IUCrJ
ISSN 2052-2525
MATERIALS | COMPUTATION

Received 24 July 2018
Accepted 21 September 2018

Edited by X. Zhang, Tsinghua University, China

Keywords: data mining; materials informatics; first-principles calculations; physical properties of materials; machine learning; similarity.

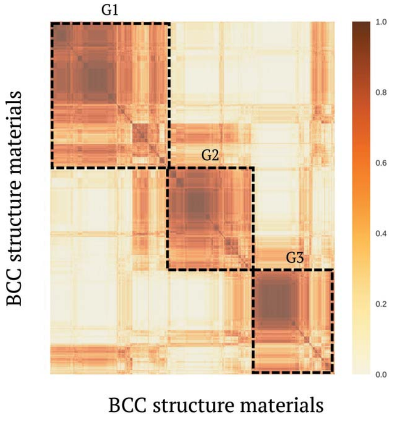
Supporting information: this article has supporting information at www.iucrj.org

Committee machine that votes for similarity between materials

Duong-Nguyen Nguyen,[a] Tien-Lam Pham,[a,b] Viet-Cuong Nguyen,[c] Tuan-Dung Ho,[a] Truyen Tran,[d] Keisuke Takahashi[e] and Hieu-Chi Dam[a,e,f]*

[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]HPC Systems Inc., 3-9-15 Kaigan, Minato-ku, Tokyo 108-0022, Japan, [d]Applied Artificial Intelligence Institute, Deakin University, Geelong, Australia, [e]Center for Materials Research by Information Integration, National Institute for Materials Science 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan, and [f]JST, PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama, 332-0012, Japan. *Correspondence e-mail: dam@jaist.ac.jp

A method has been developed to measure the similarity between materials, focusing on specific physical properties. The information obtained can be utilized to understand the underlying mechanisms and support the prediction of the physical properties of materials. The method consists of three steps: variable evaluation based on nonlinear regression, regression-based clustering, and similarity measurement with a committee machine constructed from the clustering results. Three data sets of well characterized crystalline materials represented by critical atomic predicting variables are used as test beds. Herein, the focus is on the formation energy, lattice parameter and Curie temperature of the examined materials. Based on the information obtained on the similarities between the materials, a hierarchical clustering technique is applied to learn the cluster structures of the materials that facilitate interpretation of the mechanism, and an improvement in the regression models is introduced to predict the physical properties of the materials. The experiments show that rational and meaningful group structures can be obtained and that the prediction accuracy of the materials' physical properties can be significantly increased, confirming the rationality of the proposed similarity measure.



1. Introduction

Computational materials science encompasses a range of methods to model materials and simulate their responses on different length and time scales (Sumpter *et al.*, 2015). The majority of problems addressed by computational materials science are related to methods that focus on two central tasks. The first aims to predict the physical properties of materials, and the second aims to describe and interpret the underlying mechanisms (Liu *et al.*, 2017; Lu *et al.*, 2017; Ulissi *et al.*, 2017). In the first task of predicting physical properties, computer-based quantum mechanics techniques (Jain *et al.*, 2016; Kohn & Sham, 1965; Jones & Gunnarsson, 1989; Jones, 2015) in the form of well established first-principles calculations are generally performed with high accuracy and are applicable to any material, but with high computational cost. Recently, the increase in the use of advanced machine-learning techniques (Murphy, 2012; Hastie *et al.*, 2009; Le *et al.*, 2012) and the volume of computational materials databases (Jain *et al.*, 2013; Saal *et al.*, 2013) have provided new opportunities for researchers to construct prediction models automatically (from a huge amount of precomputed data) that predict specific physical properties with the same level of high accuracy, while dramatically reducing the computational costs (Behler & Parrinello, 2007; Snyder *et al.*, 2012; Pilania *et al.*, 2013; Fernandez *et al.*, 2014; Smith *et al.*, 2017). By contrast, the second task, *i.e.* describing and interpreting the mechanisms underlying the physical properties of materials, relies mostly on the experience, insight and even luck of the experts involved. In fact, comprehension of multivariate data with nonlinear correlations is typically extremely challenging, even for experts. Thus, the utilization of data-mining and machine-learning techniques to discover hidden structures and latent semantics in multidimensional data (Lum *et al.*, 2013; Landauer *et al.*, 1998; Blei, 2012) of materials is promising, but only limited work has been reported so far (Kusne *et al.*, 2015; Srinivasan *et al.*, 2015; Goldsmith *et al.*, 2017).

To apply well established machine-learning methods to solve problems in materials science, the primitive representation of materials must usually be converted into vectors, in such a way that the comparison and calculations using the new representation reflect the nature of the materials and the underlying mechanisms of the chemical and physical phenomena. However, real-world applications, especially for solving the second task, often focus on physical properties of which the mechanism is not fully understood (Rajan, 2015; Ghiringhelli *et al.*, 2015). In these cases, it is almost impossible to represent the materials appropriately as vectors of features so that comparisons using well established mathematical calculations can reflect the similarity/dissimilarity between them. Therefore, a true data-driven approach for solving materials science problems still requires much further fundamental development.

In this study, we focus on establishing a data-driven protocol for solving the second task of computational materials science. Focusing on a specific physical property, we aim to develop a method to measure the similarity between materials from the viewpoint of the underlying mechanisms that act in these materials. The method for measuring this similarity consists of three steps: (i) variable evaluation based on nonlinear regression, (ii) regression-based clustering and (iii) similarity measurement with a committee machine (Tresp, 2001; Opitz & Maclin, 1999) constructed based on the clustering results. The variable evaluation (Liu & Yu, 2005; Blum &Langley, 1997) aims to identify and remove irrelevant and redundant variables from the data (Duangsoithong & Windeatt, 2009; Almuallim & Dietterich, 1991; Biesiada & Duch, 2007). We carried out this analysis in an exhaustive manner by testing all combinations of predicting variables to find those variables with the potential to yield good prediction accuracy (PA) for the target variable. The regression-based clustering method is developed from the well known *K*-means clustering method (Lloyd, 1982; MacQueen, 1967; Kanungo *et al.*, 2002) with major modifications for breaking down a large data set into a set of separate smaller data sets, in each of which the target variables can be predicted by a different linear model. Regression-based clustering models are then constructed for all the selected potential combinations of predicting variables, so as to construct a committee machine that votes for the similarity between the materials.

