Acta Materialia 243 (2023) 118497

Contents lists available at ScienceDirect

Acta Materialia

journal homepage: www.elsevier.com/locate/actamat

Machine learning versus human learning in predicting glass-forming ability of metallic glasses

Guannan Liu[a], Sungwoo Sohn[a], Sebastian A. Kube[a], Arindam Raj[a], Andrew Mertz[a], Aya Nawano[a], Anna Gilbert[b,c], Mark D. Shattuck[d], Corey S. O'Hern[a,e,f], Jan Schroers[a,*]

[a] Department of Mechanical Engineering and Materials Science, Yale University, New Haven, CT 06520, USA
[b] Department of Mathematics, Yale University, New Haven, CT 06520, USA
[c] Department of Statistics & Data Science, Yale University, New Haven, CT 06520, USA
[d] Benjamin Levich Institute and Physics Department, The City College of the City University of New York, New York, NY 10031, USA
[e] Department of Applied Physics, Yale University, New Haven, CT 06520, USA
[f] Department of Physics, Yale University, New Haven, CT 06520, USA

ARTICLE INFO

Article history:
Received 18 July 2022
Revised 27 October 2022
Accepted 31 October 2022
Available online 2 November 2022

Keywords:
Machine learning
Human learning
Materials design
Metallic glass
Glass-forming ability

ABSTRACT

Complex materials science problems such as glass formation must consider large system sizes that are many orders of magnitude too large to be solved by first-principles calculations. The successful application of machine learning (ML) in various other fields suggests that ML could be useful to address complex problems in materials science. To test its efficacy, we attempt to predict bulk metallic glass formation using ML. Surprisingly, we find that a recently developed ML model based on 201 alloy features constructed using simple combinations of 31 elemental features is indistinguishable from models that are based on unphysical features. The 201ML-model performs better than the unphysical model only when significant separation of training and testing data is achieved. However, it performs significantly worse than a human-learning based three-feature model. The limited performance of the 201ML-model originates from the inability to accurately represent alloy features through elemental features, showing that physical insights about mixing behavior are required to develop predictable ML models.

1. Introduction

The transformative success of machine learning (ML) strategies in a wide range of fields, including facial recognition, speech recognition, consumer behavior, and drug discovery, has triggered the consideration of such strategies in materials science. Even though much scarcer than in other fields, its application is rapidly increasing. Materials science problems that have been addressed using ML can be categorized very generally into two categories. The first category includes problems that can be reduced to a small number of atoms. Such problems can be, to a large extent, addressed through *ab initio* approaches and include formation energies [1,2], band gaps [3–5], elastic moduli [3,4,6], and crystal structures [7,8]. Even though there are still limitations in the representation, synthesizability, and accuracy of *ab initio* approaches [9,10], combining these with ML models has revealed numerous examples of accurate predictions at low cost, and further led to the discovery of materials at unconventional chemical compositions at an accelerated speed [11–15]. The other category is complex materials science problems in which properties and mechanisms originate from a large number of atoms. Examples of complex materials science problems include predicting the liquidus temperature of an alloy, the viscosity of a liquid [16], the plastic region of the stress-strain curve, the microstructure and resulting properties [17], and the glass forming ability [16,18,19] of an alloy. In this case, property-based features can no longer be calculated using *ab initio* approaches. As the data space is generally vast for complex problems and features are only approximated, a large number of training data is required for a sufficient representation and training of ML models.

A canonical example of a complex materials science problem is the prediction of the glass forming ability of an alloy. This ability is quantified in the critical cooling rate R_c, which is the minimum cooling rate required to avoid crystallization during solidification, resulting in the formation of a glass with an amorphous atomic structure [20]. A particular focus has been on bulk metallic glass (BMG) formation, which takes place in alloys with $R_c < 1000$ K/s [18,19,21]. Technologically exciting due to their superb properties

[*] Corresponding author.
E-mail address: jan.schroers@yale.edu (J. Schroers).

https://doi.org/10.1016/j.actamat.2022.118497
1359-6454/

[22,23] and unique processability [24–27], these BMGs can be vitrified into geometries exceeding at least one millimeter and often centimeters even in their smallest dimension because of their low critical cooling rate [18,21,28–30].

Motivated by their technological potential, significant research has been carried out to understand metallic glass formation [31–35] and to develop models, rules, and indicators [19,32,36,37] that guide the development of such alloys. Such models are based on thermodynamics, kinetics, rheology, and atomic/electronic structure. For example, based on the suppression of nucleation as a means to avoid crystallization, Turnbull proposed $T_{rg} = T_l/T_g$ (T_l: liquidus temperature and T_g: glass transition temperature) as an indicator for GFA [31]. Extensions to Turnbull's seminal work have been suggested to also consider additional aspects controlling GFA [32,34,38,39]. Further, theoretical concepts such as the "confusion principle" [35] have provided insights into mechanisms of GFA, revealing the number of constituting elements [19], atomic size variation [19,40,41], atomic packing density [36], atomic/short-to-medium range topological ordering [42,43] as important contributions to GFA. However, even though such approaches have significantly contributed to the understanding of metallic glass formation, their use in the discovery of BMGs has been limited. The main reason for this is that they rely on the knowledge of properties that are not *a priori* known, and their measurements are often as involved as determining the GFA through R_c directly. Such properties include viscosity, fragility, atomic packing, T_g, and structure and density of states of competing crystalline phases [44].

To accommodate for the limited predictive power of the above approaches, the slow sequential trial-and-error sample fabrication and characterization has been replaced by fast combinatorial synthesis paired with high-throughput characterization strategies [45–51]. However, even with these techniques, the potential composition space of BMG formation is by many orders of magnitude too large that a reasonable fraction can be determined [44].

An effective model that allows predicting BMGs would have to rely on *a priori* known properties. This approach has been pursued by data-driven ML strategies, pioneered by Wolverton et al. [52,53] where (i.) data on metallic glass formation were collected from the literature, (ii.) a large number of elemental features were considered that possibly affect GFA, (iii.) alloy features were derived by simple statistical functions, and (iv.) a random forest ML model was developed and evaluated by 10-fold cross-validation (CV). Despite the high accuracy achieved by the ML model, novel BMG alloys and new insights into glass formation have not been developed.

