nature communications

Article

https://doi.org/10.1038/s41467-022-35766-5

Toward the design of ultrahigh-entropy alloys via mining six million texts

Received: 13 April 2022

Accepted: 28 December 2022

Published online: 04 January 2023

Check for updates

Zongrui Pei[1,2,5], Junqi Yin[2,5], Peter K. Liaw[3] & Dierk Raabe[4]

It has long been a norm that researchers extract knowledge from literature to design materials. However, the avalanche of publications makes the norm challenging to follow. Text mining (TM) is efficient in extracting information from corpora. Still, it cannot discover materials not present in the corpora, hindering its broader applications in exploring novel materials, such as high-entropy alloys (HEAs). Here we introduce a concept of "context similarity" for selecting chemical elements for HEAs, based on TM models that analyze the abstracts of 6.4 million papers. The method captures the similarity of chemical elements in the context used by scientists. It overcomes the limitations of TM and identifies the Cantor and Senkov HEAs. We demonstrate its screening capability for six- and seven-component lightweight HEAs by finding nearly 500 promising alloys out of 2.6 million candidates. The method thus brings an approach to the development of ultrahigh-entropy alloys and multi-component materials.

Text mining (TM) is an artificial intelligence method to analyze and discover scientific knowledge in literature. It has been used in several fields, such as materials science[1–5], political science[6,7], public health[8–11], etc. TM has the potential for automatic materials discovery given sufficiently large corpora, such as for the material group of high- and medium-entropy alloys (HEAs, MEAs)[12–18], where more than 10,000 papers have been published[19]. Several TM methods have been suggested that build on corpora as training data[20]. One group of TM algorithms uses vectors to represent words, known as word-embedding algorithms[21–24]. Operations on the vectors provide meaningful information. For example, the difference between vector "FCC" and vector "Al" is approximately equal to that between vector "W" and vector "BCC", since the chemical element "Al" is commonly found with a face-centered-cubic (FCC) crystal structure and "W" with a body-centered-cubic (BCC) structure. These vectors are determined by maximizing the co-occurrence probability of an embedded word and its neighbors within the corpora. The cosine of two vectors measures the similarity of the words they represent. When increasing the frequency of the word "CoCrFeNiV" as the neighbor of "CoCrFeMnNi" by 10 times in a TM (skip-gram) model, its similarity ranking increases by 13 (Supplementary Fig. 1). TM models trained on specially selected corpora are predictive, as the presence of less relevant text items can reduce the relative frequency of keywords[1].

Here we have developed a highly optimized TM model for metallic materials focusing on HEAs. Unfortunately, TM methods can only identify targeted materials that are in principle already present in the corpora, a fact that does not per se include the discovery of materials. A key challenge in designing HEAs, however, is searching for similar elements with high mutual solubility. To this end, we propose a design concept of "context-similar elements" to overcome this limitation of existing TM methods in this field. The context-similar elements approach aims to capture the similarity of chemical elements in the alloy-design context used by scientists. The similarity in this context is not a metric calculable from simple elemental properties but a more comprehensive one that also reflects researchers' experience in materials research and design. This approach will enrich the portfolio of existing alloy-design methods and can accelerate the alloy discovery process by replacing the laborious literature search, review, and knowledge extraction with TM models. With this approach researchers with less domain-specific experience can design complex HEAs with many components assisted by TM models that not only "read" huge

[1]New York University, New York, NY 10012, USA. [2]Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. [3]University of Tennessee, Knoxville, TN 37996, USA. [4]Max-Planck-Institut für Eisenforschung, Düsseldorf 40237, Germany. [5]These authors contributed equally: Zongrui Pei and Junqi Yin.
e-mail: peizongrui@gmail.com

amounts of publications but also ”analyze” them more context-sensitive.

Results

Figure 1a shows a schematic for the machine-learning model. We adopt the skip-gram algorithm for our model since the algorithm provides a good compromise between efficiency and accuracy[1,21,22]. It has a neural-network structure with only one hidden layer. Words in the training corpora are firstly encoded into one-hot vectors w_i. This means that only one component of each vector assumes a value of “1", which records the word’s location in the whole vocabulary, and the remaining components are “0". These vectors are fed into the neural network as training data. We feed 6.4 million materials-related abstracts plus abstracts on metallic materials into the machine-learning model [see Supplementary Note and Methods]. Here we do not take the weight of the abstracts for metallic materials as a tunable parameter. If we increase its weight, there will not be a convergence trend to test the predictions. Increasing the weight will eventually result in a model without any benefit from other scientific papers. Instead, we tuned different hyperparameters to check the reliability of our models. Constructing a model specifically custom-designed for metallic alloys would be ideally realized by feeding only abstracts of papers that deal with metallic materials. However, papers related to metallic materials represent only a small portion of all scientific papers and thus provide only an insufficient basis for such data-hungry language models. To overcome this problem, our model adopts the commonly used transfer-learning method. We feed all available texts into the skip-gram model, equivalent to constructing a general model, and then feed only texts related to metallic materials, equal to slightly tuning the model for metallic materials. Accurate extraction of the named entities is essential[25], and particular attention is needed for HEAs. Researchers use different orders of constituents for the same alloy that can be mistaken as different ones. For example, the Cantor alloy, a CoCrFeNiMn compound, can be written in 5! = 120 different name variants by simply rearranging the sequence of the elements, such as CrCoFeMnNi, CoCrFeMnNi, etc. Also, one alloy can be written in various formats like Co-Cr-Fe-Ni-Mn, Co20Cr20Fe20Ni20Mn20, etc. This problem needs particular attention here due to its critical impact on identifying novel massive solid solutions. We alphabetize the chemical elements of one alloy in our skip-gram model [Fig. 1] and knowledge-graph model [Supplementary Fig. 3]. Once a model is trained, the corpora information is encoded into the matrix M in the hidden layer. As a simple demonstration of our algorithm, the word vectors v_i for two words “Fe" and “Ni" are used to calculate their respective cosine similarity, i.e., $S_{ij} = \cos(\boldsymbol{v}_i, \boldsymbol{v}_j) = \boldsymbol{v}_i \cdot \boldsymbol{v}_j / |\boldsymbol{v}_i||\boldsymbol{v}_j|$. Similar words are ranked and presented in Fig. 1b, c. According to the cosine similarity, the four words that are most similar to Fe are the chemical elements Mn, Co, Ni, and Cr. Likewise, the top four similar words of Ni are Co, Cu, Sn, and Mo. This trend analysis shows that the model can well capture the similarity of words in the context of the chemistry of solid solutions.