We evaluated the proposed protocol on three data sets of well characterized crystalline materials represented by appropriate predicting variables, together with their physical properties as determined through first-principles calculations or measured experimentally. Our experiments show that the proposed similarity measure can derive rational and meaningful material groupings and can significantly improve the prediction accuracy (PA) of the physical properties of the examined materials.

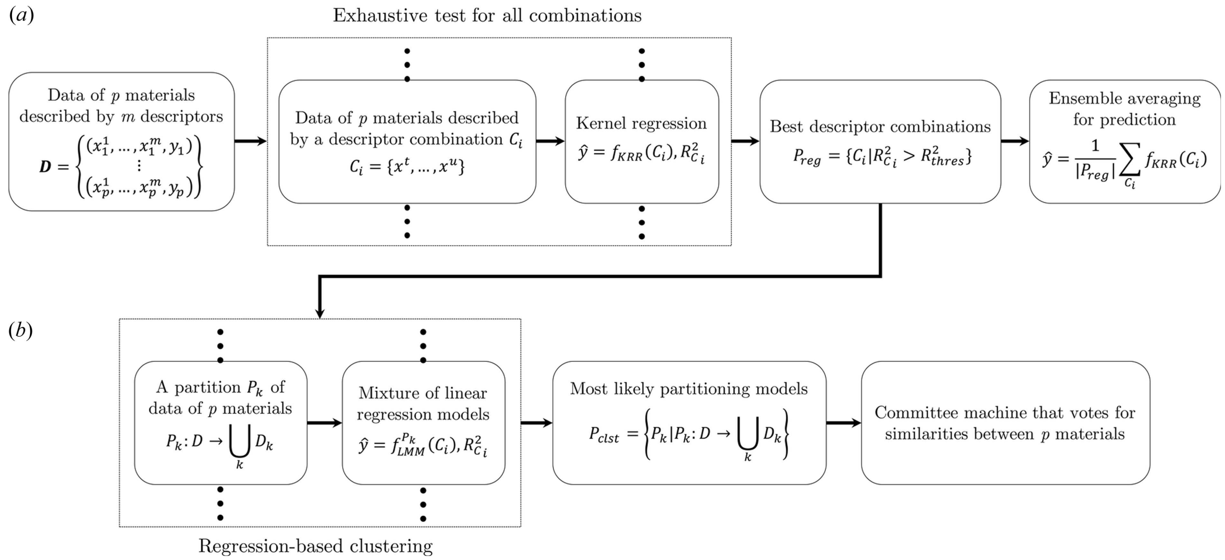

(a)
Exhaustive test for all combinations
Data of p materials described by m descriptors
$D = \{(x_1^1, \ldots, x_1^m, y_1) \vdots (x_p^1, \ldots, x_p^m, y_p)\}$
Data of p materials described by a descriptor combination C_i
$C_i = \{x^t, \ldots, x^u\}$
Kernel regression
$\hat{y} = f_{KRR}(C_i), R^2_{C_i}$
Best descriptor combinations
$P_{reg} = \{C_i | R^2_{C_i} > R^2_{thres}\}$
Ensemble averaging for prediction
$\hat{y} = \frac{1}{|P_{reg}|}\sum_{C_i} f_{KRR}(C_i)$
(b)
A partition P_k of data of p materials
$P_k: D \to \bigcup_k D_k$
Mixture of linear regression models
$\hat{y} = f_{LMM}^{P_k}(C_i), R^2_{C_i}$
Most likely partitioning models
$P_{clst} = \{P_k | P_k: D \to \bigcup_k D_k\}$
Committee machine that votes for similarities between p materials
Regression-based clustering

Figure 1
The data flow in our proposed method to measure similarity between materials, focusing on specific target physical properties and using the MapReduce representation language. The process consists of two subprocesses: (*a*) an exhaustive test for all predicting variable combinations, from which we can select the best combinations yielding the most likely regression models, and (*b*) a utilization of the regression-based clustering technique to search for partition models that can break down the data set into a set of separate smaller data sets, so that each target variable can be predicted by a different linear model. We can obtain a prediction model with higher predictive accuracy by taking an ensemble average of the models yielded in (*a*). We use the obtained partitioning models in (*b*) to construct a committee machine that votes for the similarity between materials.

2. Methods

We consider a data set $\mathcal{D}$ of p materials. Assume that a material with index i is described by an m-dimensional predicting variable vector $\mathbf{x}_i = (x_i^1, x_i^2, \ldots, x_i^m) \in \mathbb{R}^m$. The data set $\mathcal{D}$ is then represented using a $(p \times m)$ matrix. The target physical-property values of the materials are stored as a p-dimensional target vector $\mathbf{y} = (y_1, y_2 \ldots y_p) \in \mathbb{R}^p$. The entire data-analysis flow is shown in Fig. 1.