Surprised by the limited success of such ML strategies in developing novel scientific insights or materials for complex materials science problems, in this work we compared the previously developed ML model by Wolverton et al. [52] with (i.) a model we generate based on random and unphysical features, (ii.) a model only considering the chemical composition, and (iii.) a model where we consider human learning insights. For this purpose, we first reconstruct the previous ML model [52]. Specifically, we use literature data on GFA, and construct 201 alloy features through 6 simple statistical functions from 31 elemental features to build an ML model. A 10-fold CV test yields a similar high accuracy to the previously reported results [52]. To benchmark this model and its high accuracy determined through the CV test, we create another model where we choose random features that are unphysical. Surprisingly, we found that the unphysical model's 10-fold CV accuracy is as high as the previous model. In fact, even when leaving all features out and only considering the chemical composition information as input data to build an ML model, the same high accuracy is achieved. In other words, models with unphysical features or no features perform as well as the reconstructed ML model with 201 features. The only information model (i.) and (ii.) are built on is the chemical composition. Prediction through these models is based on the approach to predict a new BMG in the close chemical proximity to an existing BMG. This trivial knowledge of composition is sufficient for the commonly used 10-fold CV where "predictions are made" only by interpolation but not by extrapolation. Only if training and testing data are distinct and extrapolation instead of interpolation is required, differences in the 201-feature and unphysical models are revealed. However, all the models described above perform significantly worse than a simple, 3-(alloy)feature model based on physical insights.

2. Methods

The ML method for predicting the GFA can be broken down into four steps (Fig. 1). First, alloy data indicating chemical compositions and corresponding GFA are collected. In the second step, relevant features are determined. Here, elemental properties are first determined that need to be considered for the problem. From the elemental features, alloy features are constructed using either simple statistical functions (mean, range, standard deviation, etc.) or physics-based models, approximating the mixing process of the elements. Construction and even identification of features requires some degree of physical insights into the complex problem. Data and features are then used to build and train an ML model, for example, a random forest classification model. Finally, the ML model predicts BMGs in the unknown composition space. The details of these steps are described in detail in the following sections.

2.1. Data collection and processing

Our database is compiled by collecting all experimentally reported data from the Landolt-Bornstein Handbook on "Nonequilibrium Phase Diagrams of Ternary Amorphous Alloys" [54] and, additionally the peer-reviewed literature following the same approach as Wolverton et al. [52]. Here, we categorize alloys as BMG formers ($R_c < 10^3$ K/s), ribbon formers ($R_c < 10^6$ K/s), and non-ribbon formers ($R_c > 10^6$ K/s). All ribbon and non-ribbon data have been determined through melt-spinning experiments and taken from the Landolt-Bornstein Handbook. It should be noted that the ribbon data have been only tested at a cooling rate of 10^6 K/s, meaning the ability to form a bulk glass was not tested during the ribbon forming experiments. Therefore, the label "ribbon" does not necessarily indicate that the alloy cannot form BMG using bulk glass-preparation techniques.

In summary, the database contains 6816 unique alloy compositions with 1027 BMG, 4076 ribbon, and 1713 non-ribbon formers (Fig. 2). The database covers a wide range of chemistries, containing 55 different elements (Fig. 2a) and considered alloys ranging from binary to octonary alloys, with ternary as the majority of alloys (Fig. 2b). We also categorized all alloy data in subgroups based on the major binary element pair (Fig. 2c). One example of a specific subgroup is the ZrBe alloy family which is visualized here in a tree graph (Fig. 2d), showing example alloy compositions and their critical casting diameters D_c.

2.2. Feature identification and construction

The goal of feature identification and construction is to represent alloy compositions in a set of features that are relevant for the property of interest, which in our case is the GFA of the alloy. Features are essentially a set of quantitative and qualitative attributes that describe the alloy, which serves as the basis for the ML model. A set of features is expressed as a feature vector which has a one-to-one correspondence to the alloy composition and label (i.e., BMG, ribbon, or non-ribbon). To identify features,

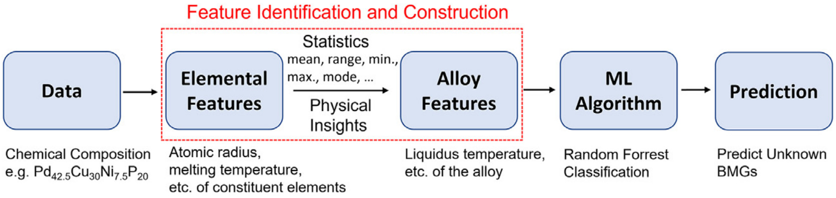
Feature Identification and Construction
Data
Elemental Features
Statistics
mean, range, min., max., mode, ...
Physical Insights
Alloy Features
ML Algorithm
Prediction
Chemical Composition e.g. Pd42.5Cu30Ni7.5P20
Atomic radius, melting temperature, etc. of constituent elements
Liquidus temperature, etc. of the alloy
Random Forrest Classification
Predict Unknown BMGs

Fig. 1. Process flow of the ML method for predicting the GFA of alloys. For the feature identification and construction step, we use either simple statistical functions of elemental features or physics-based models with physical insights into the elemental interactions.

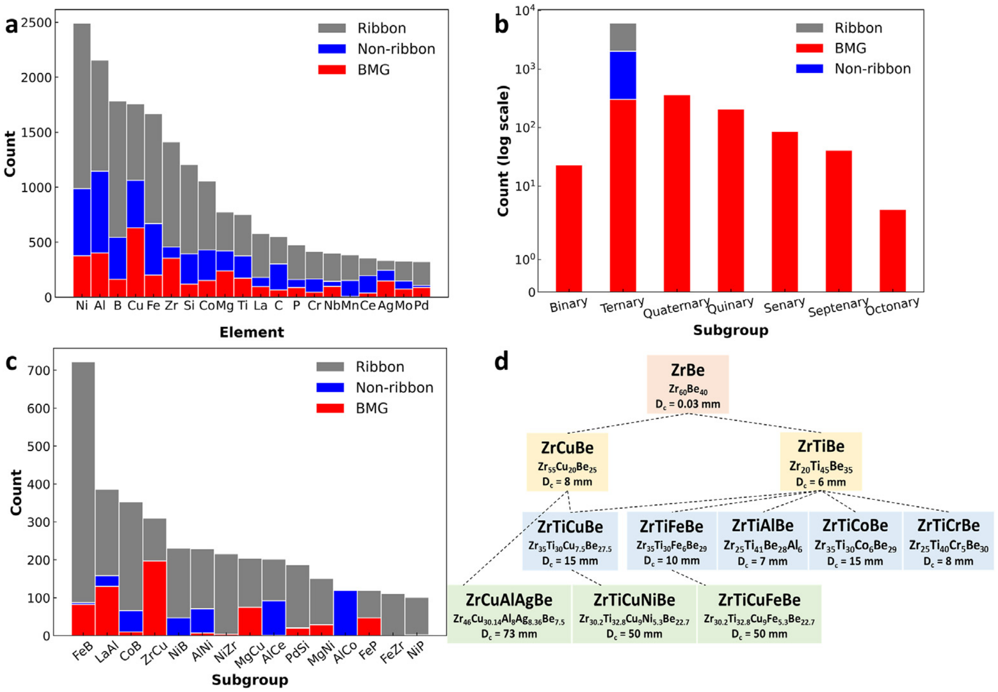

Fig. 2. Alloy database. (a) The counts of the 20 most frequent individual elements found in BMG, ribbon, and non-ribbon formers in the database. (b) The counts of alloy systems. (c) The counts of alloy subgroups (binary element pair based). (d) Tree graph of the ZrBe alloy family as an example of a specific alloy subgroup. D_c is the reported critical casting diameter of the alloy.