Chemical elements with context similarity

The distribution of elements is visualized in Fig. 2a, using the color map defined in Fig. 2b. Elements with similar chemical features are grouped together. The context-similar version of the Periodic Table of the Elements (PTE) includes rich information about how they were used in the enormous amount of literature. Suppose two chemical elements appear in a similar context. They are close in the vector space but not necessarily in Mendeleev’s PTE. For example, Al is not the neighbor of Mn and Cr in Mendeleev’s PTE. However, they are neighbors in the latent space of our word-embedding model. Mendeleev’s PTE has its physical origin in quantum mechanics, while our machine-learning model is based on the appearance of the elements in various research contexts. We take Al and Mn as an example pair to show how they can be grouped in the two representation forms (in the latent space of our machine-learning model and Mendeleev’s PTE) [Fig. 2c]. Elements Al and Mn share many similar neighboring words, such as Fe and Mg, since they are often used as solute atoms in steels and

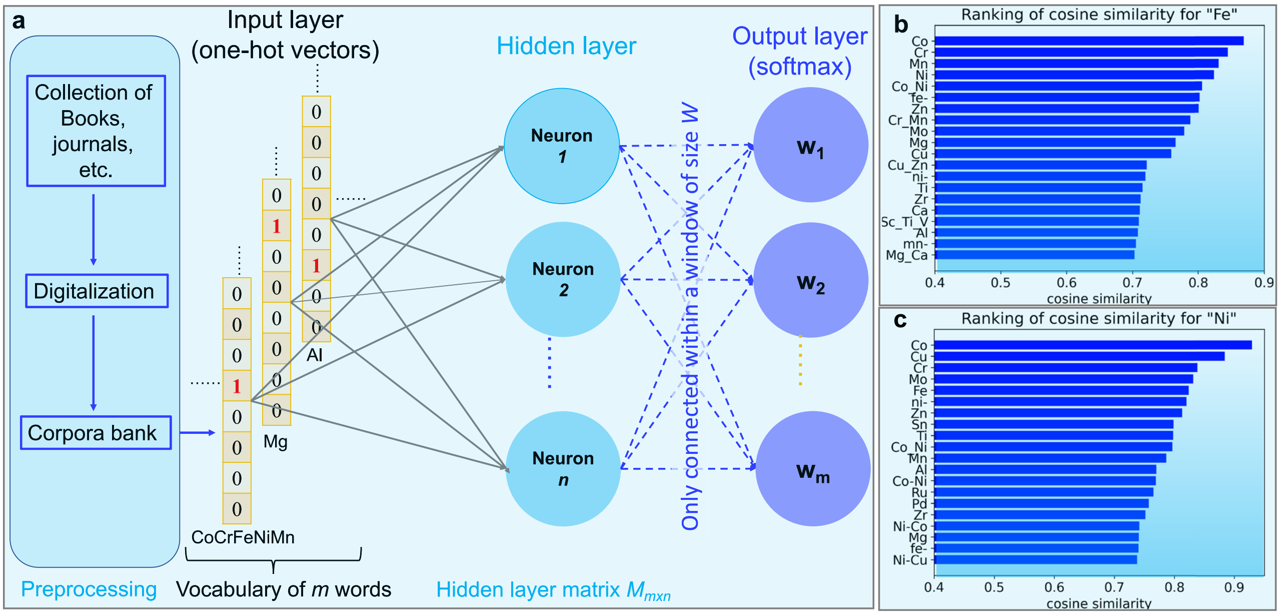

Fig. 1 | Schematic for the word-embedding model (skip-gram). It has a neural-network structure but with only one hidden layer between the input and output layers[21,22] (**a**). The training data fed into the model are the processed corpora downloaded from an online database[41]. The corpora are first tokenized into separate words or phrases (combinations of two or more words with unique meanings) and then translated into vectors. In the one-hot representative of a word vector, each word is represented by a sparse vector with only one nonzero element. The word vectors are connected to all neurons in the hidden layer; the latter is also fully connected with the output layer which represents the appearance probabilities of words in their context. For a given window size of the words that define their context, the skip-gram algorithm maximizes the probability of the word that appeared in that context. Once the neural network is optimized, the key information is stored in the hidden layer. As examples of its application, similar words of “Fe" and “Ni" are shown in **b**, **c**, respectively.