2.1. Kernel regression-based variable evaluation

To develop a better understanding of the processes that generated the data, we first utilize an exhaustive search to evaluate all variable combinations (Liu & Yu, 2005; Blum & Langley, 1997; Kohavi & John, 1997) to identify and remove irrelevant and redundant variables (Duangsoithong & Windeatt, 2009; Almuallim & Dietterich, 1991; Biesiada & Duch, 2007). We begin by learning nonlinear functions to predict the values of a specific physical property (target quantity) of the materials. We apply the Gaussian kernel ridge regression (GKR) technique (Murphy, 2012), which has recently been applied successfully to several challenges in materials science (Rupp, 2015; Botu & Ramprasad, 2015; Pilania *et al.*, 2013). For GKR, the predicted property $y = f(\mathbf{x})$ at a point $\mathbf{x}$ is expressed as the weighted sum of Gaussians:

$$f(\mathbf{x}) = \sum_{i=1}^{p} c_i \exp\left(\frac{-||\mathbf{x}_i - \mathbf{x}||_2^2}{2\sigma^2}\right), \tag{1}$$

where p is the number of training data points, σ^2 is a parameter corresponding to the variance of the Gaussian kernel function, and $||\mathbf{x}_i - \mathbf{x}||_2^2 = \sum_{\alpha=0}^{m}(x_i^\alpha - x^\alpha)^2$ is the squared L^2 norm of the difference between the two m-dimensional vectors $\mathbf{x}_i$ and $\mathbf{x}$. The coefficients c_i are determined by minimizing

$$\sum_{i=1}^{p}\left[f(\mathbf{x}_i) - y_i\right]^2 + \lambda \sum_{i=1}^{p} |c_i|_2^2, \tag{2}$$

where y_i is the observed physical property for material i. The hyper-parameters σ and the regularization parameter λ are selected with the help of cross-validation, *i.e.* by excluding some of the materials as a validation set during the training process and measuring the coefficient of determination R^2, which is defined (Kvalseth, 1985) as

$$R^2 = 1 - \frac{\sum_{j=1}^{p_{\rm vld}}\left[f\left(\mathbf{x}_j\right) - y_j\right]^2}{\sum_{j=1}^{p_{\rm vld}}\left[\overline{y} - y_j\right]^2}. \tag{3}$$

Here, $p_{\rm vld}$ is the number of validation points and $\overline{y}$ is the average of the validation set used to compare the values predicted for the excluded materials with the known observed values. In this study, we use R^2 as a measure of PA.

To estimate the PA accurately, we cross-validate the GKR (Stone, 1974; Picard & Cook, 1984; Kohavi, 1995) repeatedly using the collected data. To obtain a set of proper variable combinations that can accurately predict the target variable, we train the GKR models for all possible combinations of numerical predicting variables. It should be noted that, since we do not yet know the effect of each predicting variable on the target quantity, all the numerical predicting variables are normalized in the same manner in this analysis. With each combination, we search for the regularization parameters to maximize the PA of the corresponding GKR model. Note that each of the selected combinations contributes a perspective on the correlation between the target and the predicting variables. Thus, an ensemble averaging (Tresp, 2001; Dietterich, 2000; Zhang & Ma, 2012) technique can be applied to combine all the pre-screened regression models to improve the PA. Further, the similarity between materials regarding the mechanisms of the chemical and physical phenomena associated with the target quantity can be investigated more comprehensively if we consider all the perspectives. Consequently, we need to construct regression-based clustering models for each obtained potential combination to build the committee machine.

2.2. Regression-based clustering

In practice, a single linear model is often severely limited for modelling real data, because the data set can be nonlinear or the data themselves can be heterogeneous and contain multiple subsets, each of which fits best to a different linear model. However, in traditional data analysis, linear models are often preferred because of their interpretability. Within a linear model, one can intuitively understand how the predicting variables contribute to the target variable. Therefore, much effort has been devoted to developing subspace segmentation techniques to deconvolute a high-dimensional data set into a set of separate small data sets, each of which can be approximated well by different linear subspaces by employing principal component analysis (Fukunaga & Olsen, 1971; Vidal *et al.*, 2015; Einbeck *et al.*, 2008).

In this study, our primary interest is the local linearity between the predicting variables and the target variable, which may reflect the nature of the underlying physics around the point of observation. Therefore, we employ a simple strategy, in which the subspace segmentation is an integration of a conventional clustering method and linear regression analysis. It should be noted that the subspaces may have fewer dimensions than the whole space. Hence, we apply sparse linear regression analysis using *L1* regularization (Tibshirani, 1996) instead of the original one.

Our proposed regression-based clustering method is based on the well known K-means clustering method with two major modifications. (i) The sparse linear regression model derived from data associated with materials in a particular cluster (group) is considered to be its common characteristic (centre). The dissimilarities in the characteristics of each material in a group relative to the shared (common) nature of that group (the distance to the centre) are measured according to their deviation from the corresponding linear regression model. (ii) The sum of the differences of all materials in a group from the corresponding linear regression model of another group is used to measure the dissimilarity in the characteristics of that group with regard to the other group. The sum of the dissimilarities between one group and another and that determined in the reverse direction are used to assess the divergence between the two groups.

After performing the variable evaluation, we assume we have selected combinations of predicting variables that yield nonlinear regression models of high PA. With one of the selected combinations, m' numerical variables are selected from the original m numerical variables. A material in the data set is then described by an m'-dimensional predicting variable vector $\mathbf{x}_i' = (x_i^1, x_i^2, \ldots, x_i^{m'}) \in \mathbb{R}^{m'}$, and the data are represented using a $(p \times m')$ matrix.

Given the set $\mathcal{D}$ of p data points represented by m'-dimensional numerical vectors, a natural number $k \le p$ represents the number of clusters for a given experiment. We assume that there are k linear regression models and that each data point in $\mathcal{D}$ follows one of them. The aim is to determine those k linear regression models accordingly, to divide $\mathcal{D}$ into k non-empty disjoint clusters. Our algorithm searches for a partition of $\mathcal{D}$ into k non-empty disjoint clusters $(\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k)$ that minimize the overall sum of the residuals between the observed and predicted values (using the corresponding models) of the target variable. The problem can be formulated in terms of an optimization problem as follows.

