IUCrJ
ISSN 2052-2525
MATERIALS | COMPUTATION

Explainable machine learning for materials discovery: predicting the potentially formable Nd–Fe–B crystal structures and extracting the structure–stability relationship

Tien-Lam Pham,[a,b] Duong-Nguyen Nguyen,[a] Minh-Quyet Ha,[a] Hiori Kino,[b,c] Takashi Miyake[b,c,d] and Hieu-Chi Dam[a,c,e]*

[a]Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan, [b]ESICMM, National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan, [c]Center for Materials Research by Information Integration, National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan, [d]CD-FMat,AIST, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan, and [e]JST, PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan. *Correspondence e-mail: dam@jaist.ac.jp

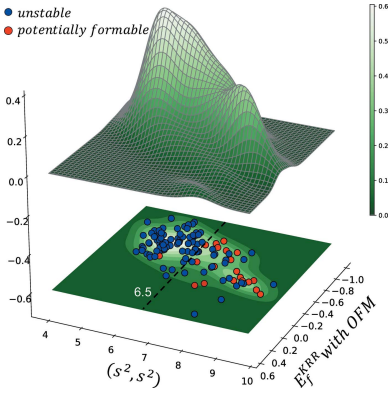
Received 3 February 2020
Accepted 21 July 2020

Edited by Y. Murakami, KEK, Japan

Keywords: data mining; machine learning; materials informatics; first-principles calculations; new magnets.

Supporting information: this article has supporting information at www.iucrj.org

New Nd–Fe–B crystal structures can be formed via the elemental substitution of *LA*–*T*–*X* host structures, including lanthanides (*LA*), transition metals (*T*) and light elements, *X* = B, C, N and O. The 5967 samples of ternary *LA*–*T*–*X* materials that are collected are then used as the host structures. For each host crystal structure, a substituted crystal structure is created by substituting all lanthanide sites with Nd, all transition metal sites with Fe and all light-element sites with B. High-throughput first-principles calculations are applied to evaluate the phase stability of the newly created crystal structures, and 20 of them are found to be potentially formable. A data-driven approach based on supervised and unsupervised learning techniques is applied to estimate the stability and analyze the structure–stability relationship of the newly created Nd–Fe–B crystal structures. For predicting the stability for the newly created Nd–Fe–B structures, three supervised learning models: kernel ridge regression, logistic classification and decision tree model, are learned from the *LA*–*T*–*X* host crystal structures; the models achieved maximum accuracy and recall scores of 70.4 and 68.7%, respectively. On the other hand, our proposed unsupervised learning model based on the integration of descriptor-relevance analysis and a Gaussian mixture model achieved an accuracy and recall score of 72.9 and 82.1%, respectively, which are significantly better than those of the supervised models. While capturing and interpreting the structure–stability relationship of the Nd–Fe–B crystal structures, the unsupervised learning model indicates that the average atomic coordination number and coordination number of the Fe sites are the most important factors in determining the phase stability of the new substituted Nd–Fe–B crystal structures.



1. Introduction

The major challenge in finding new stable material structures in nature requires high-throughput screening of an enormous number of candidate structures, which are generated from different atomic arrangements in three-dimensional space. In fact, only a handful of structures among these candidates are likely to exist. Therefore, for the non-serendipitous discovery of new materials, candidate structures must be generated strategically so that the screening space is reduced without overlooking potential materials.

Multiple strategies have been proposed for the high-throughput screening processes (Butler *et al.*, 2018; Curtarolo *et al.*, 2013; Saal *et al.*, 2013) for finding various new materials. Almost all well known screening methods consider first-prin-

ciples calculations as the basis for the estimation of physical properties. Screening processes have been successfully developed for theoretically understanding rare-earth-lean intermetallic magnetic compounds (Körner *et al.*, 2016, 2018), Heusler compounds (Ma *et al.*, 2017; He *et al.*, 2018; Balluff *et al.*, 2017), topological insulators (Yang *et al.*, 2012; Li *et al.*, 2018), perovskite materials (Emery *et al.*, 2016; Michalsky & Steinfeld, 2017), cathode coatings for Li-ion batteries (Aykol *et al.*, 2016) and M_2AX compounds (Ashton *et al.*, 2016). In recent years, various screening processes have been used to replace canonical approaches by machine learning (ML) methods. A few notable works based on ML models involve searching for hard-magnetic phases (Möller *et al.*, 2018), Heusler compounds (Kim *et al.*, 2018), bimetallic facet catalysts (Ulissi *et al.*, 2017), BaTiO3-based piezoelectrics (Xue *et al.*, 2016*b*), polymer dielectrics (Mannodi-Kanakkithodi *et al.*, 2016), perovskite halides (Pilania *et al.*, 2016) and low-thermal-hysteresis NiTi-based shape memory alloys (Xue *et al.*, 2016*a*).

ML is expected to play three different roles in performing screening processes. The first role is to replace the density functional theory (DFT) calculation and reduce the calculation cost of physical property estimation, *e.g.* convex hull distance (Kim *et al.*, 2018) and adsorption energy (Ulissi *et al.*, 2017). The reported models have achieved reasonable results in statistical evaluation tests such as cross validation. However, ensuring the reliability of extrapolating the physical properties of new materials is a major problem because the new screening materials do not always possess the same distribution as the training materials.

The second role of ML is to increase the success rate in screening processes. Given a list of hypothetical structures, ML methods are utilized for recommending the most likely new potential materials using probabilistic models [*e.g.* Bayesian optimization techniques (Yamashita *et al.*, 2018; Xue *et al.*, 2016*b*)]. This approach requires a list of potential candidates to be prepared as input, which is primarily based on human intuition. The bottleneck of the current recommendation methods is that a large number of known property materials are required as references for the system to start an effective recommendation process. This number increases dramatically with the material description dimension. Furthermore, the computational cost of the recommended process increases significantly with the number of reference materials.

The third role of ML is to effectively generate new structure candidates. The notable algorithms for this purpose are random search-based algorithms (Pickard & Needs, 2006, 2007, 2011; Wang *et al.*, 2010; Zhang *et al.*, 2017), evolutionary-algorithm-based algorithms such as USPEX (Glass *et al.*, 2006; Oganov *et al.*, 2011; Lyakhov *et al.*, 2013), XtalOpt (Lonie & Zurek, 2011) and recent deep-learning-based models (Noh *et al.*, 2019; Ryan *et al.*, 2018). In practice, it is possible to generate random structures by forcibly combining different crystal structures *in silico*. The successful discovery of novel material structures under high pressure demonstrates the effectiveness of this approach when certain constraints can be set. However, it is not easy to rationally combine different crystal structures with different compositions and symmetry in a plausible manner. Therefore, oversight in the search for a small number of potential materials cannot be controlled. The combination of first-principles calculations and ML is required for creating effective methods for exploring materials.

One of the most common strategies for generating possible crystal structure candidates is to appropriately combine or apply the atomic substitution method to previously known structures. Beginning with a dataset of host crystal structures with known physical properties and predefined substitution operators, we can employ the atomic substitution method to create new hypothetical crystal structures with the same skeleton as that of the host crystal structure. Widely used substitution operators such as single-site, multisite or element substitution operators are selected depending on the host dataset and experts' suggestions. These suggestions are typically based on domain knowledge about the physicochemical similarity between elements, atom–atom interactions, structural stability mechanisms and target physical property mechanisms. Consequently, the substitution method can work well with knowledge about material synthesis and lead directly to material synthesis ideas. Finally, an 'understanding' of the structure–stability relationship can be directly obtained from screening results, which can help in systematically correcting researchers' suggestions.

1.1. Our contribution

In this study, we propose a protocol for exploring new crystal structures under a given combination of constituent elements and the use of data mining to elucidate the structure–stability relationship (Fig. 1). As a demonstration example, we search for the new crystal structures of Nd–Fe–B materials by applying the atomic substitution method to a dataset containing host crystal structures composed of lanthanides, transition metals and light elements. We apply high-throughput first-principles calculations (Fig. 1, block A) to estimate the formation energy. Based on this, we evaluate the phase stability (hereinafter referred to as stability) of all generated Nd–Fe–B crystal structures (Section 2.2). The new Nd–Fe–B structures discovered after the screening steps are presented in Section 2.3. Supervised models are trained to mimic first-principles calculations from the host and substitution crystal structures and their calculated formation energy. Based on results from supervised learning models, relevance analysis is performed to extract the hidden structural descriptors that determine the formation energy of the generated Nd–Fe–B crystal structures (Fig. 1, block B). Finally, we trained an unsupervised learning model (Fig. 1, block C) that uses the obtained relevant descriptors to appropriately group newly generated crystal structures. We compare the obtained group labels and potentially formable states of all crystal structures to determine the relationship between the structure and stability of the Nd–Fe–B crystal structures.