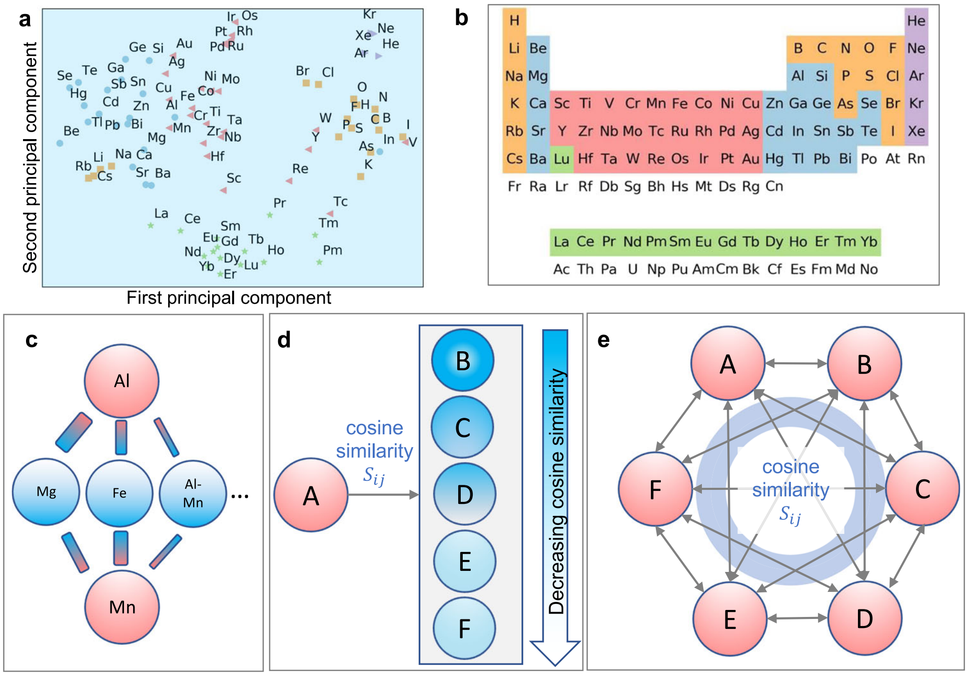
a
Second principal component
First principal component
Ir Os
Pt Rh
Pd Ru
Kr
Xe Ne
Ar He
Ge Si Au
Ag
Se Te Ga
Hg Sb Sn Cu
Cd Zn Al
Ni Mo
Fe Co
Cr Ti Ta
Br Cl
O
H N
C B
Be Tl Pb Bi
Mg Mn Zr Nb
W P S
Y I
In V
As
K
Rb Li Na Ca
Cs Sr Ba
Hf
Sc
Re
La Ce Sm Pr Tc
Eu Gd Tb Tm
Nd Dy Ho Pm
Yb Lu
Er
b
H He
Li Be B C N O F Ne
Na Mg Al Si P S Cl Ar
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te I Xe
Cs Ba Lu Hf Ta W Re Os Ir Pt Au Hg Tl Pb Bi Po At Rn
Fr Ra Lr Rf Db Sg Bh Hs Mt Ds Rg Cn
La Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb
Ac Th Pa U Np Pu Am Cm Bk Cf Es Fm Md No
c
Al
Mg
Fe
Al-Mn
...
Mn
d
A
cosine similarity S_{ij}
B
C
D
E
F
Decreasing cosine similarity
e
A
B
C
D
E
F
cosine similarity S_{ij}

Fig. 2 | The context-similar elements and their applications to design high-entropy alloys. a Chemical elements in the latent space by the principal component analysis (PCA) based on their word vectors. The horizontal and vertical axes are the first two components of PCA. The elements are colored according to the scheme shown in the Periodic Table of Elements (**b**). **c** We explain the reasons for the difference between Mendeleev's PTE and the chemical elements in the latent space by taking elements Al and Mn as examples. The thicknesses of the sticks represent the connection strengths of words (similarity). We propose two methods to design high-entropy alloys. These methods are illustrated in **d**, **e**, taking six-component alloys as an example. The first method starts with one element that must be included and select its four most similar elements according to the cosine similarity. The second method considers all participating elements equally. The cosine similarity S_{ij} of any two elements are averaged to measure its potency as a candidate.

magnesium alloys, albeit with different prevalence. This fact increases the co-occurrence probability that puts them close in the latent space. Therefore, the context similarity of elements reflects better how researchers used them to synthesize actual materials, which is precisely the experience needed to design alloys.

Design of alloys based on context similarity

Exploration of the vast space for HEA design challenges traditional alloy-design strategies[19] and requires an intelligent choice of elements, particularly when targeting solid solutions. It is important to overcome this challenge, driven by the need for high-performance materials, for which HEAs with multiple components are promising candidates. In this regard, the concept of publication-based context-similar elements provides unexplored opportunities. We propose two different methods to use word vectors in the HEA design [see Fig. 2d, e]. One is to start with a specific element A, which is for some reason preferred to be included as a chemical component, and identify the most similar elements according to the cosine similarity $\{S_{ij}\}$. For example, for a Fe-including quaternary HEA, the top three candidates are Mn, Co, and Ni according to S_{ij} [Fig. 1b]. Instead, if our target is a HEA with Ni, the top three candidates are Co, Cu, and Sn. Sn is not close to Ni in the PTE but often appears in Ni-Sn alloys, reflecting the difference between our text similarity and the PTE again. In a second method, we first select M promising elements $\{E_i\}_{i=1}^{M}$ and consider their cosine similarity equally [Fig. 2e]. For each N-component HEA, we average over the similarity of each element pair S_{ij}, i.e.,

$$\bar{S} = \frac{1}{N(N-1)/2}\sum_{i=0}^{N-1}\sum_{j=i+1}^{N-1}\cos(\boldsymbol{v}_i,\boldsymbol{v}_j). \qquad (1)$$

Here $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are the word vectors of elements i and j in the N-component HEA. We rank the alloy candidates according to $\bar{S}$ and pick the top candidates as the most promising ones. This method treats all elements equally and thus is consistent with the alloy-design spirit of multi-principal HEAs.

Body-centered-cubic high-entropy alloys

Exploration of refractory BCC HEAs with beneficial properties attracts high attention currently[26–29]. We demonstrate the approach of the "context similar" elements in Fig. 3 by limiting our candidate elements to common transition-metal elements. As an example and to demonstrate the predictive strength of our method, the newly defined $\bar{S}$ parameter in Eq. (1) is adopted to design five-component HEAs. We focus on seven elements, which yield 21 combinations (alloys). Ti, Zr, Nb, Mo, Hf, Ta, and W are among the most common transition metals, and some of them appear in the Senkov alloy TiZrNbHfTa[28]. In order to critically test the predictive power of our approach and understand the trend of HEAs, we train individual models for the representative years. Each model adopts training data of the publications that appeared only in its corresponding year. The only exception is the model of 2003 that adopts all abstracts up to that year. The idea behind that is to test if our modified TM model could have predicted an alloy with only the knowledge that was available up to the year 2003 (i.e., the Senkov alloy had not yet been discovered, but it was found only 8 years later). The results predicted by the models are shown in Fig. 3b. Among the 21 candidates, the Senkov alloy is continually ranked among the top three materials. In 2011 when the alloy was finally discovered and synthesized, and more recently 2016 and 2018, the Senkov alloy is at the absolute top of the list. This means that our context-sensitive TM model would have suggested this specific HEA at least eight years before conventional alloy-design approaches found it.