For a given experiment with cluster number k, minimize

$$P(W, M) = \sum_{i=1}^{k} \sum_{j=1}^{p} w_{ij} || y_j - y_j^{M_i} ||, \qquad (4)$$

subject to

$$\forall j : \sum_{i=1}^{k} w_{ij} = 1, w_{ij} \in \{0, 1\}, \qquad (5)$$

$$1 \le k \le p, 1 \le i \le k, 1 \le j \le p, \qquad (6)$$

where y_j and $y_j^{M_i}$ are, respectively, the observed value and the value predicted by model M_i (of k models) for the target property of the material with index j, $W = [w_{ij}]_{p\times k}$ is a partition matrix (w_{ij} takes a value of 1 if object x_j belongs to cluster $\mathcal{D}_i$ and 0 otherwise) and $M = (M_1, M_2, \ldots, M_k)$ is the set of regression models corresponding to clusters $(\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k)$.

P can be optimized by iteratively solving two smaller problems:

(i) Fix $M = \hat{M}$ and solve the reduced problem $P(W, M)$ to find $\hat{W}$ (reassign data points to the cluster of the closest centre); and

(ii) Fix $W = \hat{W}$ and solve the reduced problem $P(W, M)$ to find $\hat{M}$ (reconstruct the linear model for each cluster).

Our regression-based clustering algorithm comprises three steps and iterates until $P(W, M)$ converges to some local minimum values:

(i) The data set is appropriately partitioned into k subsets, $1 \le k \le p$. Multiple linear regression analyses are performed independently with the *L1* regularization method (Tibshirani, 1996) on each subset to learn the set of potential candidates for the sparse linear regression models $M^{(0)} = \{M_1^{(0)}, M_2^{(0)}, \ldots, M_k^{(0)}\}$. This represents the initial step $t = 0$;

(ii) $M^{(t)}$ is retained and problem $P(W, M^{(t)})$ is solved to obtain $W^{(t)}$, by assigning data points in $\mathcal{D}$ to clusters based upon models $M_1^{(t)}, M_2^{(t)}, \ldots, M_k^{(t)}$;

(iii) $W^{(t)}$ is fixed and $M^{(t)}$ is generated such that $P(W, M^{(t+1)})$ is minimized. That is, new regression models are learned according to the current partition in step (ii). If the convergence condition or a given termination condition are fulfilled, the result is output and the iterations are stopped. Otherwise, t is set to $t + 1$ and the algorithm returns to step (ii).

The group number k is chosen considering two criteria: high linearity between the predicting and target variables for all members of the group, and no model representing two different groups. The first criterion has higher priority and can be quantitatively evaluated using the Pearson correlation scores between the predicted and observed values for the target variable of the data instances in each group, by applying the corresponding linear model. The second criterion is implemented to avoid the case in which one group with high linearity is further divided into two subgroups that can be represented by the same linear model. The determination of k, therefore, can be formulated in terms of an optimization problem as follows:

$$k = \arg\min_{k \le p} \left[\log \frac{1 - \min_{1 \le i \le k} R_{i,i}^2}{\min_{1 \le i \le k} R_{i,i}^2} + \max_{1 \le i \ne j \le k} R_{i,j}^2 \right], \qquad (7)$$

where $R_{i,i}^2$ and $R_{i,j}^2$ are the Pearson correlation scores between the predicted and observed values for the target variable when we apply the linear model M_i to data instances in clusters i and j, respectively.

The first term in this optimization function decreases monotonically with respect to the range of $\min_{1 \le i \le k} R_{i,i}^2$ varying from 0 to 1. When $\min_{1 \le i \le k} R_{i,i}^2$ approaches 1 (the entire cluster exhibits almost perfect linearity between the target and predicting variables), the optimization function drops on a log scale to emphasize the expected region. In contrast, the optimization function increases exponentially when $\min_{1 \le i \le k} R_{i,i}^2$ approaches 0 (one of the clusters shows no linearity between the target and predicting variables). The second term in this optimization function is introduced to avoid overestimation of k, in which a group with high linearity further divides into two subgroups that can be represented by the same linear model. It should be noted that the criterion for determining k is also the criterion for evaluating a regression-based clustering model. Further, cluster labels can be assigned for a material without knowing the value of the target physical property, using the estimated value obtained from a prediction model, *e.g.* a nonlinear regression model.

2.3. Similarity measure with committee machine

A clustering model, obtained through regression-based clustering for a particular combination of predicting variables, represents a specific partitioning of the data set into groups in which the linear correlations between the predicting and target variables can be observed. Materials belonging to the same group potentially have the same actuating mechanisms for the target physical property. However, materials that actually have the same actuating mechanisms for a specific physical property should be observed similarly in many circumstances. Therefore, the similarity between materials, focusing on a specific physical property, should be measured in a multilateral manner. For this purpose, for each prescreening of the sets of predicting variables that yield nonlinear regression models of high PA (Section 2.1), we construct a regression-based clustering model. A committee machine that votes for the similarity between materials is then constructed from all obtained clustering models. The similarity between two materials can be measured naïvely using the committee algorithm (Seung *et al.*, 1992; Settles, 2010), by counting the number of clustering models that partition these two materials into the same cluster. The affinity matrix A of all pairs of materials in the data set is then constructed as follows:

$$A_{a,b} = \frac{1}{|S_h|} \sum_{\forall S \in S_h} \sum_{i=1}^{k_S} w_{ia}^{S} w_{ib}^{S}, \qquad (8)$$

where S_h is the set of all prescreened combinations of predicting variables that yield nonlinear regression models of high PA and k_s is the cluster number. Further, $W^S = [w_{ij}^S]_{p \times k_S}$ is the partition matrix of the clustering models obtained through regression-based clustering analysis using the combination of predicting variables S (w_{ia}^S takes a value of 1 if material a belongs to cluster i and 0 otherwise). Using this affinity matrix, one can easily implement a hierarchical clustering technique (Everitt *et al.*, 2011) to obtain a hierarchical structure of groups of materials that have similar correlations between the predicting and target variables.