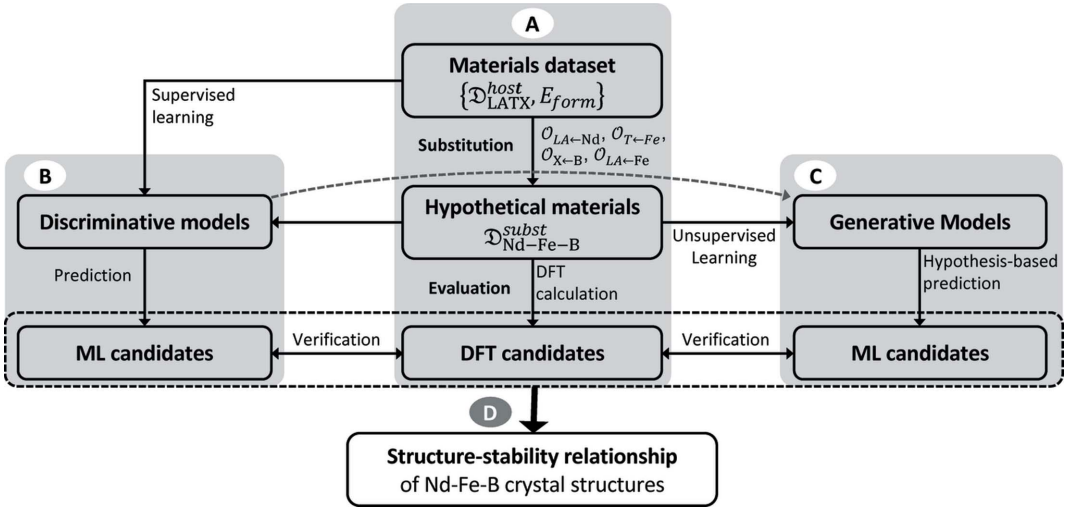


Figure 1
Workflow for extracting the structure–stability relationship of Nd–Fe–B crystal structures by integrating high-throughput first-principles calculations, supervised learning and unsupervised learning techniques.

2. Screening for potential Nd–Fe–B crystal structures

2.1. Creation of new crystal structure candidates

In this study, we focus on crystalline magnetic materials comprising a lanthanide (LA), a transition metal (T) and a light element (X). We selected the LA atoms from Y, La, Ce, Pr, Nd, Pm, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb and Lu; T from Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd, Hf, Ta, W, Re, Os, Ir, Pt, Au and Hg; and X from H, B, C, N and O. We collected the details of 5967 well known crystal structures with formation energies from the Open Quantum Materials Database (OQMD) (Saal *et al.*, 2013) (version 1.1) to form the host material dataset, denoted $\mathcal{D}^{\mathrm{host}}_{LA-T-X}$. Each host crystal structure consists of one or two rare-earth metals, one or two transition metals and one light element. Additionally, from $\mathcal{D}^{\mathrm{host}}_{LA-T-X}$ we selected a subset of all the crystal structures comprising Nd, Fe and B, denoted $\mathcal{D}^{\mathrm{host}}_{\mathrm{Nd-Fe-B}}$.

We create new candidates for crystal structures consisting of Nd, Fe and B with the same skeleton as the host crystal structures in $\mathcal{D}^{\mathrm{host}}_{LA-T-X}$ using a substitution method. For each host crystal structure, a substituted crystal structure is created by substituting all lanthanide sites with Nd, all transition metal sites with Fe and all light-element sites with B. The new structures are compared with each other and with the crystal structures in the $\mathcal{D}^{\mathrm{host}}_{LA-T-X}$ dataset to remove duplication. We follow the comparison procedure proposed by qmpy (the python application programming interface of OQMD) (Saal *et al.*, 2013*b*). The structures of the materials are transformed into reduced primitive cells to compare the two lattices; all lattice parameters are compared. The internal coordinates of the structures are compared by examining all rotations allowed by each lattice and searching for rotations and translations to map the atoms of the same species into one another within a given level of tolerance. Here, any two structures in which the percentage deviation in lattice parameters and angles are smaller than 0.1 are considered to be identical. Furthermore, we apply our designed orbital field matrix (OFM) (Section 3.1) to eliminate duplication. Two structures are considered to be the same if the L_2 norm of the difference in the OFM is less than 10^{-3}. Note that two structures with the same shape but slightly different in size are considered to be identical. Finally, we obtain a dataset for the substituted crystal structures, denoted $\mathcal{D}^{\mathrm{subst}}_{\mathrm{Nd-Fe-B}}$, with 149 new non-optimized Nd–Fe–B crystal structures. These structures are then optimized using the first-principles calculations described in detail in Section 2.2.

2.2. Assessment of phase stability

First-principles calculations based on DFT (Kohn & Sham, 1965; Hohenberg & Kohn, 1964) are one of the most effective calculation methods used in materials science. DFT calculations can accurately estimate the formation energy of materials, which is used to build phase diagrams for systems of interest. Hence, the phase stability of a material – in other words, the decomposition energy of a material (C—H distance) – is obtained via the convex hull analysis of phase diagrams and the decomposition of the material into other phases. We used the formation energy obtained from OQMD (Saal *et al.*, 2013*b*; Kirklin *et al.*, 2015) of $\mathcal{D}^{\mathrm{host}}_{LA-T-X}$ to build phase diagrams and calculate the C—H distance. The C—H distance of a material is defined as follows:

$$\Delta E = \Delta E_{\mathrm{f}} - E_{\mathrm{H}}, \tag{1}$$

where ΔE_{f} is the formation energy and E_{H} is determined by projecting from the chemical composition position to an end point appearing on the convex hull facets. Details of the algorithm for finding these convex hull facets from hull points can be found in the work by Barber *et al.* (1996) and Saal *et al.* (2013). Hereafter, we consider the C—H distance ΔE as the degree of the phase stability of a material. A material that lies below or on the C—H surface, $\Delta E = 0$, is a potentially formable material in nature, and a material associated with

Table 1
Properties of new Nd–Fe–B materials: formation energy by DFT $E_f^{\rm DFT}$ (eV per atom), stability by DFT $\Delta E^{\rm DFT}$, magnetization M ($\mu_{\rm B}$ per formula unit and $\mu_{\rm B}$ Å^{-3} in parentheses) and mean displacement Δr, estimated by hypothesized structures and final-optimized structures.

Formula	$E_f^{\rm DFT}$ (eV per atom)	$\Delta E^{\rm DFT}$ (eV per atom)	M [$\mu_{\rm B}$ ($\mu_{\rm B}$ Å^{-3})]	Δr (Å)	Host materials	OQMD id of host materials
Nd_2FeB_{10}	−0.522	−0.011	13.11 (0.050)	0.038	Ce_2NiB_{10}	2025052 (Jeitschko *et al.*, 2000)
$NdFe_2B_6$	−0.473	0.008	3.30 (0.040)	0.150	$CeCr_2B_6$	94775 (Kuzma & Svarichevskaya, 1972)
Nd_4FeB_{14}	−0.506	0.030	26.30 (0.063)	0.069	Ho_4NiB_{14}	2107958 (Geupel *et al.*, 2001)
$NdFe_2B_2$-α	−0.343	0.046	4.41 (0.067)	0.085	$DyCo_2B_2$	1852452 (Niihara *et al.*, 1987)
$NdFeB_4$-α	−0.462	0.052	17.42 (0.073)	0.041	$CeNiB_4$	2023354 (Akselrud *et al.*, 1984)
$NdFeB_4$-β	−0.455	0.060	18.73 (0.072)	0.050	$CeCrB_4$	2023373 (Kuzma *et al.*, 1973)
$Nd_2Fe_3B_5$	−0.374	0.066	6.85 (0.055)	0.143	$Eu_2Os_3B_5$	180411 (Schweitzer & Jung, 1986)
$Nd_2Fe_5B_4$	−0.284	0.069	10.31 (0.077)	0.206	$Eu_2Rh_5B_4$	183842 (Jung, 1990)
$NdFe_4B$-α	−0.092	0.070	21.64 (0.134)	1.769	$CeCo_4B$	185365 (Kuzma & Bilonizhko, 1973*a*)
$NdFe_{12}B_6$	−0.231	0.072	45.56 (0.117)	1.012	$CeNi_{12}B_6$	2077072 (Akselrud *et al.*, 1985)
$Nd_5Fe_{21}B_4$	−0.052	0.077	57.73 (0.140)	2.342	$Nd_5Co_{21}B_4$	126928 (Liang *et al.*, 2001)
$Nd_5Fe_{19}B_6$	−0.115	0.080	50.02 (0.128)	1.820	$Nd_5Co_{19}B_6$	125302 (Liang *et al.*, 2001)
$NdFe_4B$-β	−0.081	0.081	65.19 (0.135)	0.241	$NdNi_4B$	2069928 (Salamakha *et al.*, 2003)
$Nd_3Fe_{13}B_2$	−0.027	0.081	36.12 (0.144)	2.961	$Ce_3Ni_{13}B_2$	1778822 (Kuzma, 1981
$Nd_3Fe_{11}B_4$	−0.131	0.085	28.22 (0.122)	0.150	$Ce_3Co_{11}B_4$	1852403 (Kuzma & Bilonizhko, 1973*b*)
$Nd_2Fe_3B_6$	−0.375	0.088	16.02 (0.066)	0.132	$Ce_2Re_3B_6$	1966804 (Kuzma *et al.*, 1989)
$NdFe_4B_4$	−0.342	0.090	17.30 (0.048)	0.140	$CeRu_4B_4$	2074891 (Poettgen *et al.*, 2010)
$NdFe_2B_2$-β	−0.297	0.092	7.25 (0.057)	0.142	$CeIr_2B_2$	180315 (Jung, 1991)
$Nd_3Fe_8B_6$	−0.249	0.094	16.06 (0.079)	0.543	$Eu_3Rh_8B_6$	1771853 (Jung, 1990)
$Nd_2Fe_7B_3$	−0.147	0.096	35.04 (0.116)	0.209	$Ce_2Co_7B_3$	2016489 (Kuzma & Bilonizhko, 1974)

$\Delta E > 0$ is unstable. A material associated with ΔE slightly above the C—H surface is considered to be in a metastable phase.