In order to explore if any physical meaning is connected to our $\bar{S}$ parameter, we adopt our simple, approximate method to predict solid solutions and calculate the so-called γ parameter[30]. Simply put, the parameter is a ratio between the approximate Gibbs free energy (usually a negative number) of the HEA G_N and that of its binary

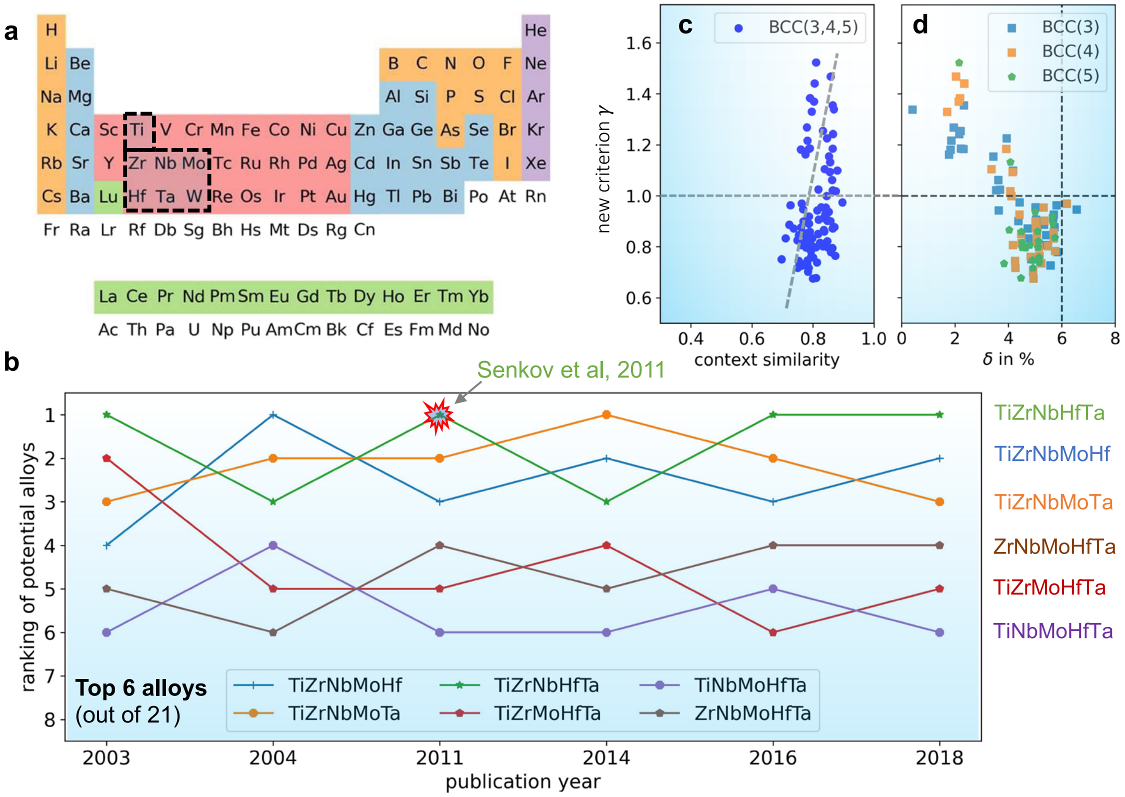

Fig. 3 | The context-similar elements and their applications for BCC high-entropy alloys. a Taking the transition elements (Ti, Zr, Nb, Mo, Hf, Ta, and W) as an example, we show that the tendency to form solid solutions is correlated to the newly defined context similarity. For better visualization, we only consider here three-, four-, and five-component equiatomic alloys. **b** The five-component alloys are ranked by their context similarity for different publication years. The top three alloys include the Senkov alloy of TiZrNbHfTa, TiZrNbMoHf, and TiZrMoHfTa. The Senkov alloy was proposed in 2011 by Senkov et al.[48]. Other promising ones include ZrNbMoHfTa, TiNbMoHfTa, TiZrNbMoTa. **c** The context similarity $\bar{S}$ is strongly correlated with the thermodynamics-based γ parameter. **d** More details on the γ parameter when using it along with the lattice distortion δ to predict the solid solutions.

subsystems G_2. If an alloy has a γ-value larger than 1, it is likely to be a solid solution rather than a multi-phase alloy. Furthermore, the parameter is found to be linearly and positively correlated with $\bar{S}$ [Fig. 3c]. This feature indicates that a larger value of $\bar{S}$ offers more promising candidates as solid solutions. The prediction of solid solutions using the γ parameter can be improved with the lattice-misfit parameter δ as an additional physics-based descriptor[30], which measures the lattice distortion due to the different atomic sizes [see Fig. 3d]. The use of these two additional descriptors reveals that only a tiny fraction of the ternary, quaternary, and quinternary alloys are solid solutions with the highest similarity scores. The majority of them are multi-phase alloys.

Face-centered-cubic high-entropy alloys

Similar to the BCC HEAs, the averaged context similarity $\bar{S}$ is calculated for a group of FCC HEAs and shown in Supplementary Figure 2. Again, we limit the constitutional elements to the transition-metal elements from V to Cu of the third group. Taking the five-component alloys as an example, we show the similarity $\bar{S}$ for individual years in Supplementary Fig. 2a. This test protocol shows that the concept effectively identifies HEAs long before they were found by conventional alloy-design methods. The Cantor alloy was first reported in 2004, but it was ranked as the second most promising solid-solution HEA by our method already before 2004. The seminal paper of Cantor et al. did not receive much attention immediately after its publication, but its impact has increased exponentially since the last decade[19]. This trend is correctly reflected by its ranking in Supplementary Fig. 2a. The second and third most promising HEAs are MnFeCoNiCu[31] and CrFeCoNiCu[13], which were also synthesized. We also calculate their tendency to form solid solutions by using the γ parameter[30]. As presented in Supplementary Fig. 2b, the two quantities are linearly correlated, similar to the case of the BCC HEAs. This trend further confirms the significance of the $\bar{S}$ parameter in screening for high-entropy solid solutions.

Combination with Integrated Computational Materials Engineering (ICME) methods to design ultrahigh-entropy alloys

The method of "context similarity" picks element candidates for HEAs, which is the first step for designing high-entropy solid solutions. Then, various procedures can be developed for further screening, refining, and filtering the results, assisted by the methods grouped under the umbrella of ICME (integrated computational materials engineering)[32,33] and included in the materials genome initiative[34]. ICME is an approach for designing materials and microstructures using mean-field thermodynamics and kinetics tools as well as ab-initio and structure-property simulation methods. A few examples are provided below to show how to integrate the context-similarity method with ICME to accelerate the design process.