3. Results and discussion

We applied the methods described above to a sequential analysis for automatic extraction of physicochemical information relating to considered materials from three available data sets. For each data set, a brute-force examination of all combinations of numerical predicting variables was conducted using a nonlinear regression technique, to identify combinations of predicting variables that yielded regression models of high PA for the later analysis process. For each of the prescreened combinations, physically meaningful patterns in the form of material groups, as well as the linear relationships between the selected predicting and target variables, could be detected automatically for the materials in each group utilizing the regression-based clustering technique. The committee machine was then constructed from the obtained clustering models. Subsequently, a hierarchical structure of material groups similar to each other could be extracted using the hierarchical clustering technique. We evaluated the obtained results from both qualitative and quantitative perspectives. The qualitative evaluations were based on the rationality and interpretability of the obtained hierarchy with reference to the domain knowledge; the quantitative evaluations were performed based on the PA of the predictive models constructed with reference to the obtained similarity between materials.

Table 1
The designed predicting variables describing the intrinsic properties of the constituent elements and the structural properties of the materials in the E_{form} prediction problem.

The A and B elements comprise the AB materials with a binary cubic structure identical to that of the $Fm\overline{3}m$ symmetry group.

Category	Predicting variables
Atomic properties of A element	Z_A, $r_{\mathrm{ion}A}$, r_A, IP$_A$, χ_A, $n_{\mathrm{e}A}$, $T_{\mathrm{b}A}$, $T_{\mathrm{m}A}$
Atomic properties of B element	Z_B, $r_{\mathrm{ion}B}$, r_B, IP$_B$, χ_B, $n_{\mathrm{e}B}$, $T_{\mathrm{b}B}$, $T_{\mathrm{m}B}$
Structural information	V_{cell}

The exhaustive search for variable selection based on kernel regression consumes a lot of computing resources, such as memory and CPU time, due to combinatorial explosion. We performed our experiments using Apache Spark (Zaharia *et al.*, 2016) on a high-performance cluster with 256 processor cores and 1.1 TB of RAM in total. The calculation cost depends on various factors, such as the number of instances of data, the number of features and the cross-validation estimate parameters. With our system, the exhaustive search task takes 36, 41 and 28 h, respectively, to perform the first, second and third experiments.

3.1. Experiment 1: mining the quantum calculated formation energy data for $Fm\overline{3}m$ AB materials

In this experiment, we collected computational data for 239 binary AB materials from the Materials Project database (Jain *et al.*, 2013). The A atoms were virtually all metallic forms: alkali, alkaline earth, transition and post-transition metals, as well as lanthanides. The B elements, by contrast, were mostly all metalloids and non-metallic atoms. We set the computed formation energy E_{form} of each AB material as the physical property of interest. To simplify the demonstration of our method, we limited the collected compounds to those possessing the same cubic structure as the $Fm\overline{3}m$ symmetry group (*i.e.* the NaCl structure).

To represent each material, we used a set of 17 predicting variables divided into three categories, as summarized in Table 1. The first and second categories pertained to the predicting variables of the atomic properties of the element A and element B constituents; these included eight numerical predicting variables: (i) atomic number (Z_A, Z_B); (ii) atomic radius (r_A, r_B); (iii) average ionic radius ($r_{\mathrm{ion}A}$, $r_{\mathrm{ion}B}$); (iv) ionization potential (IP$_A$, IP$_B$); (v) electronegativity (χ_A, χ_B); (vi) number of electrons in the outer shell ($n_{\mathrm{e}A}$, $n_{\mathrm{e}B}$); (vii) boiling temperature ($T_{\mathrm{b}A}$, $T_{\mathrm{b}B}$); and (viii) melting temperature ($T_{\mathrm{m}A}$, $T_{\mathrm{m}B}$) of the corresponding single substances. The boiling and melting temperatures were as measured under standard conditions (0°C, 10^5 Pa). Information related to crystal structure is very valuable for understanding the physical properties of materials. Therefore, we designed the third category with structural predicting variables whose values were calculated from the crystal structures of the materials. In this experiment, owing to the similarities in the crystal structures of the collected materials, we utilized only the unit-cell volume (V_{cell})

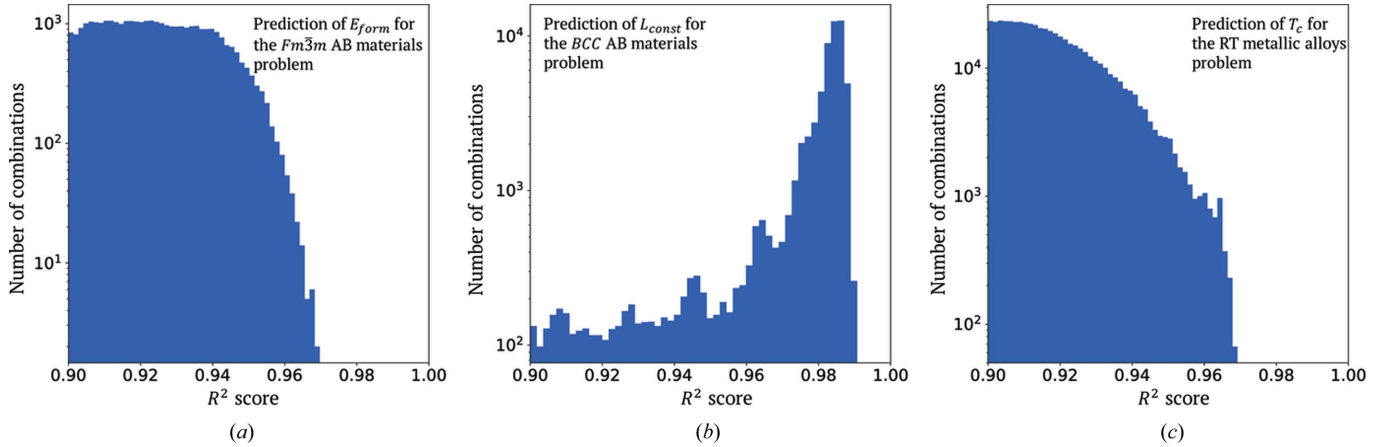


Figure 2
The numbers of predicting variable combinations that yield corresponding prediction models with R^2 larger than 0.90 for different problems: (*a*) the prediction of $E_{\rm form}$ for the $Fm\bar{3}m$ AB materials, (*b*) the prediction of $L_{\rm const}$ for the b.c.c. AB materials and (*c*) the prediction of magnetic phase-transition temperature $T_{\rm C}$ for the rare earth–transition metal alloys.

as the structural predicting variable. The computed $E_{\rm form}$ of each material was set as the target variable.