Metastable phases are synthesized in numerous cases, for which we consider a reasonable range for the C—H distance (Balachandran *et al.*, 2018). Referring to the prediction accuracy of formation energy [∼0.1 eV per atom by OQMD (Saal *et al.*, 2013)], we define all materials with $\Delta E \leq 0.1$ eV per atom as potentially formable structures and as unstable materials otherwise. Following this definition, $\mathcal{D}^{\rm host}_{LA-T-X}$ can be divided into subsets $\mathcal{D}^{\rm host\_stb}_{LA-T-X}$ and $\mathcal{D}^{\rm host\_unstb}_{LA-T-X}$ for potentially formable crystal structures and unstable crystal structures, respectively.

$\mathcal{D}^{\rm host}_{\rm Nd-Fe-B}$ includes 35 Nd–Fe–B crystal structures, which can be used as references to construct the Nd–Fe–B phase diagram. Seven materials were found for ternary materials, which were comprised of Nd, Fe and B. To verify the reliability of the dataset used to construct the phase diagram as well as the stability definition, we removed each ternary material and used the remaining materials in $\mathcal{D}^{\rm host}_{\rm Nd-Fe-B}$ to estimate its corresponding convex hull distance. Under this test, among the seven ternary crystal structures, there are two formable ternary materials, $NdFe_4B_4$ and $Nd_5Fe_2B_6$, which lie on the surface of the CH of the phase diagram with ΔE = 0.0. Additionally, one material, $NdFe_{12}B_6$, is potentially formable (metastable) with a stability of less than 0.1 eV per atom, as shown in Table VI of the supporting information. It should be noted that the important magnetic material, $Nd_2Fe_{14}B$, did not exist in the OQMD database at the time when we conducted this study. Based on the Nd–Fe–B phase diagram and the formation energy of −0.057 eV per atom calculated using DFT, the corresponding $\Delta E^{\rm DFT}$ is 1.4×10^{-4} eV per atom. This result implies that $Nd_2Fe_{14}B$ is in the stable phase. To conclude, we confirm that the experimentally synthesized structures all satisfy the stability definition given in equation (1) in this section.

We followed the computational settings of OQMD (Saal *et al.*, 2013*b*; Kirklin *et al.*, 2015) for estimating the formation energy of the newly created Nd–Fe–B crystal structures in $\mathcal{D}^{\rm subst}_{\rm Nd-Fe-B}$. The calculations were performed using the *Vienna Ab initio Simulation Package* (VASP) (Kresse & Hafner, 1993, 1994; Kresse & Furthmüller, 1996*a*,*b*) by utilizing projector-augmented wave method potentials (PAW) (Blöchl, 1994; Kresse & Joubert, 1999) and the Perdew–Burke–Ernzerhof (PBE) (Perdew *et al.*, 1996) exchange-correlation functional.

We employed DFT + U for Fe, and all calculations were spin-polarized with ferromagnetic alignment of the spins and with initial magnetic moments of 5, 0 and 0 $\mu_{\rm B}$ for Fe, Nd and B, respectively. For each newly created structure, we performed coarse optimization, fine optimization and a single-point calculation, following the 'coarse relax', 'fine relax' and 'standard' procedures of the OQMD. The k-grid for these calculation series is selected by the k-points per reciprocal atom (KPRA): 4000, 6000 and 8000 for 'coarse relax', 'fine relax' and 'standard', respectively. We used a cutoff energy of 520 eV for all calculations. The total energies of the standard calculations are used for the formation energy calculations, $\Delta E_f^{\rm DFT}$. The C—H distance of a newly created structure can be estimated from $\Delta E^{\rm DFT} = \Delta E_f^{\rm DFT} - E_{\rm H}$.

After calculating the formation energy, we found 20 new Nd–Fe–B crystal structures that are not in $\mathcal{D}^{\rm host}_{LA-T-X}$, in which the C—H distance of the corresponding optimized structure is less than 0.1 eV. These structures originate from different host structures with different skeletons. Note that we found one structure, Nd_2FeB_{10}, with a stability of less than −0.01 eV per atom. Thus, this structure is also used as a reference to construct the Nd–Fe–B phase diagram. Among the 20 new Nd–Fe–B structures, there are three pairs of indistinguishable structures sharing the same chemical compositions ($NdFe_2B_2$, $NdFeB_4$ and $NdFe_4B$). Details about these structures are given in Table 1. The phase diagram of the Nd–Fe–B materials, including the 20 new substituted structures, is shown in Fig. 2.

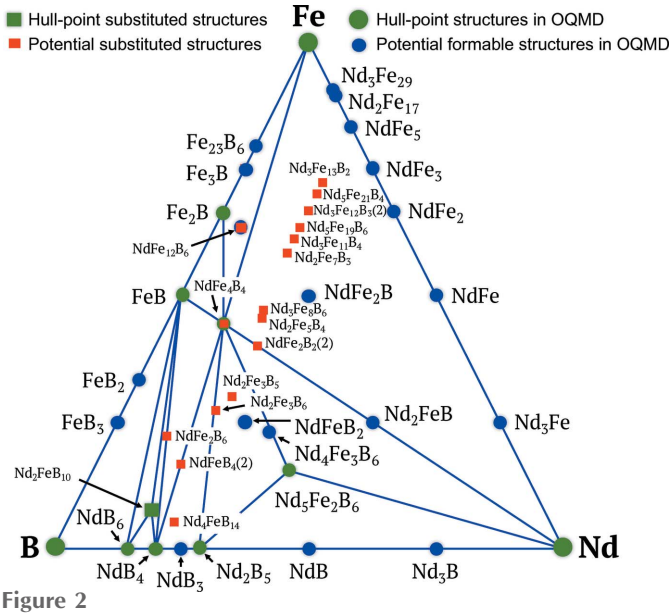


Figure 2
Phase diagram of Nd–Fe–B including materials obtained from OQMD (blue circles) and 20 new substituted structures confirming it is potentially formable (red squares). Hull points are denoted in green. The total number of disparate structures with the same chemical composition is shown in parentheses.

We also calculated the magnetization of these materials. We used open-core approximation to treat the 4*f* electrons of Nd. The contribution of 4*f* electrons to the magnetization is $J_{g_J} = 3.273$. The magnetization is normalized to the volume of a unit cell:

$$M = M_{\mathrm{DFT}} + J_{g_J} n_{\mathrm{Nd}} = M_{\mathrm{DFT}} + 3.273 n_{\mathrm{Nd}}, \qquad (2)$$

where M_{DFT} is the magnetization given by DFT and n_{Nd} is the number of Nd atoms in the unit cell. All calculation results are summarized in Table 1.

2.3. Newly discovered Nd–Fe–B crystal structures

Fig. 3 shows five specific crystal structures of the predicted formable crystal structures. A common characteristic of these crystal structures is that boron atoms form a network structure and Nd and Fe atoms are surrounded by the cages formed by the boron atom network. In the Nd_4FeB_{14} crystal structure, these boron cages are arranged in parallel and Fe atoms are sandwiched between two halves of the boron atom octahedron. In the crystal structure of Nd_2FeB_{10}, which is confirmed by DFT calculations and selected as the hull point in the phase diagram, Nd and Fe atoms are trapped in the boron atom cages; however, these cages are arranged in herringbone patterns. Interestingly, two stable crystal structures of $NdFeB_4$ were found as the proportion of Fe increased. One $NdFeB_4$-α structure was obtained by the elemental substitution of the original $CeNiB_4$ crystal structure [id: 2023354 (Akselrud *et al.*, 1984)]. This crystal structure is similar to the Nd_4FeB_{14} crystal structure, with cages formed by boron networks that trap Nd and Fe atoms and are arranged in parallel. In contrast, in the other predicted crystal structure for $NdFeB_4$ {$NdFeB_4$-β structure obtained by the elemental substitution of the $CeCrB_4$ [id: 2023373 (Kuzma *et al.*, 1973)] crystal structure}, the boron atoms form a planar network structure comprised of heptagon–pentagon ring pairs. Another form of boron cage is found in the $NdFe_2B_6$ crystal structure. All potentially formable crystal structures are shown in detail in the supporting information.