In the first example, we screen for alloys based on their mechanical properties. Critical mechanical properties include for instance creep, ductility, and yield stress. Here we focus on one important mechanism behind these features, i.e., solid-solution strengthening σ_y, as it provides an essential contribution to the yield stress[35,36]. We adopt a model developed recently by Varvenne et al.[35]. The full details of the model are available in the reference and supplementary material. Supplementary Fig. 2d shows the solid solution strengthening predicted at 300 K for the top six FCC alloy candidates. The CrFeCoNiCu alloy has the largest strengthening effect σ_y of ~290 MPa, followed by CrMnFeCoCu with a comparable value. The strengthening effect of about 146 MPa for the Cantor alloy is fairly consistent with the experimental measurement of 125 MPa[35,37]. More validations of the methods can be found elsewhere[35,38]. These results show that alloys with optimal mechanical properties can be designed jointly with the TM-based $\bar{S}$ parameter.

In the second example, we aim to use our method and go beyond the established high entropy systems and design six- and seven-component, lightweight, single-phase equiatomic HEAs. This example has been partly motivated by Cantor, who calls for a bolder design of HEAs, also considering materials beyond five components[19]. To tackle this challenge, we design a workflow comprised of multiple steps and elaborate on the application and its statistical aspects in Fig. 4a, b. The screening consists of three steps with three adjustable criteria, i.e., (i) context similarity $\bar{S} > 0.6$; (ii) thermodynamics-based solid solution parameter $\gamma > 1$, and (iii) the alloy's mass density $\rho < 7.8\,\text{g/cm}^3$ (density of iron). In the current example, these three criteria have been selected for the sake of demonstration. The high-throughput screening has been limited to 30 transition-metal elements [see supplementary text for the list]. The total number of alloy candidates before screening has been 2.6 million. After each step, we excluded sets with 1–2 orders of magnitude of alloys. Eventually, only 494 HEAs remained and are promising for synthesis in experiments.

More specifically, the top-ranked alloys along these three categories are TiCrFeCoNiMo (six-component), TiCrFeCoNiCuZn (seven-component), and TiFeCoNiCuZn (six-component) according to the context similarity $\bar{S}$; six-component VCrMnFeCoNi, VCrMnFeCoCu, and VCrMnFeNiCu following the thermodynamics-based parameter γ; and six-component ScTiZnZrAgCd, TiVCrMnFeZn, and TiVCrMnCuZn regarding the lowest mass density. These results demonstrate concrete design suggestions identified directly and autonomously via our TM modeling approach from the existing wealth of published literature. The list of the 494 identified alloys can be found in the supplementary data.

Searching for existing HEAs to avoid redesigning alloys

One of the crucial tasks for materials designers is to check if the targeted alloys have been already proposed and synthesized before. This task becomes increasingly important in the ever-growing information avalanche. When studying HEAs, an additional challenge arises from their non-standardized naming system. Therefore, there is an urgent need to build TM models in which all HEAs are standardized with unique names, as in chemistry. It would then be much easier to check if an alloy has already been synthesized before or not. The knowledge graph (KG) approach has demonstrated its usefulness in quickly retrieving the required information[2]. KG is a graph-structured data model that links entities such as alloys and their properties through various relations (i.e., edge words) [see Supplementary Fig. 3a]. The Google knowledge graph is for general-purpose applications, yet, it is not specialized for extracting useful information from research corpora[39]. So, for example, it cannot tell if CoCrFeMnNi and NiMnFeCrCo are the same materials or not. Here we propose a KG for alloys, focusing on HEAs, i.e., an alloyKG, as an acronym for the knowledge graph for alloys. Their constituent elements are ordered alphabetically according to their symbols. We connect the HEAs with the DOI's (Digital Object Identifiers) of the papers in which they appeared. Essential authors in the field and their specific contributions can also be identified for further processing. In Supplementary Fig. 3b, we show the results for an exemplary search using our alloyKG approach with the keyword "CoCrFeMnNi" and the edge phrase "mentioned by". Since the naming system is standardized in alloyKG, every arrangement of the 120 possibilities for a five-component alloy gives the same results. The retrieval yields the papers (represented by their DOI's) that mentioned the alloys.