A kernel regression-based variable evaluation was performed for these data with 3 × 10-fold cross-validations. We first examined how $E_{\rm form}$ can be predicted from the designed predicting variables for all collected materials. We performed a screening for all possible ($2^{17} - 1 = 131\,071$) variable combinations. Hence, we found a total of 34 468 variable combinations deriving GKR models with R^2 scores exceeding 0.90 (Fig. 2). Among these, there were 139 variable combinations deriving GKR models with R^2 scores exceeding 0.96. These predicting variable combinations were then considered as candidates for the next step of the analysis. The highest prediction accuracy (PA) in this experiment is 0.967 (mean of absolute error, abbreviated as MAE: 0.122 eV), obtained using the combination $\{V_{\rm cell}, \chi_A, n_{{\rm e}A}, n_{{\rm e}B}, {\rm IP}_A, T_{{\rm b}A}, T_{{\rm m}A}, r_B\}$. Moreover, we could obtain a superior PA with an R^2 score of 0.972 (MAE: 0.117 eV) by taking ensemble averages (Tresp, 2001; Dietterich, 2000; Zhang & Ma, 2012) of GKR models, which were constructed using the 139 selected variable combinations.

We performed regression-based clustering analyses for all 139 selected variable combinations with 1000 initial randomized states. Using evaluation criteria similar to those for determining the number of clusters [formula (5)], the 200 best clustering results among these trials were selected to construct a committee machine that voted for the similarity between materials. The obtained affinity matrix for all the $Fm\bar{3}m$ AB materials is shown in Fig. 3(*a*). The similarity between each material pair varies from 0 to 1. A cell of the affinity matrix takes a value of 0 when the corresponding two materials are never included in the same cluster by a regression-based clustering model. In contrast, a cell of the affinity matrix takes a value of 1 when the corresponding two materials always appear in the same cluster according to every regression-based clustering model. Using this similarity, we could roughly divide

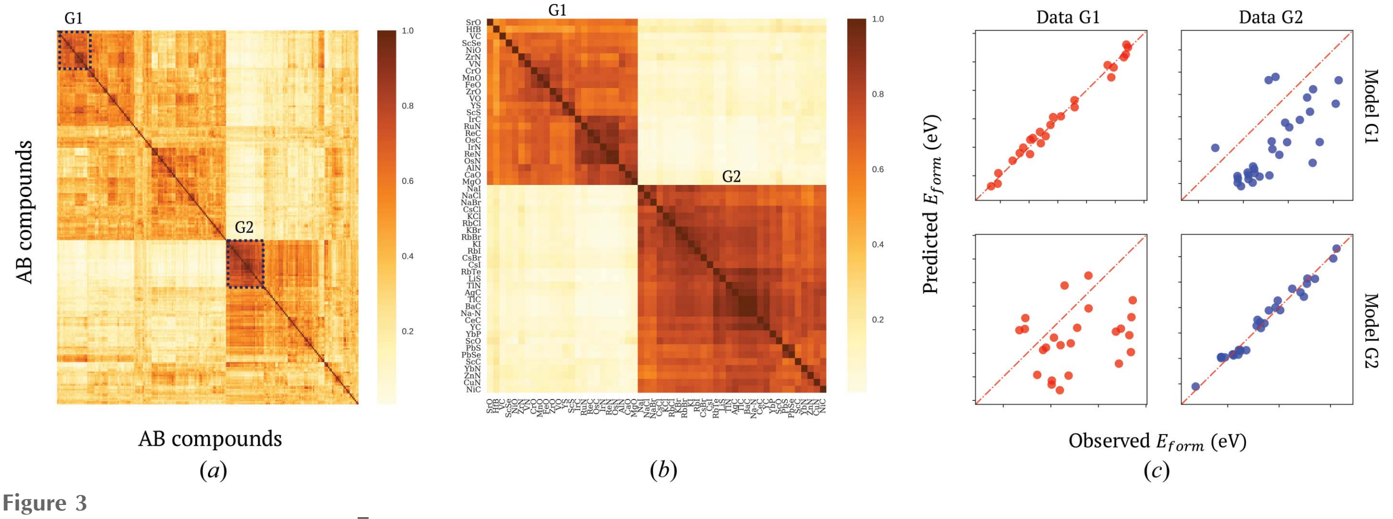


Figure 3
(*a*) The affinity matrix between the $Fm\bar{3}m$ AB materials yielded by the regression-based committee voting machine. (*b*) Enlarged views of highly similar elements in the G1 and G2 regions of the affinity matrix shown with dashed lines in panel (*a*). (*c*) Confusion matrices measuring linear similarities among materials in G1 and G2, as well as dissimilarities between models generated for materials in different groups.

Table 2
PA values for the $E_{\rm form}$, $L_{\rm const}$ and $T_{\rm C}$ prediction problems.

The results obtained with and without using the similarity measure (SM) information are shown for comparison.

		$E_{\rm form}$ (eV)		$L_{\rm const}$ (Å)		$T_{\rm C}$ (K)	
Prediction method		Without SM	With SM	Without SM	With SM	Without SM	With SM
GKR with all variables	R^2	0.929	0.954	0.982	0.986	0.893	0.929
	MAE	0.189	0.154	0.022	0.018	78.80	58.09
GKR with the best variable combination	R^2	0.967	0.978	0.989	0.992	0.968	0.988
	MAE	0.122	0.110	0.014	0.013	42.74	25.76
Ensemble of GKRs with top selected best variable combinations	R^2	0.972	0.982	0.991	0.992	0.974	0.991
	MAE	0.117	0.101	0.013	0.011	37.87	24.16

all the materials into two groups, as represented by the upper left and bottom right of Fig. 3(*a*).

