou, Y. Zhang, F. J. Wang, and G. L. Chen, Appl. Phys. Lett. **92**, 241917 (2008).
[33]Q. F. He, Y. F. Ye, and Y. Yang, J. Appl. Phys. **120**, 154902 (2016).
[34]C. Chattopadhyay, A. Prasad, and B. S. Murty, Acta Mater. **153**, 214 (2018).
[35]O. N. Senkov and D. B. Miracle, J. Alloys Compd. **658**, 603 (2016).
[36]D. H. Wolpert and W. G. Macready, IEEE Trans. Evol. Comput. **1**, 67 (1997).
[37]M. Al-Baali and R. Fletcher, J. Oper. Res. Soc. **36**, 405 (1985).
[38]Y. Zhang and C. Ling, npj Comput. Mater. **4**, 25 (2018).
[39]P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist, Nature **533**, 73 (2016).