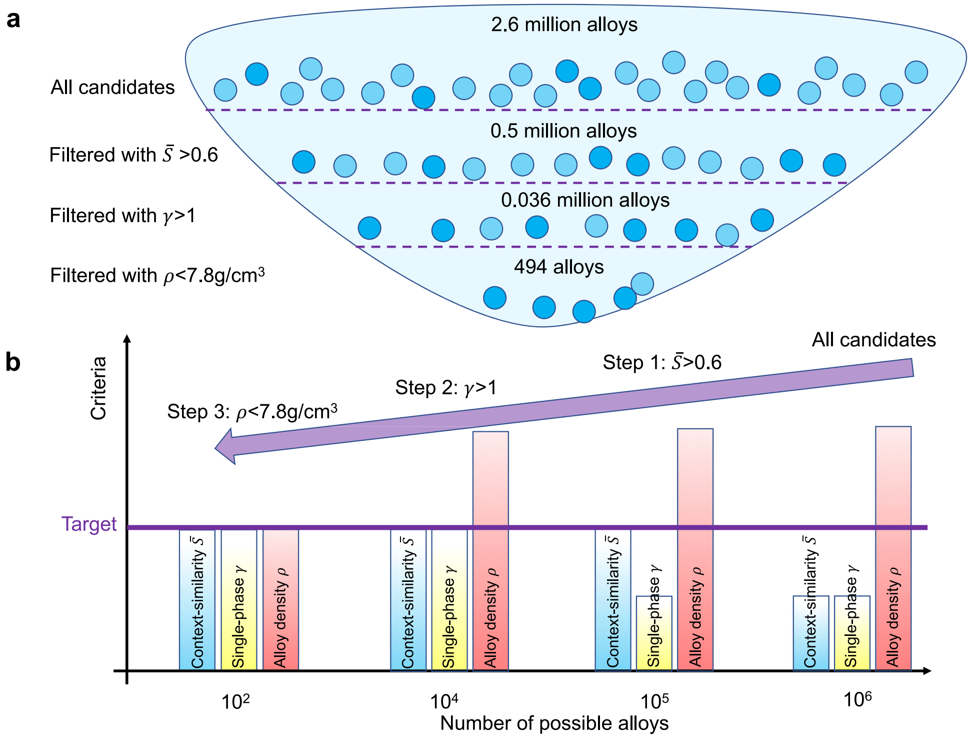

Fig. 4 | An exemplary design of lightweight high-entropy alloys with more than five components. a We show the example study conducted for six- and seven-component single-phase, equiatomic alloys, limited to 30 transition-metal elements. The first step is to calculate the average context similarity $\bar{S}$ and select alloys with a $\bar{S} > 0.6$. The distributions of $\bar{S}$ for all these candidates are shown in Supplementary Fig. 5. The second filtering step is to find in this subset those alloys that are likely in the form of solid solutions assisted by the thermodynamics-based rule $\gamma > 1$[30]. The third step is to select the alloys with a density smaller than iron, i.e., $\rho < 7.8\,\text{g/cm}^3$. Here we picked all alloys with a mass density below that of iron. **b** Along this filtering workflow, the number of possible alloys decreases from 10^6 to 10^2. Alloys in the shortlist are candidate materials for synthesis and testing.

Discussion

One of the most important directions for the HEA community is to explore the vast compositional space of HEAs with more components, such as six or seven components, and not limit the search to five or fewer components. Irrespective of the success of traditional TM methods, one of their principal shortcomings is that they cannot readily design alloys that do not appear in the corpora. This challenge has been overcome by our current "context similarity element" concept. One unique and characteristic feature of HEAs is that they are (ideally) solid solutions, which means we can screen for similar element candidates and ignore their specific concentrations first. This strategy is one of the main reasons why the concept works specifically well for the current alloy-design task. The transferability of our method has been carefully tested for a wide range of alloys from medium, high, and even ultrahigh-entropy alloys, with both BCC and FCC crystal structures. The method is applicable to the design of different HEAs, and there is no specific additional need to fine-tune the model. Another benefit of the approach is that the selected alloy groups can be further refined, after pre-screening promising composition spaces by our TM approach, by using further filtering criteria from the established ICME toolbox„ such as thermodynamic, kinetic, structural, and/or property databases and simulations. This hybrid alloy-design concept, combining TM-based pre-screening of the practically infinite chemical composition space and subsequent physics-based filtering, paves a pathway towards a closed-loop materials design approach that is characterized by the following specific steps: (i) fully automatically reading and autonomously analyzing millions of papers, (ii) searching for a specific set of chemical elements and suited alloy ingredients, (iii) proposing alloy candidates, (iv) calculating properties of the alloys, (iv) selecting alloys based on the targeted properties, and (v) identifying and excluding alloys that were already synthesized and casting all results into a recommendation list. With this approach, even inexperienced users with less domain knowledge in the field of alloy design can develop complex materials with many components assisted by TM models and a huge body of scientific publication corpora.

The word vectors of the chemical elements can reflect the rise of specific HEAs, quantified by changes in the cosine similarity S_{ij} [Supplementary Fig. 4]. The Cantor alloy of CoCrFeMnNi was proposed in 2004[14]. Prior to that year, the most similar elements to Fe, according to our similarity index, are Cr, Mn, Mg, and Al. Given the increasing relevance of the Cantor alloy and its subsystems, the top four most similar elements to Fe, when extracted from context mining, are only the elements in the Cantor alloy, at least since 2014. In 2014, several milestone papers appeared about the Cantor alloy. For example, it was found to have good ductility and toughness even at cryogenic temperatures[40]. In a different study of the same year, several compositional subsystems of the Cantor alloy were explored[37]. Stable FCC systems were identified, including the equiatomic CrCoNi solid solution. These investigations and many others pushed Mn, Co, Ni, and Cr into the top similarity list for Fe. We trained different models to test the stability of our method. When the window size is changed from 8 to 10, and the dimension (number of neurons in the hidden layer) from 200 to 300, the most similar words of "alloy_HEA" remain almost the same [see Supplementary Fig. 6], but the training time increases significantly. We also applied the models to calculate the context similarity, taking BCC HEAs as an example [see Supplementary Fig. 7]. The representative Senkov alloy TiZrNbHfTa is ranked number one by these models. All these HEAs follow the same order except for the TiZrNbMoHf and TiZrNbMoTa to switch their positions. This trend again shows a model with 200-dimensional word vectors and a window size of 8 is sufficient for designing alloys.

In summary, we proposed a concept for the systematic and automatic search for "context similarity elements" and demonstrated its successful application for the design of high-component high-entropy alloys. The method overcomes the common problem of traditional text mining methods that can only explore existing materials and enables us to design alloys that do not appear in the training corpora. As a demonstration, we show that the approach would have successfully identified the representative FCC Cantor and BCC Senkov alloys as the most promising high-entropy alloys, long before they had been actually discovered and synthesized. We also find that the context similarity is strongly correlated to a thermodynamics-based rule proposed by us in a previous study[30]. This trend indicates that this thermodynamic alloy-design parameter adequately captures the tendency of solid solution formation. We also show that this method can be integrated with other ICME methods deemed vital for the materials genome initiative. Furthermore, we designed a workflow for high-throughput screening of lightweight six- and seven-component HEAs. We show that the method has the potential to realize the ambitious aim to find high-component HEAs, as recently proposed by Cantor[19]. It also equips the research community with a general tool for the efficient discovery of materials.

Methods

Data collection and processing

Scientific texts appear in various formats, such as books, journals, etc., either in printed or electronic versions. The first step for corpora collection is to unify all these texts in a single digital format that can be directly used in machine-learning models (Fig. 1a of the main text). Here the training corpora of 6.4 million abstracts are downloaded through the ELSEVIER Scopus API[41]. The latter can retrieve abstracts in bulk with the journal ISSN and publishing year as input. We use the ISSN list generated by Tshitoyan et al.[1] as the starting point. The abstracts are stored in JSON format along with the metadata, such as authors, years of publication, keywords, journals, etc. In addition, we also manually add important journals and abstracts for HEAs that are absent in the first round of abstract collections. The representative journals for metallic materials of the past two decades include Acta Mater., Journal of Alloys and Compound, Materials Science and Engineering: A, and Advanced Engineering Materials. Note that there is a weekly download quota for regular Scopus developer API. The entire collection process of 6.4 million abstracts can take several months.