3. Mining structure–stability relationship of Nd–Fe–B crystal structures

3.1. Materials representation

We must convert the information regarding the materials into descriptor vectors. We employ the OFM (Lam Pham *et al.*, 2017; Pham *et al.*, 2018) descriptor with a minor modification. The OFM descriptors are constructed using the weighted product of the one-hot vector representations, **O**, of atoms. Each vector **O** is filled with zeros, except those representing the electronic configuration of the valence electrons of the corresponding atom. The OFM of a local structure, named θ, is defined as follows:

$$\Theta = \mathbf{O}^{\top}_{\mathrm{central}} \times \left(1.0, \sum_k \frac{\theta_k}{\theta_{\max}} \mathbf{O}_k\right), \qquad (3)$$

where θ_k is the solid angle determined by the face of the Voronoi polyhedra between the central atom and the index k neighboring atom, and $\theta_{\max}$ is the maximum solid angle between the central atom and neighboring atoms. By removing the distance dependence in the original OFM formulation (Lam Pham *et al.*, 2017; Pham *et al.*, 2018), we focus exclusively on the coordination of valence orbitals and the shape of the crystal structures. The mean over the local

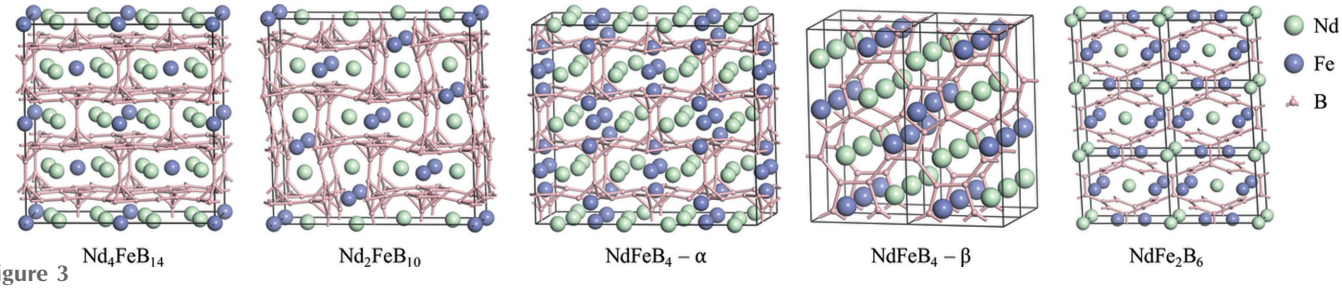


Figure 3
Representative Nd–Fe–B structures discovered by applying the elemental substitution method to the lanthanide, transition metal and rare-earth material dataset. Left to right: Nd_4FeB_{14}, Nd_2FeB_{10}, $NdFeB_4$-α, $NdFeB_4$-β and $NdFe_2B_6$. All 20 structures discovered are shown in the supporting information.

structure descriptors is used as the descriptor of the entire structure:

$$\mathrm{OFM}_p = \frac{1}{N_p}\sum_{l=1}^{N_p}\Theta_p^l, \tag{4}$$

where p is the structure index, and l and N_p are the local structure indices and the number of atoms in the unit cell of the structure p, respectively.

Note that owing to the designed cross product between the atomic representation vectors of each atom, each element in the matrix represents the average number of atomic coordinates for a certain type of atom. For example, an element of a descriptor obtained by considering the product of a d^6 element of the center atom representation and an f^4 element of the environment atom representation, denoted (d^6, f^4), shows an average coordination number of f^4 (Nd) sites surrounding all d^6 (Fe) sites. As the term s^2 appears at all descriptors for Fe, Nd and B sites, the element (s^2, s^2) represents the average coordination number of a given structure. All of these OFM elements provide a foundation for the intuitively interpretable investigation of the structure–stability relationship.

3.2. Mining of formation energy data of *LA–T–X* crystal structures with a supervised learning method

We trained the ML models that can predict the formation energy of the crystal structures, ΔE_f, from $\mathcal{D}^{\mathrm{host}}_{LA-T-X}$, which is represented using the OFM descriptor and the corresponding known formation energy data. We applied kernel ridge regression (KRR) (Murphy, 2012), which is demonstrated to be useful for predicting material properties. In the KRR algorithm, the target variable, $y = \Delta E_\mathrm{f}$, is represented by a weighted kernel function as follows:

$$\hat{y}_p = \sum_k c_k K(\mathbf{x}_p, \mathbf{x}_k) = \sum_k c_k \exp(-\gamma|\mathbf{x}_p - \mathbf{x}_k|), \tag{5}$$

where $\hat{y}_p$ is the predicted formation energy of crystal structure p; $\mathbf{x}_p$ and $\mathbf{x}_k$ are the representation vectors of crystal structures p and k based on the OFM descriptor, respectively; k runs over all crystal structures in the training set; $K(\mathbf{x}, \mathbf{x}_k)$ is the Laplacian kernel function. The c_k coefficients are estimated by minimizing the total square error regularized by the L_2 norm as follows: $\sum_k(\hat{y}_k - y_k)^2 + \lambda\sum_k c_k^2$, where y_k and $\hat{y}_k$ are the observed and predicted target values of the structure k, respectively. We perform a ten-times tenfold cross-validation process to determine parameters λ and γ in the KRR models. These parameters are selected by minimizing the mean absolute error (MAE) of the validation set.

Fig. 4 shows the ten-times tenfold cross-validated comparison of the formation energies calculated using DFT and those predicted by the KRR model for the crystal structures in $\mathcal{D}^{\mathrm{host}}_{LA-T-X}$ (blue circles). Fig. 4 also shows a comparison of the formation energies calculated using DFT and those predicted using the KRR model (trained using all crystal structures in $\mathcal{D}^{\mathrm{host}}_{LA-T-X}$) for the crystal structures in $\mathcal{D}^{\mathrm{subst}}_{\mathrm{Nd-Fe-B}}$ (red circles). In the cross-validated comparison of materials in $\mathcal{D}^{\mathrm{host}}_{LA-T-X}$, the formation energies predicted via KRR show good agreement with those calculated using DFT, with an R^2 (Kvålseth, 1985) value of 0.990 (1), see Table 2.

Table 2
Ten-times tenfold cross-validation results provided by the KRR model in predicting formation energy.

Model	R^2	MAE (eV per atom)	RMSE (eV per atom)
Kernel ridge	0.990 (1)	0.094 (2)	0.137 (1)

It should be noted that this predictive model is learned from the data ($\mathcal{D}^{\mathrm{host}}_{LA-T-X}$) containing only the optimized crystal structures. Thus, when applied to a newly generated non-optimized crystal structure (in $\mathcal{D}^{\mathrm{subst}}_{\mathrm{Nd-Fe-B}}$), it is clear that the possibility of correctly predicting the formation energy is low. The MAE of the KRR-predicted formation energy of the crystal structures in $\mathcal{D}^{\mathrm{subst}}_{\mathrm{Nd-Fe-B}}$ after structure optimization is approximately 0.3 (eV per atom), which is three times larger than the cross-validated MAE result. The results of applying the KRR prediction model to estimate the stability of these hypothetical materials are shown in detail in Section 3.5.

3.3. Descriptor-relevance analysis

Furthermore, we focus on $\mathcal{D}^{\mathrm{host}}_{\mathrm{Nd-Fe-B}}$ and evaluate the relevance (Nguyen *et al.*, 2019; Yu & Liu, 2004; Visalakshi & Radha, 2014) of each element in the OFM descriptor with respect to the formation energy of the crystal structure. We utilize the change in prediction accuracy when removing or adding a descriptor [from the full set of descriptors (Nguyen *et al.*, 2018) in the OFM] to search for the descriptors that are strongly relevant (Nguyen *et al.*, 2019; Dam *et al.*, 2018) to the

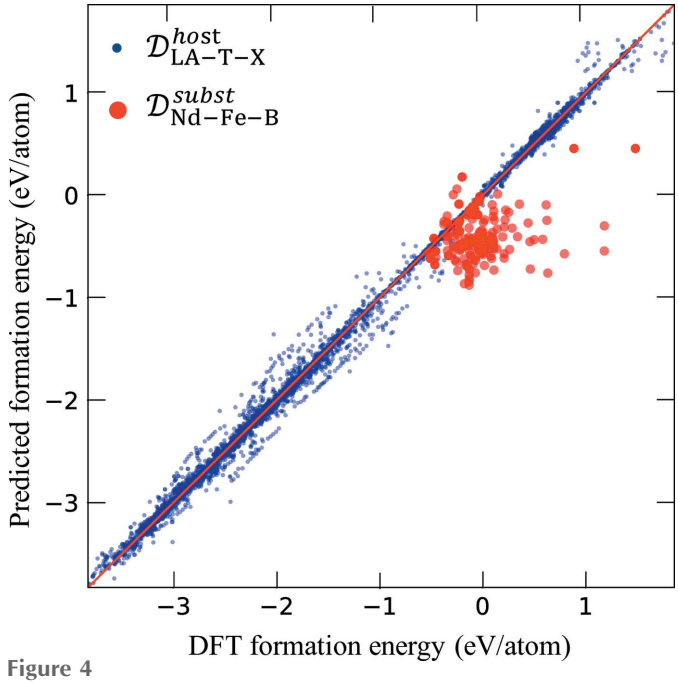


Figure 4
Comparison of formation energies calculated using DFT and those predicted through ML using the KRR model with the OFM descriptor. The blue and red solid circles represent the cross-validated results for $\mathcal{D}^{host}_{\mathrm{LA-T-X}}$ and the prediction results for $\mathcal{D}^{subst}_{\mathrm{Nd-Fe-B}}$, respectively.

formation energy (*i.e.* C—H distance and phase stability) of the Nd–Fe–B crystal structures.

In detail, for a given set S of descriptors, we define the prediction capacity $PC(S)$ of S by the maximum prediction accuracy that the KRR model can achieve using the variables in a subset s of S as follows:

$$PC(S) = \max_{\forall s \subset S} R_s^2;\ s_{PC} = \text{argmax}_{\forall s \subset S} R_s^2, \tag{6}$$

where R_s^2 is the value of the coefficient of determination R^2 (Kvålseth, 1985) achieved by the KRR using a set s as the independent variables. s_{PA} is the subset of S that yields the prediction model having the maximum prediction accuracy.