one can choose a strategy of considering essentially all possible material properties as features and use machine learning strategies such as feature selection to evaluate which features are important [55]. This strategy, in principle, does not require any understanding of the investigated problem. However, in reality, this strategy often overfits the data and results in poor performance (low accuracy) when applied to a new data set [56]. More effectively, one can use physical insights into the problem to determine features. To benchmark the predictive power of the models, we also construct features that are entirely unphysical and, in addition, develop models without features where we only use the chemical composition of alloys, i.e., the concentration of each respective element, as “features”. In summary, we test and pursue the extremes of features here where we use (i.) a large number of general-material features without any specific physical insight, (ii.) random, unphysical features, (iii.) no feature, only composition information, and (iv.) physical insight-based features from human learning. These features are described below in detail:

1. General-material features: We started with an expansive general-material feature set consisting of 201 features developed by Wolverton et al. [52]. These features originate from 31 elemental features (elemental properties defined for a constituent element). Six simple statistical functions are used to construct the alloy feature, including the minimum, maximum, and range of the values of the properties of each element present in the material, along with the fraction-weighted mean, mean absolute deviation, and mode (i.e., the property of the most prevalent element). Details on the features can be found in the supplementary materials.
2. Random, unphysical features: We randomly generated values for each element for five elemental properties with no physical meaning. We used the same six statistical functions described above to translate elemental features to alloy features.
3. No feature: We only use the composition information, i.e., the atomic percent (at.%) of the constituent elements, as input to the ML model.
4. Human learning features: It has been widely confirmed that BMG formers generally exhibit a) a composition close to deep eutectics b) an atomic size difference of > 12%, and c) a large negative heat of mixing among at least some constituent elements. These features reflect the state-of-the-art understanding of what characterizes a BMG forming alloy [19]. To represent these empirical rules by properties that are *a priori* known, we constructed three features:
   (1) Liquidus temperature reduction ΔT: To determine ΔT for an alloy we first separate the alloy in all binary combinations. For those binary combinations, liquidus temperatures are known. To construct the liquidus temperature of the alloy, we use the ratio of the binary combinations. We extrapolate the liquidus temperature of the alloy T_{alloy} by calculating from constituent binary pairs' liquidus temperatures,

e.g., T_{AB} at composition $A_{\frac{a}{a+b}}B_{\frac{b}{a+b}}$. For a ternary alloy $A_aB_bC_c$, T_{alloy} is calculated as:

$$T_{alloy} = \frac{(a+b) \times T_{AB} + (a+c) \times T_{AC} + (b+c) \times T_{BC}}{2 \times (a+b+c)}$$

To determine the liquidus temperature reduction, T_{alloy} is normalized by the mean liquidus temperature T_{mean} among the constituent elements, e.g., $T_A \times a + T_B \times b + T_C \times c$ for the ternary alloy $A_aB_bC_c$. ΔT is expressed as:

$$\Delta T = \frac{T_{alloy}}{T_{mean}}$$

(2) Atomic size difference δ:

$$\delta = 100\% \times \sqrt{\sum_i x_i(1 - r_i/\bar{r})},\ \bar{r} = \sum_i x_i r_i$$

where r_i is the atomic radius of the constituent element and x_i is the atomic fraction of the element.

(3) Maximum heat of mixing ΔH_{max}: We find the maximum (absolute value) binary mixing enthalpy $|\Delta H|$ among constituent binary pairs within the alloy. For this pair, we use ΔH multiplied by a factor as our feature. For example, for an alloy $A_aB_bC_c$, if $|\Delta H_{AB}|$ is the maximum binary pair value, ΔH_{max} is calculated as $\Delta H_{max} = \frac{2 \times a \times b}{a+b} \times \Delta H_{AB}$. ΔH_{AB} is obtained from the Miedema model [57]. The factor $\frac{2 \times a \times b}{a+b}$ considers the fractional number of A-B bonds in the alloy.

2.3. Machine learning algorithm

In this study we choose the random forest algorithm to build classification models that map different sets of features (described in Section 2.2) to the GFA of the alloys. Random forest is robust, easy-to-understand, and it handles high dimensional data well. In short, a random forest classification model constructs a multitude of decision trees at training (illustrated in detail in Fig. 3), and the output of the model is the label selected by most decision trees. Open-source python package Scikit-learn is used to build the random forest ML model. Hyper-parameter choices such as the number of decision trees, the number of features to choose at each tree node, or the maximum depth of each tree are optimized by grid search in the training process to achieve the best classification accuracy. The trained model can classify any new alloy (beyond the training data) into different categories of GFA and computes the relative likelihood for a new alloy to be in a certain category. Therefore, we can use the ML model to make predictions for the unknown composition space.

3. Results and discussion

3.1. Performance of machine learning models

We first evaluate and compare model performance using a 10-fold CV test. In a 10-fold CV, the data set is partitioned randomly into ten parts. Each time, 9/10 is used for training and the remaining 1/10 for testing. We average the test accuracies of the 10 models to get an average accuracy. Such test accuracy shows how accurately the predictive model will perform in practice when facing unseen data. When considering all data, including BMG ($R_c < 10^3$ K/s), ribbon ($R_c < 10^6$ K/s), and non-ribbon ($R_c > 10^6$ K/s) formers for modelling, we found that high accuracy of 89% (Fig. 4a) is reproduced for the ML model based on general-material features. Surprisingly, the ML model based on unphysical, artificial features results in an 87% accuracy, essentially indistinguishable from the general-material model in terms of accuracy. In fact, if we only consider the chemical composition information as input to build an ML model by using the atomic percent of each element as features, we get the same high accuracy of 89%. When only considering BMG and non-ribbon data for building ML models, which essentially reduces the problem to a binary classification problem, similar behaviors are observed from these models with model performance results summarized in Fig. 4b. This finding that the unphysical feature model and the model based only on composition results in the same accuracy than the general-material feature based model is surprising and requires further investigation. The only information the unphysical model and the composition model are built on is the chemical composition. Therefore, their predictions are based on the trivial approach that it predicts new BMG compositions in close vicinity of existing BMGs used in training. We also characterized the performance of the classification models using a receiver operating characteristic (ROC) curve, indicating no significant difference among four models (Fig. 4c). The confusion matrix of the various ML models is also analyzed and shown in the supplementary materials (Figs. S1–4).