Machine-learning model

The skip-gram algorithm is adopted in this study. It has a neural-network structure but only with one hidden layer. Words in the training corpora are firstly encoded into one-hot vectors w_i. One component of each vector is "1", which records the word's location in the whole vocabulary, and the rest are zeros. These vectors are fed into the neural network as the training data. The training objective is to maximize the probability of each word in their context defined by a window size n (8), which is also the cost function to be optimized. The left and right eight words are considered the neighbors of the word in one basic unit, i.e., one abstract. The skip-gram model is trained for 30 epochs, and more epochs do not significantly improve the model performance. In the trained model, the key information of the corpora is encoded into the matrix M in the hidden layer. Multiplying w_i by M we obtain a representative vector v_i of 200 dimensions for the word i. There is no need to revisit the neural network for applications of the model. Words that are semantically or grammatically similar correspond to vectors that can reflect the similarity. The vector v_i has many interesting properties, such as compositionality and cosine similarity, as mentioned previously.

Here we feed the 6.4 million abstracts plus abstracts on metallic materials into the machine-learning model. The list of journals whose abstracts are duplicated is described in the supplementary material. The consequence is we place a double weight on the metallic materials. As a result, the specially tailored model is expected to work better for metallic materials than previous models.

Thermodynamic rule

Our thermodynamics-based rule was initially derived and published in ref. 30. It provides a systematic method to calculate the free energies ΔG_N for a given N-component system and ΔG_2 for all its binaries. If ΔG_N is the lowest, the multicomponent system is a single-phase alloy; otherwise, it is a multi-phase alloy. For convenience, we define a parameter γ to describe this criterion, i.e.,

$$\gamma := \begin{cases} \Delta G_N / \min(\Delta G_2) & \text{if } \min(\Delta G_2) < 0; \\ -\Delta G_N / \min(\Delta G_2) & \text{if } \Delta G_N < 0 \text{ and } \min(\Delta G_2) > 0. \end{cases} \quad (2)$$

The criterion now becomes $\gamma \geq 1$.

Density functional theory calculations

Density functional theory[42,43] simulations are carried out using Vienna Ab-initio Simulation Package (version 5.4.4)[44] to obtain the optimal volumes in a designated crystal structure (here FCC). The generalized gradient approximation parametrized by Perdew-Burke-Ernzerhof[45] is used to calculate the electronic exchange-correlation interaction, and the Kohn−Sham equation is solved using the projector augmented wave method[46], where the Brillouin zone is sampled using Monkhorst-Pack scheme[47]. The atomic configurations of elements in the pseudo-potentials used in our calculations are V [Ne3s^2]3p^{6}3d^{3}4s^2, Cr [Ar] 3d^{5}4s^1, Mn [Ar]3d^{6}4s^1, Fe [Ar]3d^{7}4s^1, Co [Ar]3d^{8}4s^1, Ni [Ar]3d^{8}4s^2, and Cu [Ar]3d^{9}4s^1. The relaxation stops when the energy difference between ionic steps is smaller than 10^{-5} eV. A plane wave cutoff of 400 eV and the k-point meshes of $10 \times 10 \times 10$ for the Brillouin zone are used. A supercell size of four atoms is adopted for pure elements and FCC structure in this study. In these calculations, only volume relaxation is needed.

Data availability

The article DOIs used to generate the training corpora in this study have been deposited in our GitHub repository under the accession link (https://github.com/peizong/alloy2vec). The raw training corpora data are protected and not shared due to the data privacy rules of Elsevier. Users can download it after they open an Elsevier account since all the papers are stored in their database. Details and guidelines to use the API and papers provided by Elsevier are here: https://dev.elsevier.com. Any reader can register there and receive an API account to reproduce the results. All copyright rules explained by Elsevier on that webpage must be followed.

Code availability

The *alloy2vec* code is available in GitHub (see Data Availability), and can also be downloaded through https://doi.org/10.5281/zenodo.7337034.