Fig. 3(*b*) shows enlarged views of the affinity matrix for two groups of typical materials denoted G1 and G2. We can clearly see that the affinities between materials within each of the two groups, G1 and G2, exceed 0.7, showing high intra-group similarities. In contrast, the affinities between materials in different groups are smaller than 0.2, showing significant dissimilarity between G1 and G2. Further detailed investigation reveals that the materials in G1 are oxides, nitrides and carbides. The maximum common positive oxidation number of the *A* elements is greater than or equal to the maximum common negative oxidation number of the *B* elements for the compounds in this group. On the other hand, the materials in G2 are halides of alkaline metals, oxides, nitrides and carbides, for which the maximum common positive oxidation number of the *A* elements is less than or equal to the maximum common negative oxidation number of the *B* elements. Further investigation shows that only seven among 24 compounds in G1 have computed electronic structures with a band gap. In contrast, half of the compounds in G2 have computed electronic structures with a band gap. The obtained results suggest that the bonding nature of compounds in G1 is different from that of compounds in G2. The linearities between the target variable and the predicting variables for the two groups are summarized in Fig. 3(*c*). The diagonal plots show the correlations between the observed and predicted values for the target variables obtained using linear models of the predicting variables for the materials in the two groups. The off-diagonal plots show the correlations between the observed and predicted values for the target variables obtained using the linear models of the other groups. We could again confirm the intra-group similarity, and the dissimilarity between different groups, in terms of the linearity between the target and predicting variables for the compounds in the two groups.

To evaluate the validity of the analysis process quantitatively, we embedded the similarity measured by the committee machine into the regression of $E_{\rm form}$ of the $Fm\overline{3}m$ *AB* materials. To predict the value of the target variable for a new material, instead of using the entire available data set, we used only one third of the available materials having the highest similarity to the new material. It should again be noted that the similarity between the materials in the data set and the new material can be determined without knowing the value of the target physical property, using the value predicted by ensemble averaging of the nonlinear regression models.

Table 2 summarizes the PA in predicting $E_{\rm form}$ values of the $Fm\overline{3}m$ materials obtained using several regression models with

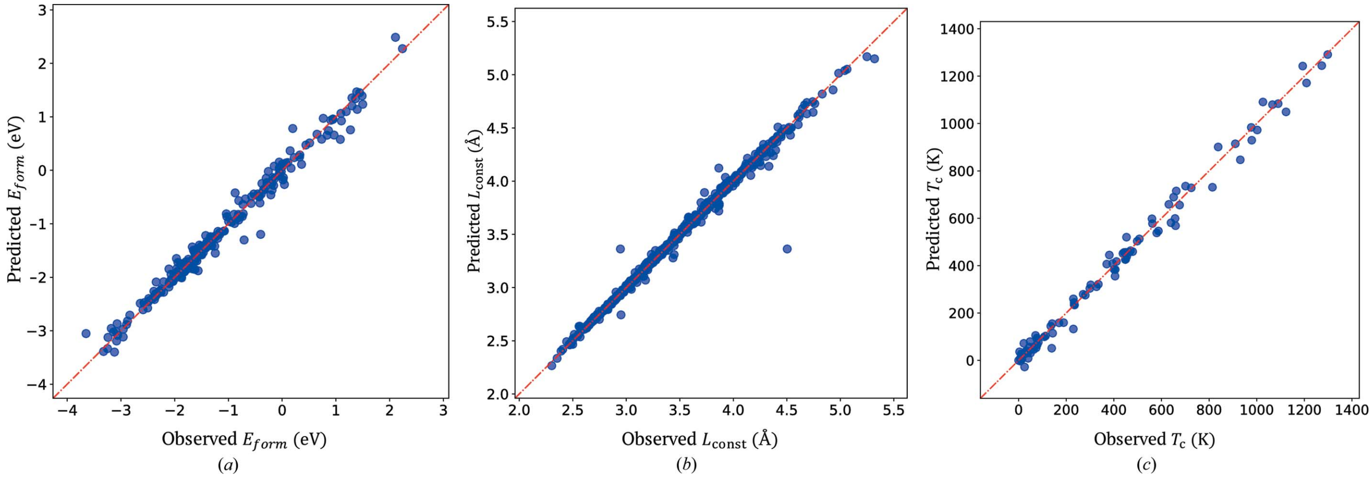


Figure 4
(From left to right) Observed and predicted target variables taking ensemble averaging of 139 ($E_{\rm form}$ problem), 57 ($L_{\rm const}$ problem) and 59 ($T_{\rm C}$ problem) best prediction models including similarity measure information. Ensemble models yield PAs with R^2 scores of 0.982 (MAE: 0.101 eV) for predicting the $E_{\rm form}$ problem, 0.992 (MAE: 0.011 Å) for predicting the $L_{\rm const}$ problem and 0.991 (MAE: 24.16 K) for predicting the $T_{\rm C}$ problem.

the designed predicting variables. The nonlinear model obtained using ensemble averaging of the best nonlinear regression models, having an R^2 score of 0.972 (MAE: 0.117 eV), could be improved significantly to an R^2 score of 0.982 (MAE: 0.101 eV) by considering the information from the similarity measurement (Fig. 4*a*). Therefore, the obtained results provide significant evidence to support our hypothesis that the similarity measured by the committee machine reflects the similarity in the actuating mechanisms of the target material physical property.

Table 3
The designed predicting variables describing the intrinsic properties of the constituent elements and the structural properties of the materials in the lattice parameter prediction problem.

A and *B* are elements of the binary *AB* b.c.c. materials.