The training and test data sets typically contain alloys from the same alloy system for the considered data. Therefore, the 10-fold CV test is limited to an interpolation of the training data rather than a true test into a significantly different composition space which would require extrapolation. To address this, Wolverton *et al.* previously proposed an extrapolation test called the “leave-binary-out” CV test. The authors systematically withheld data subsets containing each pair of elements (458 binary pairs in total) in the training data and use the withheld subset as a test [52]. For example, all alloys containing the ZrBe binary pair are withheld in the training data and are used as test data. The classification accuracy is then calculated for the test data, and the average accuracy across all binary pairs is used as the final measure. This “leave-binary-out” test is designed to evaluate the ML model’s ability to predict GFA for alloy systems different from the systems present in training, addressing the limitation of the 10-fold CV test. Following this argument, one should expect the unphysical and composition models to fail and result in low prediction accuracy as they only predict a BMG in the composition vicinity of an existing, used-in-training BMG. However, unphysical and composition models yield average classification accuracies of 77 and 75%, essentially indistinguishable from the accuracy of the general-material model of 76%. When only considering BMG and non-ribbon data for building ML models, similar results are obtained, revealing high accuracy above 91% for all four models. The model performance results are summarized in the supplementary Fig. S5. We argue that the average accuracy is not informative since for many of the binary pairs there are only a few alloys containing them in the database; thus, their “leave-binary-out” accuracies are less reliable and not indicative of the model’s ability. Instead, we argue here that one should only select to leave out more prominent binary systems and known BMG formers to reveal the differences among models. More crucially, the classification is discrete but not continuous (i.e., BMG or non-ribbon) since alloys are classified as BMG with a probability to form BMG > 0.5 given by the model. This classification is also insufficient. We argue that an effective model would predict known BMGs not just with probability $p > 0.5$ but with a very high probability, i.e., $p > 0.95$. To distinguish between these two criteria of $p > 0.5$ from $p > 0.95$, one would need to predict all possible alloys in the composition space and determine for each alloy the probability to form a BMG. For example, as we will show below, calculating for all potential ternary alloys (~2.6 million) $p > 0.5$ yields 0.5–1 million BMGs. This number is much too large to experimentally verify, even with high throughput methods, and it is also unreasonably large; it has been estimated previously that only

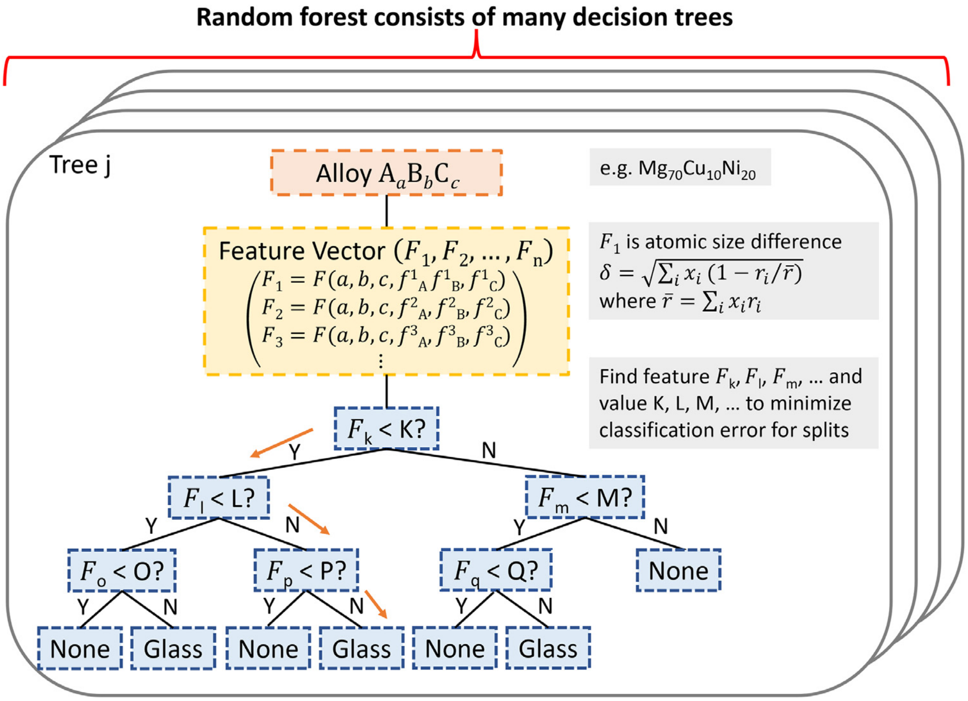
Random forest consists of many decision trees
Tree j
Alloy $A_aB_bC_c$
e.g. $Mg_{70}Cu_{10}Ni_{20}$
Feature Vector $(F_1, F_2, \dots, F_n)$
$F_1 = F(a,b,c,f^1_A f^1_B, f^1_C)$
$F_2 = F(a,b,c,f^2_A, f^2_B, f^2_C)$
$F_3 = F(a,b,c,f^3_A, f^3_B, f^3_C)$
F_1 is atomic size difference
$\delta = \sqrt{\sum_i x_i (1 - r_i/\bar{r})}$
where $\bar{r} = \sum_i x_i r_i$
Find feature F_k, F_l, F_m, ... and value K, L, M, ... to minimize classification error for splits
F_k < K?
Y
N
F_l < L?
F_m < M?
F_o < O?
F_p < P?
F_q < Q?
None
Glass

Fig. 3. Random Forest Classification. In our data, an alloy is represented by the feature vector in which each feature is constructed based on a specific function with composition and elemental properties as input. The random forest classification algorithm builds many decision trees (tree 1, 2, …, n), with tree j shown as an example here. For the constructing of each decision tree, at any node of the tree, the algorithm finds the feature F_k, F_l, $F_m, \dots$ (from a set of features) and the cutoff value K, L, M, … for splitting the data to minimize the classification error rate. The classification error rate is simply the percentage of training observations in a particular region that do not belong to the most common class, as we intend to classify each observation to the most prevalent class of training data in that region. When using the developed random forest model to make predictions, data goes through every decision tree in the forest to arrive at a label (one example decision path is indicated with orange arrows). The final predicted label is the label selected by most trees (i.e. majority voting).