References

1. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
2. Nie, Z., Liu, Y., Yang, L., Li, S. & Pan, F. Construction and application of materials knowledge graph based on author disambiguation: revisiting the evolution of LiFePO$_4$. *Adv. Energy Mater.* **11**, 2003580 (2021).
3. Hakimi, O., Krallinger, M. & Ginebra, M.-P. Time to kick-start text mining for biomaterials. *Nat. Rev. Mater.* **5**, 553–556 (2020).
4. Court, C. J. & Cole, J. M. Magnetic and superconducting phase diagrams and transition temperatures predicted using text mining and machine learning. *npj Comput. Materials* **6**, 1–9 (2020).
5. Krenn, M. & Zeilinger, A. Predicting research trends with semantic and neural networks with an application in quantum physics. *Proc. Natl. Acad. Sci.* **117**, 1910–1916 (2020).
6. Grimmer, J. & Stewart, B. M. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Polit. Anal.* **21**, 267–297 (2013).
7. Ficcadenti, V., Cerqueti, R. & Ausloos, M. A joint text mining-rank size investigation of the rhetoric structures of the US presidents' speeches. *Expert Syst. Appl.* **123**, 127–142 (2019).
8. Birgmeier, J. et al. Amelie speeds mendelian diagnosis by matching patient phenotype and genotype to primary literature. *Sci. Transl. Med.* **12**, eaau9113 (2020).
9. Hoffmann, R. et al. Text mining for metabolic pathways, signaling cascades, and protein networks. *Science's STKE* **2005**, pe21–pe21 (2005).
10. Cheng, X., Cao, Q. & Liao, S. S. An overview of literature on covid-19, mers and sars: using text mining and latent dirichlet allocation. *J. Inf. Sci.* **48**, 304–320 (2020).
11. Mani, G. & Hope, T. Viral science: Masks, speed bumps, and guard rails. *Patterns* **1**, 100101 (2020).
12. Liu, X., Zhang, J. & Pei, Z. Machine learning for high-entropy alloys: progress, challenges and opportunities. *Prog. Mater. Sci.* **131**, 101018 (2022).
13. Yeh, J.-W. et al. Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes. *Adv. Eng. Mater.* **6**, 299–303 (2004).
14. Cantor, B., Chang, I., Knight, P. & Vincent, A. Microstructural development in equiatomic multicomponent alloys. *Mater. Sci. Eng.* **375**, 213–218 (2004).
15. Zhang, Y. et al. Microstructures and properties of high-entropy alloys. *Prog. Mater. Sci.* **61**, 1–93 (2014).
16. Miracle, D. B. & Senkov, O. N. A critical review of high entropy alloys and related concepts. *Acta Mater.* **122**, 448–511 (2017).
17. George, E. P., Raabe, D. & Ritchie, R. O. High-entropy alloys. *Nat. Rev. Mater.* **4**, 515–534 (2019).
18. Shi, P. et al. Hierarchical crack buffering triples ductility in eutectic herringbone high-entropy alloys. *Science* **373**, 912–918 (2021).
19. Cantor, B. Multicomponent high-entropy cantor alloys. *Prog. Mater. Sci.* **120**, 100754 (2020).
20. van de Schoot, R. et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nat. Mach. Intell.* **3**, 125–133 (2021).
21. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. https://arxiv.org/abs/1301.3781 (2013).
22. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. https://arxiv.org/abs/1310.4546 (2013).
23. Pennington, J., Socher, R. & Manning, C. D. Glove: global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543 (2014). http://www.aclweb.org/anthology/D14-1162.
24. Devlin, J., Chang, M., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. http://arxiv.org/abs/1810.04805. (2018).
25. Weston, L. et al. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *J. Chem. Inf. Model.* **59**, 3692–3702 (2019).
26. Wei, S. et al. Natural-mixing guided design of refractory high-entropy alloys with as-cast tensile ductility. *Nat. Mater.* **19**, 1175–1181 (2020).
27. Wang, F. et al. Multiplicity of dislocation pathways in a refractory multiprincipal element alloy. *Science* **370**, 95–101 (2020).
28. Senkov, O., Wilks, G., Miracle, D., Chuang, C. & Liaw, P. Refractory high-entropy alloys. *Intermetallics* **18**, 1758–1765 (2010).
29. Yin, J., Pei, Z. & Gao, M. C. Neural network-based order parameter for phase transitions and its applications in high-entropy alloys. *Nat. Comput. Sci.* **1**, 686–693 (2021).

30. Pei, Z., Yin, J., Hawk, J. A., Alman, D. E. & Gao, M. C. Machine-learning informed prediction of high-entropy solid solution formation: beyond the hume-rothery rules. *npj Comput. Mater.* **6**, 1–8 (2020).
31. Biswas, K. & Gurao, N. et al. Deciphering micro-mechanisms of plastic deformation in a novel single phase fcc-based mnfeconicu high entropy alloy using crystallographic texture. *Mater. Sci. Eng. A* **657**, 224–233 (2016).
32. Council, N. R. et al. Integrated computational materials engineering: a transformational discipline for improved competitiveness and national security (National Academies Press, 2008).
33. Sundman, B., Lukas, H. & Fries, S.*Computational thermodynamics: the Calphad method* (Cambridge university press New York, 2007).
34. de Pablo, J. J. et al. New frontiers for the materials genome initiative. *npj Comput. Mater.* **5**, 1–23 (2019).
35. Varvenne, C., Luque, A. & Curtin, W. A. Theory of strengthening in fcc high entropy alloys. *Acta Mater.* **118**, 164–176 (2016).
36. Toda-Caraballo, I. & Rivera-Díaz-del Castillo, P. E. Modelling solid solution hardening in high entropy alloys. *Acta Mater.* **85**, 14–23 (2015).
37. Wu, Z., Bei, H., Pharr, G. M. & George, E. P. Temperature dependence of the mechanical properties of equiatomic solid solution alloys with face-centered cubic crystal structures. *Acta Mater.* **81**, 428–441 (2014).
38. Pei, Z., Dutta, B., Körmann, F. & Chen, M. Hidden effects of negative stacking fault energies in complex concentrated alloys. *Phys. Rev. Lett.* **126**, 255502 (2021).
39. Introducing the knowledge graph: things, not strings. Google Official Blog (2021). Accessed: 2021-10-24.
40. Gludovatz, B. et al. A fracture-resistant high-entropy alloy for cryogenic applications. *Science* **345**, 1153–1158 (2014).
41. Elsevier research products APIs. https://dev.elsevier.com (2021). Accessed: 2021-10-24.
42. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).
43. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
44. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
45. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
46. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
47. Monkhorst, H. J. & Pack, J. D. Special points for brillouin-zone integrations. *Phys. Rev. B* **13**, 5188–5192 (1976).
48. Senkov, O., Scott, J., Senkova, S., Miracle, D. & Woodward, C. Microstructure and room temperature properties of a high-entropy TaNbHfZrTi alloy. *J. Alloys Compd.* **509**, 6043–6048 (2011).

Acknowledgements

We appreciate the proofreading and valuable comments of Dr. Anubhav Jain of Lawrence Berkeley National Laboratory and John Dagdelen of the University of California Berkeley. This research used resources from New York University's Greene supercomputer and Oak Ridge National Laboratory's Compute and Data Environment for Science (CADES) and the Oak Ridge Leadership Computing Facility (OLCF). The latter is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. One of us (P.K.L.) very much appreciates the support from (1) the National Science Foundation (DMR-1611180 and 1809640) with program directors, Drs. J. Yang, G. Shiflet, and D. Farkas and (2) the US Army Research Office (W911NF-13-1-0438 and W911NF-19-2-0049) with program managers, Drs. M.P. Bakas, S.N. Mathaudhu, and D.M. Stepp.

Author contributions

Z.P. designed the project, conceived the idea of the alloy-design method, performed the machine-learning study (construction and training of the models and analysis of results), and wrote the manuscript. J.Y. collected and pre-processed the training data, explored larger models to broaden our insights into the present study, and edited the manuscript. P.K.L. and D.R. helped fine-tune the project and edited the manuscript. All authors together finalized the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-35766-5.

Correspondence and requests for materials should be addressed to Zongrui Pei.

Peer review information *Nature Communications* thanks Simon Billinge and the other, anonymous, reviewer for their contribution to the peer review of this work.