Let S_i denote a set of descriptors after removing a descriptor x_i from the full descriptor set S; $S_i = S - \{x_i\}$. A descriptor is strongly relevant if and only if

$$PC(S) - PC(S_i) = \max_{\forall s \subset S} R_s^2 - \max_{\forall s \subset S_i} R_s^2 > 0. \tag{7}$$

Fig. 5 summarizes the results obtained from the descriptor-relevance analysis. The black-triangled curve shows the dependence of the maximum prediction capacity (max. *PC*, in R^2 score) on the number of variables/OFM descriptors used in regression models. Other curves show the dependence of the maximum prediction capacity on the number of OFM descriptors used in regression models when a specific OFM is removed from the whole set of OFM descriptors. For example, the orange-dotted curve illustrates the max. *PC* of the OFM descriptor set without the appearance of the (p^1, s^2) descriptor. It is evident that the descriptor (s^2, s^2) (red-squared curve) is highly relevant to the prediction of the formation energy of the crystal structures in $\mathcal{D}^{\text{subst}}_{\text{Nd–Fe–B}}$. For further investigation, we project all substituted crystal structures in $\mathcal{D}^{\text{subst}}_{\text{Nd–Fe–B}}$ into the space of the KRR-predicted formation energy, E_f^{KRR} and (s^2, s^2), as shown in Fig. 6. One can easily deduce that the distribution of $\mathcal{D}^{\text{subst}}_{\text{Nd–Fe–B}}$ is a mixture of two distribution components. The larger distribution component is located in the region $(s^2, s^2) < 6.5$, whereas the other is located in the region $(s^2, s^2) \geq 6.5$. We infer the existence of two distinct groups of substituted crystal structures. The first group contains structures with average atomic coordination numbers lower than 6.5, and the second group contains structures with average atomic coordination numbers higher than 6.5. Furthermore, most newly discovered potentially formable crystal structures belong to the second group.

3.4. Mining of substituted Nd–Fe–B crystal structure data with an unsupervised learning method

In this section, we demonstrate the use of the proposed generative model, which applies the relevance analysis results and unsupervised learning, in contrast to the conventional supervised learning approach. As a result, this model performs detailed investigations at particular sites whose coordination numbers are highly correlated to the structure–stability relationship.

The underlying hypothesis of this approach is that there are various correlation patterns between crystal structure properties and their formation energies. Naturally, most of these patterns are for unstable crystal structures and only a few of these pertain to potentially formable crystal structures. These

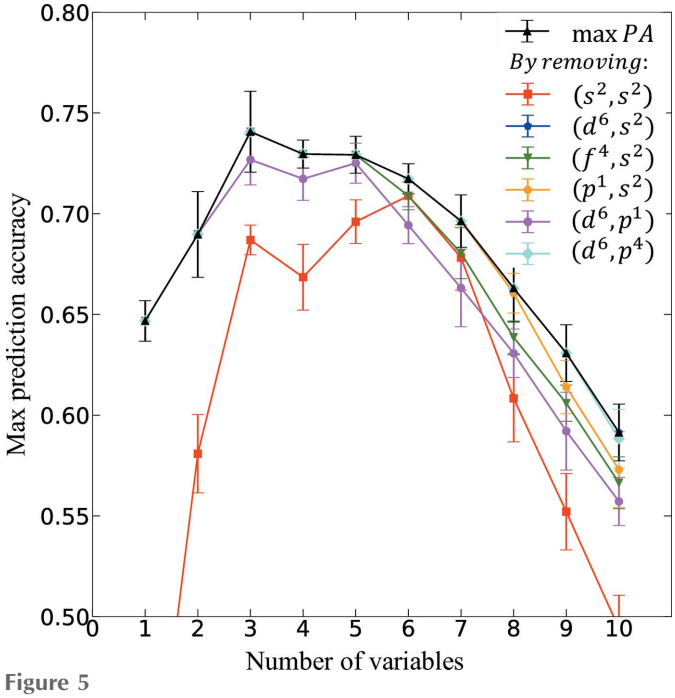


Figure 5
Results of the relevance analysis performed for predicting E_f of all Nd–Fe–B materials present in $\mathcal{D}^{\text{host}}_{LA-T-X}$. By removing the descriptor (s^2, s^2), the maximum prediction capacity (red line) is significantly reduced compared with the maximum prediction capacity line (max. *PC*) of all descriptor sets (black line).

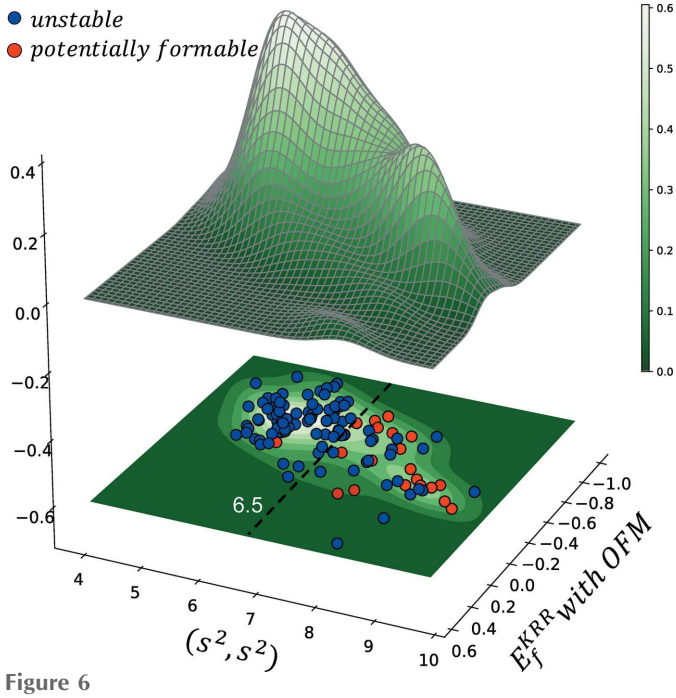


Figure 6
Distribution of substituted materials in space with the x axis showing the KRR-predicted formation energy, E_f^{KRR}, with non-optimized structures and the y axis showing the extracted strongly relevant descriptor (s^2, s^2). The black dotted line shows the limitation of (s^2, s^2), which maximizes the separation between two mixture distributions.

patterns might not be exposed directly through the feature-relevance analysis method due to the multivariate correlation between the target and description variables. The strong relevant descriptor (s^2, s^2) can appear as an extracted pattern to indicate the correlation between the structure–stability relationship. As the term s^2 appears at all descriptors for Fe, Nd and B sites, (s^2, s^2) indicates only the average atomic coordination numbers, which do not precisely represent the coordination number of any particular site. On the contrary, other OFM descriptors are designed to explicitly represent the coordination number of all pairwise elements. As the two terms d^6 and f^4 appear at only descriptors for Fe or Nd, respectively, in order to investigate the average coordination number of the Fe, Nd and B sites, in addition to (s^2, s^2), we focus on the values of the descriptors (d^6, s^2) and (f^4, s^2). These descriptors represent the average atomic coordination numbers of Fe sites and Nd sites. Furthermore, we also focus on the values of the OFM descriptors (d^6, d^6), (d^6, f^4), (f^4, d^6) and (f^4, f^4). These descriptors represent the average number of Fe sites surrounding the Fe sites, Nd sites surrounding the Fe sites, Fe sites surrounding the Nd sites and Nd sites surrounding the Nd sites. These OFM descriptors are useful in discussing not only the structure–stability relationship but also the strength of magnetic-exchange couplings between the 3*d* orbitals of Fe and the 4*f* orbitals of Nd.

Fig. 7 shows the density distribution of the newly created crystal structures, $\mathcal{D}^{\text{subst}}_{\text{Nd–Fe–B}}$, in two-dimensional space using the selected descriptors. For all pairs of descriptors, the density distribution is similar to the distribution of (s^2, s^2) and E_f^{KRR} shown in Fig. 6 with two clear peaks, one large and one small, with slight overlap. This result again confirms that (s^2, s^2) is highly relevant for expressing the nature of the distribution of the newly created crystal structures. In addition, (d^6, s^2) and (d^6, d^6) are important for identifying the characteristics of the distribution. It should be noted that these features could not be exposed using feature-relevance analysis since the prediction model can utilize the information from other highly correlated features instead, *e.g.* (s^2, s^2). In contrast, the average coordination number of the Nd sites (f^4, s^2) and the average coordination number of the Nd sites around the Nd sites (f^4, f^4) have a weak relationship with the characteristics related to the distribution of these crystal structures. These results indicate that the average coordination number of the Fe sites (d^6, s^2) and the average coordination number of the Fe sites around the Fe sites (d^6, d^6) are extremely important for characterizing the newly created Nd–Fe–B crystal structures.