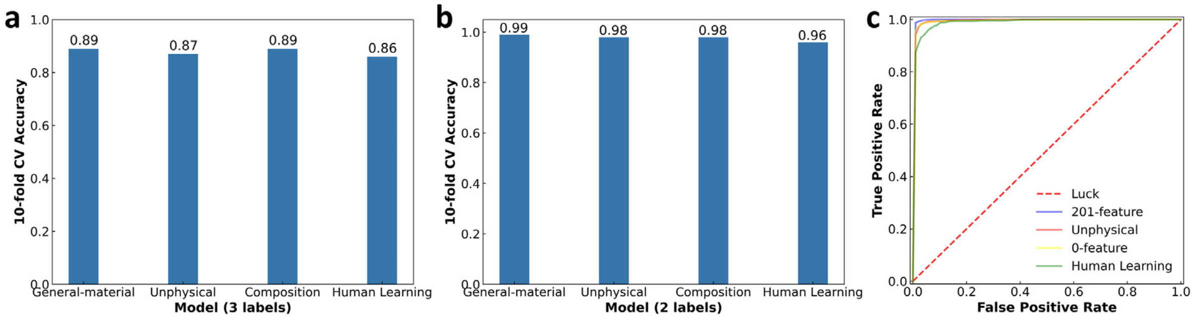

Fig. 4. ML model performance. (a) 10-fold CV test reveals high accuracy above 86% for all four models (general-material, unphysical, composition, and human learning) built on all data, including BMG ($R_c < 10^3$ K/s), ribbon ($R_c < 10^6$ K/s), and non-ribbon ($R_c > 10^6$ K/s) formers. (b) 10-fold CV test reveals high accuracy above 96% for all models built on only BMG and non-ribbon data. (c) Comparison of ROC curves among all four models built on BMG and non-ribbon data show the variations of true positive rates (TPR) and false positive rates (FPR) of the classifier as a function of the threshold at which an entry is labeled as "BMG", indicating that all the models exhibit similar performance in predicting glass formation. For all models their performance is much better than the random guess line connecting (0, 0) and (1,1) points as a baseline, but none of the models is significantly better than the others.

~1/10^6 alloys form BMGs [44]. Therefore, it is more reasonable to look at high probability predictions and test a model's predictive power for those predictions, i.e., $p > 0.95$.

3.2. Prediction of unknown metallic glasses

To test the ability of various ML models to predict BMGs that are novel and significantly different from known BMGs used in training, we use the trained model and predict into the entire composition space, which is spanned by ternary combinations of 24 practical elements. We then examine the predicted compositions and identify compositions with the highest likelihood of being a BMG, i.e., $p > 0.95$.

To construct the composition space that we consider for potential ternary BMG formation, we consider 24 practical elements (metals and metalloids, considering cost, reactivity, and toxicity): B, Mg, Al, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ge, Sr, Zr, Nb, Mo, In, Sn, Si, Ba, and Ta. These elements yield 2024 ternary combinations (24 choose 3). For each ternary system, we space alloys on a 2 at.% grid to reasonably consider the often rapid changes of GFA with composition [58]. This results in 1326 alloys per ternary and ~2.6 million alloys for all 2024 ternary systems from the here considered elements.

For all ~2.6 million alloys, we use the model to calculate the label and its probability (i.e., BMG, $p > 0.95$). To test the model's prediction performance, we leave some known BMGs out of the training set and subsequently determine if they have been pre-

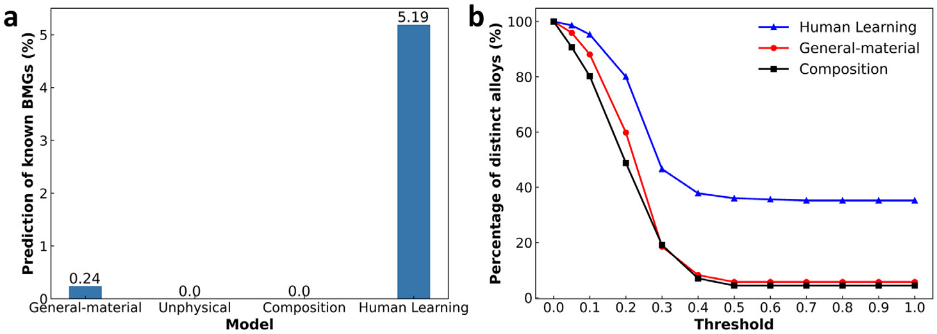

Fig. 5. ML model performance predicting unknown bulk metallic glasses. (a) Comparison of the predictive power of ML models (general-material, unphysical, composition, and human learning), quantified by the percentage of known BMG formers (the ones that contain any Cu, Pt, Ge, Sn, and Hf elements have been left out in training) that are among the top BMG predictions by the ML model with a probability > 0.95. (b) Comparison of the ability of ML models (general-material, composition, and human learning) to predict compositionally distinct alloys, quantified by the percentage of alloys that are compositionally distinct from known BMGs in training at varied at.% difference thresholds.

dicted as BMGs with high probability. We quantify the prediction performance by the percentage of known BMG formers (the ones that have been left out in training) that are among the predicted BMG formers with a $p > 0.95$. The BMG formers we leave out of the training data set are: (i.) those that contain any Cu, Pt, Ge, Sn, and Hf elements or (ii.) those that contain any ZrCu, ZrBe, FeB, and NiNb element pair, respectively. The selection of the elements chosen for (i.) is such that “similar’ elements are still in the training set whereas for (ii.) the pairs exhibit more unique behavior [59] and hence are more different from alloys in the training set.

When comparing the prediction performance of the four considered ML models with alloy features based on human learning, general-material, unphysical, and composition, by far the best prediction is achieved by the model using human learning-based features. Its predictive power, as quantified here, is more than twenty times higher than that of the general-material model, while the unphysical model and the model based solely on composition cannot predict the known BMGs that have been left out of the training set at all (Fig. 5a). The models based on unphysical features and composition do not have any predictive power in the unknown space as they only operate by predicting a BMG former in the close vicinity of a known BMG used in training. As the general-material model's performance is only insignificantly higher, we must also conclude that the general-material model does not exhibit predictive ability beyond the trivial prediction very close in composition to an already known and used-in-training BMG.

To further investigate the predictions made by the various models and how different they are from the training data, we use all BMG and non-ribbon data in training and investigate out of the top predictions (BMG, $p > 0.95$) how many alloys are compositionally distinct from known BMGs used as training data. To quantify compositional distinction between alloys in the training and in the prediction, we use the composition fraction difference, $A_aB_bC_c$ different by Δa, Δb, Δc, from $A_{a\pm\Delta a}B_{b\pm\Delta b}C_{c\pm\Delta c}$. To identify compositionally distinct alloys, we vary the threshold for Δa, Δb, Δc. As shown in Fig. 5b, the human learning model can predict significantly more alloys that are compositionally distinct from the training data than the general-material model and the composition model under any threshold. This finding further strengthens the above finding that the general-material model and the composition model only predict new BMG compositions in close vicinity of known BMGs used in the training data.

3.3. Discussion of machine learning approaches to predict BMG compositions

The presented comparisons of ML models are based on qualitatively and quantitatively very different features and feature constructions. The general-material model has the lowest requirement of physical insights when constructing alloy features, hence having the most hands-off ML characteristic. The human learning model is the extreme opposite. The considered three features had been identified as most indicative of bulk metallic glass formation through 50 years of research. To benchmark ML models for the prediction of BMGs, we constructed two models, one based on random features that are unphysical and another model that does not use features but only the information of the composition. These models have no predictive power and reveal no insights into glass formation motifs beyond the trivial prediction that in the close vicinity of a known BMG used in training, other BMGs are present. Even though BMG forming alloys can rapidly change their GFA with composition [58], one can usually, and this is what these models do, find a BMG in close vicinity of a known BMG. Hence, an ML model with performance similar to the unphysical and composition models will also not have true predictive power.