We employed a GMM (Murphy, 2012) for learning the patterns of crystal structures by clustering $\mathcal{D}^{\text{subst}}_{\text{Nd–Fe–B}}$ into groups. The GMM model is based on the assumptions that the data consist of different groups and the data in each group follow their own Gaussian distribution. In other words, in the GMM, the distribution of data are fitted to a combination of a certain number, M, of Gaussian functions (Murphy, 2012) where M represents the number of data groups. The probability distribution of a crystal structure with index p, represented using selected descriptors, $\mathbf{x}_p$ and $f(\mathbf{x}_p)$, can be approximated as follows:

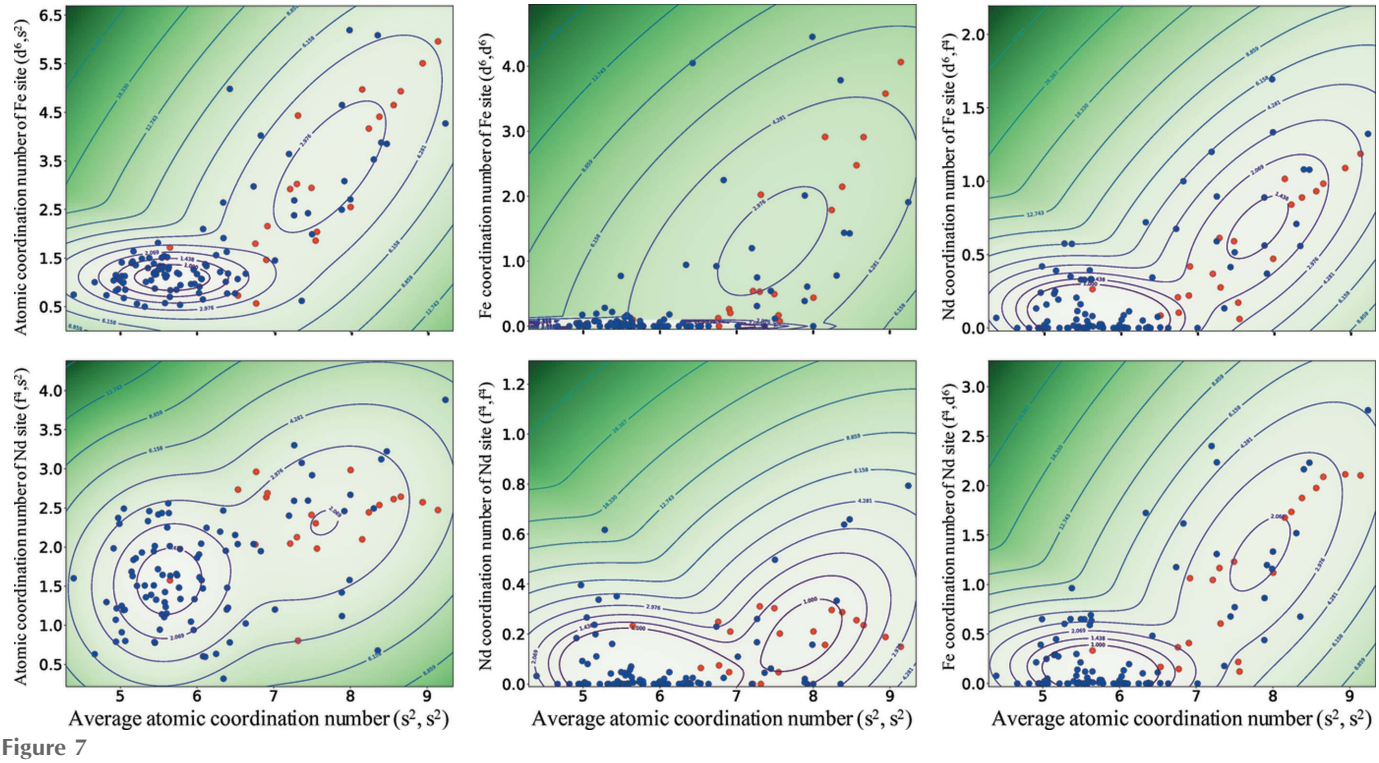


Figure 7
Density distribution of the newly generated Nd–Fe–B crystal structures in two-dimensional space obtained using selected OFM descriptors. The blue and red solid circles represent the unstable and potentially formable crystal structures verified by DFT calculations, respectively. Contour lines show the isodense surface of the distribution.

$$f(\mathbf{x}_p) = \sum_{m=1}^{M} \alpha_m \Phi(\mathbf{x}_p, \mu_m, \Sigma_m), \tag{8}$$

where

$$\Phi(\mathbf{x}_p, \mu_m, \Sigma_m) = \frac{\exp\left[-(\mathbf{x}_p - \mu_m)^T \Sigma_m^{-1} (\mathbf{x}_p - \mu_m)\right]}{(2\pi)^{d/2} |\Sigma_m|^{1/2}} \tag{9}$$

is a multivariate Gaussian distribution with mean μ_m and covariance matrix Σ_m and d is the dimension of the representation vector $\mathbf{x}_p$. The $\alpha_{\rm m}$ coefficients are the weights that satisfy the following constraint:

$$\sum_{m=1}^{M} \alpha_m = 1. \tag{10}$$

The probability that $\mathbf{x}_p$ belongs to group m can be represented as follows:

$$p(\mathbf{x}_p|m) = \frac{\alpha_m \Phi(\mathbf{x}_p, \mu_m, \Sigma_m)}{\sum_{m=1}^{M} \alpha_m \Phi(\mathbf{x}_p, \mu_m, \Sigma_m)}. \tag{11}$$

The model parameters α_m, μ_m and Σ_m are determined using an expectation-maximization algorithm (Pedregosa *et al.*, 2011). The number of data groups, M, is fixed at two in this study. It is interesting to note that the GMM provides a 'probabilistic image' of the pattern of crystal structures, wherein it provides the probability of a crystal structure remaining in a group instead of assigning the crystal structures to a specific group. The sum of the probabilities of crystal structures remaining in either of the groups is one. Therefore, the GMM is expected to discover distinctive patterns of crystal structures from the data and calculate the probability that a crystal structure belongs to a group.

We can label the newly generated crystal structures by fitting the data $\mathcal{D}^{\rm subst}_{\rm Nd-Fe-B}$ to the GMM with two Gaussian distributions and calculating the probabilities of the crystal structures belonging to each group. Given that it is not easy to find a new potential formable crystal structure, we suppose that most newly generated structures are unstable and only a few are potentially formable. Therefore, we infer that the large Gauss component corresponds to the distribution of unstable crystal structures and the small Gauss component corresponds to the distribution of potential formable crystal structures. This hypothesis can be verified through comparison with the results of the DFT calculations, and it can be seen that most of the potential formable crystal structures confirmed by DFT calculation actually belong to the small Gauss component. This implies that the phase stabilities of the Nd–Fe–B crystals are not significantly related to the coordination number of the Nd sites but are largely determined by the coordination number of the Fe sites, suggesting that, if the Nd sites can be replaced in part by Fe, the crystal structure characteristics of Nd–Fe–B which are directly related to its phase stability can be controlled. Further application of this discovery in the design of Nd–Fe–B crystal materials is promising.

Table 3
Evaluation results of KRR, LG and DT models, and unsupervised GMM in estimating the stability of materials in $\mathcal{D}^{\rm subst}_{\rm Nd-Fe-B}$.

Model	*Precision*	*Recall*	f_1
KRR model	0.533	0.534	0.376
LG model	0.629	0.687	0.599
DT model	0.704	0.676	0.687
GMM	0.729	0.821	0.735

3.5. Learning prediction models for the phase stability of crystal structures

A large number of ML applications reported to date (in materials science research) state the effectiveness and applicability of ML methods using statistical tests (such as cross validation). However, statistical tests are methods for assessing the risk in predicting the physical properties of the most optimized-structure materials, and are not appropriate for predicting and discovering novel materials. Therefore, in this study, to verify whether ML techniques are effective in searching for new potentially formable Nd–Fe–B crystal structures, we trained three supervised ML models from $\mathcal{D}^{\rm host}_{LA-T-X}$ and one unsupervised model from $\mathcal{D}^{subst}_{\rm Nd-Fe-B}$. In addition, we tested whether the models can predict the stability of the newly created crystal structures in $\mathcal{D}^{\rm subst}_{\rm Nd-Fe-B}$. The three supervised ML models are trained by considering 5967 materials in $\mathcal{D}^{\rm host}_{LA-T-X}$ with the OFM descriptor and the stability target values described in Sections 3.1 and 2.2. Then, all models are applied to predict the 149 newly hypothetical structures in $\mathcal{D}^{\rm subst}_{\rm Nd-Fe-B}$ while considering the stability calculated by the DFT as references in prediction accuracy evaluation.

In the first model (KRR model), the C—H distance is calculated using the formation energy predicted by the KRR model described in Section 3.2. Then, we applied a threshold of 0.1 eV per atom to the obtained C—H distance to determine whether the crystal structure is potentially formable. It is worth noting again that the bottleneck of this method is that the formation energy prediction model is learned from data containing only the optimal crystal structures. Therefore, the formation energy is not predicted correctly when the method is applied to a newly created non-optimal crystal structure.