Using the data set described in Section 2.1, we found that in the interpolation test (10-fold CV) the performance of the general-material model and the human learning model, quantified by the 10-fold CV accuracy, is essentially identical to the unphysical and composition models. This observation suggests that the typically used 10-fold CV accuracy [52,60] is not a useful method to measure how effective an ML model is to predict BMGs. The fact that their accuracies are essentially identical suggests that interpolation within the same composition space where training has been carried out is only based on the trivial knowledge of composition. As in the 10-fold CV for the data set used here, there typically exists (statistically) an alloy in the 90% training data that is compositionally similar to every alloy the model predicts in the 10% test data. Hence, interpolation of all models is most effective when only using the composition information. The performance of ML models in the extrapolation mode is significantly different. Whereas the general-material model does not extrapolate better than the models (unphysical and composition) with no true extrapolative ability, the human learning-based model performs significantly better (Fig. 6).

These findings are highly surprising and demand a deeper discussion on why standard ML strategies requiring essentially no physical insights are limited in studying or predicting complex ma-

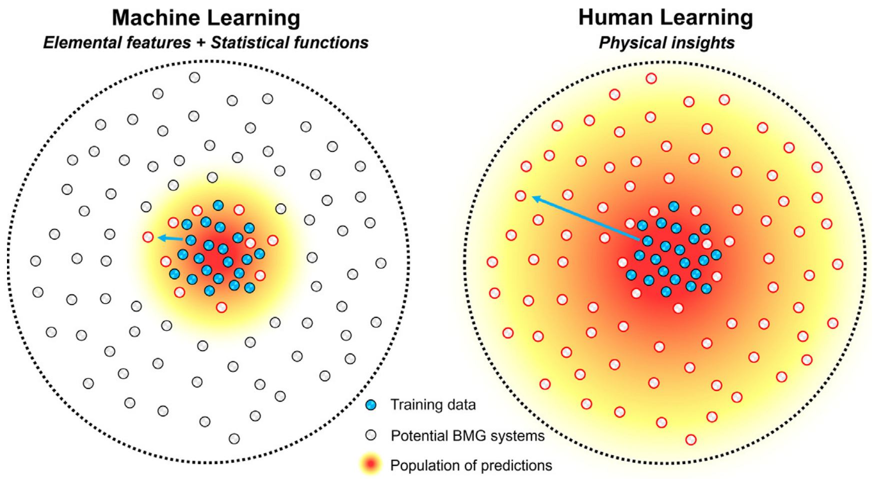

Fig. 6. Predictive power of machine learning versus human learning insights-based ML model for BMG discovery. The general-material ML model is based on elemental features and statistical functions while the human learning model is built on physical insights into GFA. The general-material model can only make accurate predictions in close vicinity of the training data versus that the human learning model is able to predict distinct BMG systems in the vast potential composition space. Number of known BMG systems in the training is ~10^2, potential BMG systems ~10^3, and multicomponent alloy systems ~10^5.

terials. It has been argued in the past that data set bias, continuous vs bimodal GFA classification, lack of reported failed or null results in the literature are among the challenges faced by ML models [61,62]. Even though not discussed in detail here, our study confirmed some of these points. However, we argue that the main challenge lies in the construction of a meaningful feature basis when describing alloys. Whereas the properties of individual elements are essentially all known, mixing is vastly richer and only for minute fractions determined. Instead, the field of materials science has been focusing and relying on physical models. Even though powerful to develop a conceptual understanding, such models can only be applied for idealized cases, are always an approximation, and are often limited to binary mixing. Hence, using elemental properties and combining them through statistical functions that do not consider the underlying physical mechanisms such as minimum, maximum, range, mean, mean absolute deviation, and mode cannot generally describe alloy properties accurately. To give just one example, when constructing the liquidus temperature feature T_L for the alloy $Au_{82}Si_{18}$, using the average values of the liquidus temperatures for Au and Si, which are the most reasonable among the statistical functions previously used, results in 1127°C compared to the actual value of $T_L(Au_{82}Si_{18}) = 364°C$. As the T_L of an alloy or more specifically the reduction of the T_L relative to the weighted average of the constitutive elements' T_L is a main contributor to bulk glass formation [31], it is not surprising that the general-material model lacks insights beyond the trivial knowledge of the data it uses.

It is important to mention that even in the human learning-based model, the features are only simplified and idealized approximations to the real mixing behavior in the alloy, and hence one cannot expect precise prediction. Further, as only idealized features are used, the model will make "expected" predictions of new BMG formers based on current understanding. "Unexpected" predictions beyond today's understanding of BMG formation are less likely to be made through the here constructed ML model based on human learning. As the composition space for alloys is vast and only a minute fraction has been considered where only a tiny fraction of potential BMG formers have been identified [44], identifying only the "expected" BMGs would already be a large success and advance BMG technology. To also identify "unexpected" BMGs beyond today's understanding of BMG formation, ML strategies are in principle capable but are unrealistic for addressing BMG formation or other complex materials science problems. This is because when considering the effectiveness of an ML approach, one also must take into account the quantity of experimental data that can be practically determined and compare this to the potential data space. The alloy space for multicomponent alloys is vast; for the here limited ternary consideration ~2.6×10^6 alloys, when considering up to quinary alloys ~10^{12} alloys [44]. Only when a sufficiently representative fraction of the potential data space can be experimentally realized and used in the training set, the ML algorithms can generate predictions with high accuracy, and even with features that poorly represent the alloy (like the alloy features used in the general-material model). However, determining such a representative fraction is a grand challenge; even when the state-of-the-art combinatorial synthesis paired with high-throughput characterization methods that can determine and characterize ~10^4-10^5 alloys per year [63] are considered, it would take over million years to determine 1% of the composition space of quinary alloys. A better strategy may be to (i.) carefully determine fewer but carefully selected alloys to represent the alloy and features space and determine for those alloys the GFA and (ii.) use models based on physical insights that describe the mixing behavior.

4. Conclusion

We built ML models to predict bulk metallic glass formation. Surprisingly we found that a general-material ML model with 201 alloy features constructed through simple statistical functions from 31 elemental features is indistinguishable from models that are unphysical or do not consider any features, when the prediction accuracy is tested in an interpolation manner. Only when significant separation of training and testing data is carried out, the general-material model performs better in this extrapolation mode than the unphysical or composition models, yet significantly worse than a human learning based 3-feature model. We explain the limited performance of the general-material model by the general inability to accurately represent alloy features through elemental features. As generally the potential data space is too large to determine a representative fraction, which would allow ML models to be effective even with poorly representative features, complex material science problems like bulk metallic glass formation require physical insights to develop effective and predictable ML models.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was financialy supported by the National Science Foundation through NSF-DMR (grant no. 2104316).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.actamat.2022.118497.