The second model is a logistic regression model (LG model). From the two subsets of $\mathcal{D}^{\rm host}_{LA-T-X}$, including the potentially formable ($\mathcal{D}^{\rm host\_stb}_{LA-T-X}$) and unstable ($\mathcal{D}^{\rm host\_unstb}_{LA-T-X}$) crystal structures, we modelled the probability of observing potentially formable ($y = 1$) and unstable ($y = 0$) class labels directly using classification models. We hypothesized that the probability of observing potentially formable materials, $p(y = 1|X = \mathbf{x}_p)$, follows:

$$p(y = 1|X = \mathbf{x}_p) = \frac{\exp \sum_i c_i x_{pi}}{1 + \exp \sum_i c_i x_{pi}}, \tag{12}$$

where $\mathbf{x}_p$ is the description vector of structure p (obtained by flattening the OFM), i is the index of vector elements in $\mathbf{x}_p$ and c_i is the coefficient of the corresponding element x_π. In our experiments, all c_i coefficients are obtained via maximum *a*

Table 4
Classification results in predicting the 'potentially formable' class label of substituted materials with KRR, LG and DT models, GMM, and ensemble models.

The AND and OR operators in these ensemble models are denoted by & and |, respectively.

	KRR	LG	DT	GMM	KRR | GMM	LG | GMM	DT | GMM	KRR & GMM	LG & GMM	DT & GMM
Precision	0.24	0.35	0.56	0.49	0.24	0.36	0.48	**0.58**	0.53	**0.58**
Recall	0.82	0.79	0.45	0.91	0.97	**1.0**	0.91	0.76	0.7	0.45
f_1	0.37	0.49	0.5	0.64	0.39	0.53	0.63	**0.66**	0.61	0.51

Table 5
Classification results in predicting the 'unstable' class label of substituted materials with KRR, LG and DT models, GMM, and ensemble models.

The AND and OR operators in these ensemble models are denoted by & and |, respectively.

	KRR	LG	DT	GMM	KRR | GMM	LG | GMM	DT | GMM	KRR & GMM	LG & GMM	DT & GMM
Precision	0.83	0.91	0.85	0.97	0.94	**1.0**	0.97	0.92	0.91	0.85
Recall	0.25	0.59	0.90	0.73	0.14	0.49	0.72	0.84	0.83	**0.91**
f_1	0.38	0.71	0.87	0.83	0.24	0.66	0.83	**0.88**	0.86	**0.88**

posteriori estimation using L_1 as the regularization term (Ng, 2004; Lee *et al.*, 2006). The third model is the decision tree model (DT model) (Murphy, 2012), which uses information gain (Breiman *et al.*, 1984; Hastie *et al.*, 2009) as the criterion to measure the quality of tree-splitting.

The unsupervised model is based on the observations of the mixture distribution of the newly created crystal structures, $\mathcal{D}^{\rm subst}_{\rm Nd-Fe-B}$. We build the fourth model (GMM) by assuming that the major and minor Gauss components obtained correspond to the 'unstable' and 'potentially formable' class labels of the crystal structures, respectively.

The evaluation results of the four models are summarized in Table 3. We use three evaluation scores: *Precision*, *Recall* and f_1. The *Precision* score (also referred to as positive predictive value) with respect to the unstable structure class is the fraction of the unstable crystal structures predicted correctly among the number of crystal structures predicted to be unstable (Perry *et al.*, 1955). The *Recall* score (also known as sensitivity) with respect to the unstable structure class is the fraction of the unstable crystal structures predicted correctly among all crystal structures that are actually unstable (Perry *et al.*, 1955). The *Precision* and *Recall* scores are combined in the f_1 score (or *f*-measure) to provide a single measurement (Derczynski, 2016). To compare the classification ability of ML models, we summarize the evaluation scores of all classes (*i.e.* 'unstable' and 'potentially formable') by utilizing a macro averaging method (Su *et al.*, 2015) which is implemented in *sklearn.metrics.average_precision_score* (version 0.21.3; Pedregosa *et al.*, 2011).

The KRR model shows the lowest values of all evaluation scores among the three supervised learning models where *Precision*, *Recall* and f_1 are 0.533, 0.534 and 0.376, respectively. In contrast, the DT model provides the most accurate prediction. This model accurately predicts the potentially 'formable unstable' label of all substituted Nd–Fe–B crystal structures with 0.704 macro *Precision* score and obtains macro *Recall* and f_1 scores of 0.676 and 0.687, respectively. The LG model shows the highest macro *Recall* score, 0.687, compared with the other two supervised learning models.

The final but most surprising result is that the unsupervised GMM is superior to the other three supervised learning models in all three evaluation scores. The average *Precision* and *Recall* scores of the GMM are 0.729 and 0.821, respectively, which are significantly higher than those of the three supervised learning models. This result shows that the integration of descriptor-relevance analysis and unsupervised learning with the GMM is superior to conventional ML models, such as KRR, LG and DT, for obtaining information about the phase stability of substituted Nd–Fe–B crystal structures. We also investigated the usefulness of ensembling models. As the prediction problem under consideration is a binary classification, we implement two well known operators, 'AND' and 'OR,' for combining classification results. The details of the results are shown in Tables 4 and 5. These results again suggest that the structure–stability relationship obtained using data mining is highly promising for the design of Nd–Fe–B materials.

4. Conclusions

We focus on discovering new Nd–Fe–B materials using the elemental substitution method with *LA–T–X* compounds, with a lanthanide, transition metal and light element (X = B, C, N, O) as host materials. For each host crystal structure, a substituted crystal structure is created by substituting all lanthanide sites with Nd, all transition metal sites with Fe and all light-element sites with B. High-throughput first-principles calculations are applied to evaluate the phase stability of the newly created crystal structures, and twenty of them are found to be potentially formable. We implemented an approach by incorporating supervised and unsupervised learning techniques to estimate the stability and analyze the relationship between the structure and stability of the newly created Nd–Fe–B crystal structures. Three supervised learning models (KRR, LG and DT) learned from *LA–T–X* host crystal structures achieved the maximum accuracy and *Recall* scores of 70.4 and 68.7%, respectively, in predicting the stability state of new substituted Nd–Fe–B crystals. The proposed unsu-

pervised learning model resulting from the integration of descriptor-relevance analysis and the GMM provides accuracy and *Recall* scores of 72.9 and 82.1%, respectively, which are significantly better than those of the supervised models. Moreover, the unsupervised learning model can capture and interpret the structure–stability relationship of the Nd–Fe–B crystal structures. The average atomic coordination number and the coordination number of the Fe sites are quantitatively shown to be the most important factors in determining the phase stability of the new substituted Nd–Fe–B crystal structures.

Funding information

Funding for this research was provided by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) (award No. ESICMM12016013 to HCD, TLP, DNN, HK, TM); MEXT (JST) Precursory Research for Embryonic Science and Technology (PRESTO) (award to HCD); MEXT, Japan Society for the Promotion of Science (JSPS) (KAKENHI grant Nos. JP19H05815 to HCD; 20K05301) (Grant-in-Aid for Scientific Research on Innovative Areas 'Interface Ionics'); Materials research was carried out by the Information Integration Initiative (MI^2I) project of the Support Program for Starting Up Innovation Hub from JST, and MEXT as a social and scientific priority issue employing the post-K computer (creation of new functional devices and high-performance materials to support next-generation industries; CDMSI) (awarded to HCD, HK and TM).