References

[1] F.A. Faber, A. Lindmaa, O.A. von Lilienfeld, R. Armiento, Machine learning energies of 2 million elpasolite $(AB{C}_{2}{D}_{6})$ crystals, Phys. Rev. Lett. 117 (13) (2016) 135502.

[2] W. Ye, C. Chen, Z. Wang, I.H. Chu, S.P. Ong, Deep neural networks for accurate predictions of crystal stability, Nat. Commun. 9 (1) (2018) 3800.

[3] T. Xie, J.C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Phys. Rev. Lett. 120 (14) (2018) 145301.

[4] C. Chen, W. Ye, Y. Zuo, C. Zheng, S.P. Ong, Graph networks as a universal machine learning framework for molecules and crystals, Chem. Mater. 31 (9) (2019) 3564–3572.

[5] C. Schattauer, M. Todorovic, K. Ghosh, P. Rinke, F. Libisch, Machine learning sparse tight-binding parameters for defects, NPJ Comput. Mater. 8 (1) (2022) 1–11.

[6] H. Levamaki, F. Tasnadi, D.G. Sangiovanni, L.J.S. Johnson, R. Armiento, I.A. Abrikosov, Predicting elastic properties of hard-coating alloys using ab-initio and machine learning methods, NPJ Comput. Mater. 8 (1) (2022) 1–10.

[7] D. Morgan, G. Ceder, S. Curtarolo, High-throughput and data mining with ab initio methods, Meas. Sci. Technol. 16 (1) (2005) 296–301.

[8] S. Kang, W. Jeong, C. Hong, S. Hwang, Y. Yoon, S. Han, Accelerated identification of equilibrium structures of multicomponent inorganic crystals using machine learning potentials, NPJ Comput. Mater. 8 (1) (2022) 1–10.

[9] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, M.A.L. Marques, Predicting the thermodynamic stability of solids combining density functional theory and machine learning, Chem. Mater. 29 (12) (2017) 5090–5103.

[10] W. Sun, S.T. Dacek, S.P. Ong, G. Hautier, A. Jain, W.D. Richards, A.C. Gamst, K.A. Persson, G. Ceder, The thermodynamic scale of inorganic crystalline metastability, Sci. Adv. 2 (11) (2016) e1600225.

[11] N.V. Orupattur, S.H. Mushrif, V. Prasad, Catalytic materials and chemistry development using a synergistic combination of machine learning and ab initio methods, Comput. Mater. Sci. 174 (2020) 109474.

[12] Y.P. He, E.D. Cubuk, M.D. Allendorf, E.J. Reed, Metallic metal-organic frameworks predicted by the combination of machine learning methods and ab initio calculations, J. Phys. Chem. Lett. 9 (16) (2018) 4562–4569.

[13] G. Hautier, C.C. Fischer, A. Jain, T. Mueller, G. Ceder, Finding nature's missing ternary oxide compounds using machine learning and density functional theory, Chem. Mater. 22 (12) (2010) 3762–3767.

[14] A.S. Rosen, S.M. Iyer, D. Ray, Z.P. Yao, A. Aspuru-Guzik, L. Gagliardi, J.M. Notestein, R.Q. Snurr, Machine learning the quantum-chemical properties of metal-organic frameworks for accelerated materials discovery, Matter 4 (5) (2021) 1578–1597 US.

[15] G.L.W. Hart, T. Mueller, C. Toher, S. Curtarolo, Machine learning for alloys, Nat. Rev. Mater. 6 (8) (2021) 730–755.

[16] S.A. Kube, S. Sohn, R. Ojeda-Mota, T. Evers, W. Polsky, N. Liu, K. Ryan, S. Rinehart, Y. Sun, J. Schroers, Compositional dependence of the fragility in metallic glass forming liquids, Nat. Commun. 13 (1) (2022) 3708.

[17] W.J. Boettinger, S.R. Coriell, A.L. Greer, A. Karma, W. Kurz, M. Rappaz, R. Trivedi, Solidification microstructures: Recent developments, future directions, Acta Mater. 48 (1) (2000) 43–70.

[18] J. Schroers, Bulk metallic glasses, Phys. Today 66 (2) (2013) 32.

[19] A. Inoue, Stabilization of metallic supercooled liquid and bulk amorphous alloys, Acta Mater. 48 (1) (2000) 279–306.

[20] K.F. Kelton, A.L. Greer, Nucleation in Condensed Matter, Pergamon Press, 2010.

[21] W.L. Johnson, Bulk glass-forming metallic alloys: Science and technology, MRS Bull. 24 (10) (1999) 42–56.

[22] W.H. Wang, C. Dong, C.H. Shek, Bulk metallic glasses, Mater. Sci. Eng. R 44 (2-3) (2004) 45–89.

[23] M.F. Ashby, A.L. Greer, Metallic glasses as structural materials, Scr. Mater. 54 (3) (2006) 321–326.

[24] J. Schroers, The superplastic forming of bulk metallic glasses, JOM 57 (5) (2005) 35–39 Us.

[25] A. Wiest, J.S. Harmon, M.D. Demetriou, R.D. Conner, W.L. Johnson, Injection molding metallic glass, Scr. Mat. 60 (3) (2009) 160–163.

[26] R.O. Mota, N. Liu, S.A. Kube, J. Chay, H. McClintock, J. Schroers, Overcoming geometric limitations in metallic glasses through stretch blowmolding, Appl. Mater. Today 19 (2020) 100567.

[27] W.L. Johnson, G. Kaltenboeck, M.D. Demetriou, J.P. Schramm, X. Liu, K. Samwer, C.P. Kim, D.C. Hofmann, Beating crystallization in glass-forming metals by millisecond heating and processing, Science 332 (6031) (2011) 828–833.

[28] A. Peker, W.L. Johnson, A highly processable metallic-glass: $Zr_{41.2}Ti_{13.8}Cu_{12.5}Ni_{10.0}Be_{22.5}$, Appl Phys Lett 63 (17) (1993) 2342–2344.

[29] V. Ponnambalam, S.J. Poon, G.J. Shiflet, Fe-based bulk metallic glasses with diameter thickness larger than one centimeter, J. Mater. Res. 19 (5) (2004) 1320–1323.

[30] A. Takeuchi, N. Chen, T. Wada, Y. Yokoyama, H. Kato, A. Inoue, J.W. Yeh, Pd20Pt20Cu20Ni20P20 high-entropy alloy as a bulk metallic glass in the centimeter, Intermetallics 19 (10) (2011) 1546–1554.

[31] D. Turnbull, Under what conditions can a glass be formed, Contemp. Phys. 10 (5) (1969) 473 -&.

[32] W.L. Johnson, J.H. Na, M.D. Demetriou, Quantifying the origin of metallic glass formation, Nat. Commun. 7 (2016) 1–7.

[33] H. Assadi, J. Schroers, Crystal nucleation in deeply undercooled melts of bulk metallic glass forming systems, Acta Mater. 50 (1) (2002) 89–100.

[34] A.L. Greer, New horizons for glass formation and stability, Nat. Mater. 14 (6) (2015) 542–546.

[35] A.L. Greer, Materials science - confusion by design, Nature 366 (6453) (1993) 303–304.

[36] D.B. Miracle, A structural model for metallic glasses, Nat. Mater. 3 (10) (2004) 697–702.

[37] M.X. Li, Y.T. Sun, C. Wang, S.S. Sohn, J. Schroers, W.H. Wang, Y.H. Liu, Data–driven discovery of a universal indicator for metallic glass forming ability, Nat. Mater. 21 (2) (2021) 165–172.

[38] Z.P. Lu, H. Tan, Y. Li, S.C. Ng, The correlation between reduced glass transition temperature and glass forming ability of bulk metallic glasses, Scr. Mater. 42 (7) (2000) 667–673.

[39] J. Orava, A.L. Greer, Fast and slow crystal growth kinetics in glass-forming melts, J. Chem. Phys. 140 (21) (2014) 214504.

[40] K. Zhang, W.W. Smith, M.L. Wang, Y.H. Liu, J. Schroers, M.D. Shattuck, C.S. O'Hern, Connection between the packing efficiency of binary hard spheres and the glass-forming ability of bulk metallic glasses, Phys. Rev. E 90 (3) (2014) 032311.

[41] K. Zhang, B. Dice, Y.H. Liu, J. Schroers, M.D. Shattuck, C.S. O'Hern, On the origin of multi-component bulk metallic glasses: atomic size mismatches and de-mixing, J. Chem. Phys. 143 (5) (2015) 054501.

[42] X.J. Liu, Y. Xu, X. Hui, Z.P. Lu, F. Li, G.L. Chen, J. Lu, C.T. Liu, Metallic liquids and glasses: atomic order and global packing, Phys. Rev. Lett. 105 (15) (2010) 155501.

[43] Z.W. Wu, M.Z. Li, W.H. Wang, K.X. Liu, Hidden topological order and its correlation with glass-forming ability in metallic glasses, Nat. Commun. 6 (2015) 1–7.

[44] Y.L. Li, S.F. Zhao, Y.H. Liu, P. Gong, J. Schroers, How many bulk metallic glasses are there? ACS Comb. Sci. 19 (11) (2017) 687–693.

[45] S. Ding, Y. Liu, Y. Li, Z. Liu, S. Sohn, F. Walker, J. Schroers, Combinatorial development of metallic glasses, Nat. Mater. 13 (2014) 494.

[46] P. Tsai, K.M. Flores, A combinatorial strategy for metallic glass design via laser deposition, Intermetallics 55 (2014) 162–166.

[47] N.J. Liu, T.X. Ma, C.Q. Liao, G.N. Liu, R.M.O. Mota, J.B. Liu, S. Sohn, S. Kube, S.F. Zhao, J.P. Singer, J. Schroers, Combinatorial measurement of critical cooling rates in aluminum-base metallic glass forming alloys, Sci. Rep. 11 (1) (2021) 1–9.

[48] F. Ren, L. Ward, T. Williams, K.J. Laws, C. Wolverton, J. Hattrick-Simpers, A. Mehta, Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments, Sci. Adv. 4 (4) (2018) eaaq1566.

[49] M.X. Li, S.F. Zhao, Z. Lu, A. Hirata, P. Wen, H.Y. Bai, M.W. Chen, J. Schroers, Y.H. Liu, W.H. Wang, High-temperature bulk metallic glasses developed by combinatorial methods, Nature 569 (7754) (2019) 99.

[50] J.M. Gregoire, P.J. McCluskey, D. Dale, S.Y. Ding, J. Schroers, J.J. Vlassak, Combining combinatorial nanocalorimetry and X-ray diffraction techniques to study the effects of composition and quench rate on Au-Cu-Si metallic glasses, Scr. Mater. 66 (3-4) (2012) 178–181.

[51] J.R. Hattrick-Simpers, J.M. Gregoire, A.G. Kusne, Perspective: composition-structure-property mapping in high-throughput experiments: turning data into knowledge, APL Mater. 4 (5) (2016) 053211.

[52] L. Ward, S.C. O'Keeffe, J. Stevick, G.R. Jelbert, M. Aykol, C. Wolverton, A machine learning approach for engineering bulk metallic glass alloys, Acta Mater. 159 (2018) 102–111.

[53] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, NPJ Comput. Mater. 2 (2016) 1–7.

[54] Y. Kawazoe, Nonequilibrium phase diagrams of termary amorphous alloys, LB: New Series Group III: Condensed, Springer 37 (1997) 1–295.

[55] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (1) (2014) 16–28.

[56] D.M. Hawkins, The problem of overfitting, J. Chem. Inf. Comput .Sci. 44 (1) (2004) 1–12.

[57] D.B. Miracle, D.V. Louzguine-Luzgin, L.V. Louzguina-Luzgina, A. Inoue, An assessment of binary metallic glasses: correlations between structure, glass forming ability and stability, Int. Mater. Rev. 55 (4) (2010) 218–256.

[58] J.H. Na, M.D. Demetriou, M. Floyd, A. Hoff, G.R. Garrett, W.L. Johnson, Compositional landscape for glass formation in metal alloys, Proc. Natl. Acad. Sci. U. S. A. 111 (25) (2014) 9031–9036.

[59] A. Takeuchi, A. Inoue, Classification of bulk metallic glasses by atomic size difference, heat of mixing and period of constituent elements and its application to characterization of the main alloying element, Mater. Trans. 46 (12) (2005) 2817–2829.

[60] J. Xiong, S.Q. Shi, T.Y. Zhang, Machine learning prediction of glass-forming ability in bulk metallic glasses, Comput. Mater. Sci. 192 (2021) 110362.

[61] D. Morgan, R. Jacobs, Opportunities and challenges for machine learning in materials science, Annu. Rev. Mater. Res. 50 (2020) 71–103.

[62] J. Schmidt, M.R.G. Marques, S. Botti, M.A.L. Marques, Recent advances and applications of machine learning in solid-state materials science, NPJ Comput. Mater. 5 (2019) 1–36.

[63] S.A. Kube, S. Sohn, D. Uhl, A. Datye, A. Mehta, J. Schroers, Phase selection motifs in high entropy alloys revealed through combinatorial methods: large atomic size difference favors BCC over FCC, Acta Mater. 166 (2019) 677–686.