References

Akselrud, L., Kuzma, Y. & Bruskov, V. (1985). *Dop. Akad. Nauk Ukr. RSR Ser. B*, **1985**, 33–35.
Akselrud, L. G., Kuzma, Yu. B., Pecharskii, V. K. & Bilonizhko, N. S. (1984). *Sov. Phys. Crystallogr.* **29**, 431–434.
Ashton, M., Hennig, R. G., Broderick, S. R., Rajan, K. & Sinnott, S. B. (2016). *Phys. Rev. B*, **94**, 054116.
Aykol, S., Kim, S., Hegde, D., Snydacker, Z., Lu, S., Hao, S., Kirklin, D., Morgan, D. & Wolverton, C. (2016). *Nat. Commun.* **7**, 13779.
Balachandran, P. V., Emery, A. A., Gubernatis, J. E., Lookman, T., Wolverton, C. & Zunger, A. (2018). *Phys. Rev. Mater.* **2**, 043802.
Balluff, K., Diekmann, G., Reiss, G. & Meinert, M. (2017). *Phys. Rev. Mater.* **1**, 034404.
Barber, C. B., Dobkin, D. P. & Huhdanpaa, H. (1996). *ACM Trans. Math. Softw.* **22**, 469–483.
Blöchl, P. E. (1994). *Phys. Rev. B*, **50**, 17953–17979.
Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees.* London: Taylor & Francis.
Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. (2018). *Nature*, **559**, 547–555.
Chen, Y., Li, X., Chen, X. L., Liang, J. K., Rao, G. H., Shen, B. G., Liu, Q. L., Jin, L. P. & Wang, M. Z. (2000). *Chem. Mater.* **12**, 1240–1247.
Curtarolo, S., Hart, G. L. W., Nardelli, M. B., Mingo, N., Sanvito, S. & Levy, O. (2013). *Nat. Mater.* **12**, 191–201.
Dam, H. C., Nguyen, V. C., Pham, T. L., Nguyen, A. T., Terakura, K., Miyake, T. & Kino, H. (2018). *J. Phys. Soc. Jpn*, **87**, 113801.
Derczynski, L. (2016). *Proceedings of the International Conference on Language Resources*, http://www.lrec-conf.org/proceedings/lrec2016/summaries/105.html.
Emery, A. A., Saal, J. E., Kirklin, S., Hegde, V. I. & Wolverton, C. (2016). *Chem. Mater.* **28**, 5621–5634.
Geupel, S., Zahn, G., Paufler, P. & Graw, G. (2001). *Z. Kristallog. New Cryst. Struct.* **216**, 175–176.
Glass, A. R., Oganov, A. R. & Hansen, N. (2006). *Comput. Phys. Commun.* **175**, 713–720.
Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed, Springer Series in Statistics. New York: Springer-Verlag.
He, J., Naghavi, S. S., Hegde, V. I., Amsler, M. & Wolverton, C. (2018). *Chem. Mater.* **30**, 4978–4985.
Hohenberg, P. & Kohn, W. (1964). *Phys. Rev.* **136**, B864–B871.
Jeitschko, W., Konrad, T., Hartjes, K., Lang, A. & Hoffmann, R. (2000). *J. Solid State Chem.* **154**, 246–253.
Jung, W. (1990). *J. Less-Common Met.* **161**, 375–384.
Jung, W. (1991). *J. Less-Common Met.* **171**, 119–125.
Kim, K., Ward, L., He, J., Krishna, A., Agrawal, A. & Wolverton, C. (2018). *Phys. Rev. Mater.* **2**, 123801.
Kirklin, S., Saal, J. E., Meredig, B., Thompson, A., Doak, J. W., Aykol, M., Rühl, S. & Wolverton, C. (2015). *npj Comput. Mater.* **1**, 15010.
Kohn, W. & Sham, L. J. (1965). *Phys. Rev.* **140**, A1133–A1138.
Körner, W., Krugel, G. & Elsässer, C. (2016). *Sci. Rep.* **6**, 24686.
Körner, W., Krugel, G., Urban, D. F. & Elsässer, C. (2018). *Scr. Mater.* **154**, 295–299.
Kresse, G. & Furthmüller, J. (1996*a*). *Comput. Mater. Sci.* **6**, 15–50.
Kresse, G. & Furthmüller, J. (1996*b*). *Phys. Rev. B*, **54**, 11169–11186.
Kresse, G. & Hafner, J. (1993). *Phys. Rev. B*, **47**, 558–561.
Kresse, G. & Hafner, J. (1994). *Phys. Rev. B*, **49**, 14251–14269.
Kresse, G. & Joubert, D. (1999). *Phys. Rev. B*, **59**, 1758–1775.
Kuzma, Y. & Bilonizhko, M. (1973*a*). *Sov. Phys. Crystallogr.* **18**, 710–714.
Kuzma, Y. & Bilonizhko, N. (1974). *Izv. Akad. Nauk SSSR Neorg. Mater.* **10**, 265–269.
Kuzma, Y. & Bilonizhko, N. S. (1973*b*). *Kristallografiya*, **18**, 710.
Kuzma, Y., Mikhalenko, S. & Chaban, N. (1989). *Sov. Powder Met. Met. Ceram.* **28**, 60–64.
Kuzma, Y. B. & Bilonizhko, N. S. (1981). *Dop. Akad. Nauk. Ukr. RSR Ser. A*, **43**, 87–90.
Kuzma, Y. B. & Svarichevskaya, S. I. (1972). *Kristallografiya*, **17**, 939–941.
Kuzma, Y. B., Svarichevskaya, S. I. & Fomenko, V. N. (1973). *Izv. Akad. Nauk. Neorg. Mater.* **9**, 1542–1545.
Kvålseth, T. O. (1985). *Am. Stat.* **39**, 279–285.
Lam Pham, T., Kino, H., Terakura, K., Miyake, T., Tsuda, K., Takigawa, I. & Chi Dam, H. (2017). *Sci. Technol. Adv. Mater.* **18**, 756–765.
Lee, S.-I., Lee, H., Abbeel, P. & Ng, A. Y. (2006). *Proceedings, 21st National Conference on Artificial Intelligence (AAAI-06)*. Palo Alto: AAAI Press.
Li, X., Zhang, Z., Yao, Y. & Zhang, H. (2018). *2D Materials*, **5**, 045023.
Liang, J., Rao, G., Chu, W., Yang, H. & Liu, G. (2001). *J. Appl. Phys.* **90**, 1931–1933.
Lonie, D. C. & Zurek, E. (2011). *Comput. Phys. Commun.* **182**, 372–387.
Lyakhov, A. O., Oganov, A. R., Stokes, H. T. & Zhu, Q. (2013). *Comput. Phys. Commun.* **184**, 1172–1182.
Ma, J., Hegde, V. I., Munira, K., Xie, Y., Keshavarz, S., Mildebrath, D. T., Wolverton, C., Ghosh, A. W. & Butler, W. H. (2017). *Phys. Rev. B*, **95**, 024411.
Mannodi-Kanakkithodi, A., Pilania, G., Huan, T. D., Lookman, T. & Ramprasad, R. (2016). *Sci. Rep.* **6**, 20952.
Michalsky, R. & Steinfeld, A. (2017). *Catal. Today*, **286**, 124–130.
Möller, J. J., Körner, W., Krugel, G., Urban, D. F. & Elsässer, C. (2018). *Acta Mater.* **153**, 53–61.
Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective.* Cambridge: MIT Press.
Ng, A. Y. (2004). *International Conference on Machine Learning*, https://doi.org/10.1145/1015330.1015435.
Nguyen, D.-N., Pham, T.-L., Nguyen, V.-C., Ho, T.-D., Tran, T., Takahashi, K. & Dam, H.-C. (2018). *IUCrJ*, **5**, 830–840.

Nguyen, D., Pham, T., Nguyen, V., Nguyen, A., Kino, H., Miyake, T. & Dam, H. (2019). *J. Phys. Conf. Ser.* **1290**, 012009.
Niihara, K., Yajima, S. & Shishido, T. (1987). *J. Less-Common Met.* **135**, 1137–1140.
Noh, J., Kim, J., Stein, H. S., Sanchez-Lengeling, B., Gregoire, J. M., Aspuru-Guzik, A. & Jung, Y. (2019). *Matter*, **1**, 1370–1384.
Oganov, A. O., Lyakhov, A. O. & Valle, M. (2011). *Acc. Chem. Res.* **44**, 227–237.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). *J. Mach. Learn. Res.* **12**, 2825.
Perdew, J. P., Burke, K. & Ernzerhof, M. (1996). *Phys. Rev. Lett.* **77**, 3865–3868.
Perry, J. W., Kent, A. & Berry, M. M. (1955). *Amer. Doc.* **6**, 242–254.
Pham, T., Nguyen, N., Nguyen, V., Kino, H., Miyake, T. & Dam, H. (2018). *J. Chem. Phys.* **148**, 204106.
Pickard, C. J. & Needs, R. J. (2006). *Phys. Rev. Lett.* **97**, 045504.
Pickard, C. J. & Needs, R. J. (2007). *Nat. Phys.* **3**, 473–476.
Pickard, C. J. & Needs, R. J. (2011). *J. Phys. Condens. Matter*, **23**, 053201.
Pilania, G., Balachandran, C., Kim, C. & Lookman, T. (2016). *Front. Mater.* **3**, 19.
Poettgen, E., Matar, S. F. & Mishra, T. M. (2010). *Z. Anorg. Allg. Chem.* **636**, 1236–1241.
Ryan, J., Lengyel, J. & Shatruk, M. (2018). *J. Am. Chem. Soc.* **140**, 10158–10168.
Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. (2013). *JOM*, **65**, 1501–1509.
Salamakha, P., Sologub, O., Mazumdar, Ch., Alleno, E., Noël, H., Potel, M. & Godart, C. (2003). *J. Alloys Compd.* **351**, 190–195.
Schweitzer, K. & Jung, W. (1986). *Z. Anorg. Allg. Chem.* **533**, 30–36.
Su, W., Yuan, Y. & Zhu, M. (2015). *Proceedings of the 2015 International Conference on the Theory of Information Retrieval ICTIR'15*, pp. 349–352, https://doi.org/10.1145/2808194.2809481.
Ulissi, Z. W., Tang, M. T., Xiao, J., Liu, X., Torelli, D. A., Karamad, M., Cummins, K., Hahn, C., Lewis, N. S., Jaramillo, T. F., Chan, K. & Nørskov, J. K. (2017). *ACS Catal.* **7**, 6600–6608.
Visalakshi, S. & Radha, V. (2014). *2014 IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1–6, https://doi.org10.1109/ICCIC.2014.7238499/.
Wang, Y., Lv, J., Zhu, L. & Ma, Y. (2010). *Phys. Rev. B*, **82**, 094116.
Xue, D., Balachandran, P. V., Hogden, J., Theiler, J., Xue, D. & Lookman, T. (2016*a*). *Nat. Commun.* **7**, 11241.
Xue, D., Balachandran, R., Yuan, T., Hu, X., Qian, X., Dougherty, E. R. & Lookman, T. (2016*b*). *Proc. Natl Acad. Sci. USA*, **113**, 13301–13306.
Yamashita, T., Sato, N., Kino, H., Miyake, T., Tsuda, K. & Oguchi, T. (2018). *Phys. Rev. Mater.* **2**, 013803.
Yang, K., Setyawan, W., Wang, S., Buongiorno Nardelli, M. & Curtarolo, S. (2012). *Nat. Mater.* **11**, 614–619.
Yu, L. & Liu, H. (2004). *J. Mach. Learn. Res.* **5**, 1205.
Zhang, Y., Wang, H., Wang, Y., Zhang, L. & Ma, Y. (2017). *Phys. Rev. X*, **7**, 011017.