Category	Predicting variables
Atomic properties of metals *A*	$r_{\mathrm{cov}A}$, m_A, Z_A, $n_{\mathrm{e}A}$, ℓ_A, χ_A, ρ_A
Atomic properties of metals *B*	$r_{\mathrm{cov}B}$, m_B, Z_B, $n_{\mathrm{e}B}$, ℓ_B, χ_B, ρ_B
Structural and additional information	ρ, d_χ, Sum_{AD}

3.2. Experiment 2: mining the quantum calculated lattice parameter for body-centred cubic structure data

In this experiment, a data set of 1541 binary *AB* body-centred cubic (b.c.c.) crystals with a 1:1 element ratio was collected from Takahashi *et al.* (2017). We focused on the computed lattice constant value L_{const} of the crystals. The *A* elements corresponded to almost all transition metals (Ag, Al, As, Au, Co, Cr, Cu, Fe, Ga, Li, Mg, Na, Ni, Os, Pd, Pt, Rh, Ru, Si, Ti, V, W and Zn) and the *B* elements corresponded to those with atomic numbers in the ranges of 1–42, 44–57 and 72–83. This data set included unrealistic materials such as the binary material AgHe, which incorporates He, an element that is known to possess a closed-shell structure and is, therefore, unlikely to form a solid.

To describe each material, we used a combination of 17 variables that related to basic physical properties of the *A* and *B* constituent elements, as summarized in Table 3. These chosen properties were as follows: (i) atomic radius (r_A, r_B); (ii) mass (m_A, m_B); (iii) atomic number (Z_A, Z_B); (iv) number of electrons in the outermost shell ($n_{\mathrm{e}A}$, $n_{\mathrm{e}B}$); (v) atomic orbital (ℓ_A, ℓ_B); and (vi) electronegativity (χ_A, χ_B). The atomic orbital values were converted from the categorical symbols *s*, *p*, *d*, *f* to numerical values representing the orbitals, *i.e.* 0, 1, 2, 3, respectively. To embed the structure information, four more properties were included: (vii) the density of atoms per unit volume (ρ_A, ρ_B); (viii) the unit-cell density ρ; (ix) the difference in electronegativity d_χ; and (x) the sum of the atomic orbital *B* and the difference in electronegativity Sum_{AD} (see Takahashi *et al.*, 2017).

A kernel regression-based variable selection with 3 × 10-fold cross-validation was performed to examine all combinations of the 17 variables. From the total number of screening variable combinations ($2^{17} - 1 = 131\,071$), we found 60 568 variable combinations for deriving regression models with R^2 scores exceeding 0.90 (Fig. 2). Among these, there were 57 variable combinations yielding regression models with R^2 scores exceeding 0.9895. The highest PA for this experiment is 0.989 (MAE: 0.014 Å), which was obtained using the combination $\{\rho, \ell_A, r_{\mathrm{cov}B}, m_A, m_B, \rho_B, n_{\mathrm{e}B}\}$. We could obtain a better PA with an R^2 score of 0.991 (MAE: 0.013 Å) by taking ensemble averaging of GKR models which derived from the 57 selected variable combinations. This result is a considerable improvement over the maximum PA (R^2 score: 0.90) of the support vector regression technique with the feature-selection strategy mentioned by Takahashi *et al.* (2017).

In the regression-based clustering analysis, the 57 selected variable combinations, accompanied by 1000 initial randomized states for each combination, were used to search for the most probable clustering results to construct the committee machine. The affinity matrix obtained for all materials is shown in Fig. 5(*a*), after rearrangement by a hierarchical clustering algorithm (Everitt *et al.*, 2011). Utilizing this similarity, we could roughly divide all materials in the data set into three groups, G1, G2 and G3. Further investigation revealed that most materials in G1 are constructed from two heavy transition metals. In contrast, the materials in G2 and G3 are constructed from a metal and a non-metal element, *e.g.* oxides and nitrides. For a given *A* element, L_{const} of the materials in G1 increases with the atomic number of the *B* element. On the other hand, L_{const} of the materials in G2 remains constant for materials sharing the same *A* element. Further, L_{const} for the materials in group G3 depends mainly on the electronegativity difference between the constituent elements *A* and *B*. Note that the materials in these three groups are visualized in detail in the supporting information. The linearities between the

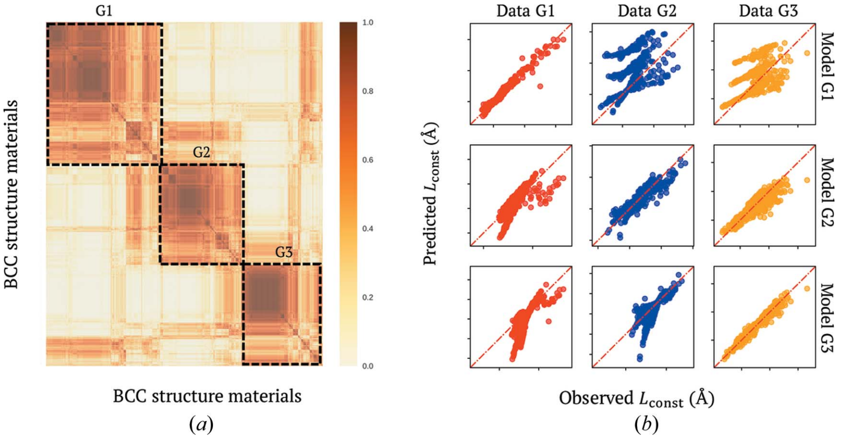


Figure 5
(*a*) The similarity matrix between materials for the L_{const} prediction problem yielded by the regression-based committee voting machine. This similarity matrix can be approximated as three disjoint groups of materials denoted G1, G2 and G3. (*b*) Confusion matrices measuring linear similarities among materials in each group, as well as dissimilarities between models generated for materials in different groups.

observed and predicting variables for these groups are shown in Fig. 5(*b*).

To predict the L_{const} of a new material, we applied the same strategy as that explained in the previous experiment. Table 2 summarizes the PA values obtained in our experiments. The nonlinear model obtained using ensemble averaging of the 57 best nonlinear regression models and having an R^2 score of 0.991 (MAE: 0.013 Å) could be marginally improved to an R^2 score of 0.992 (MAE: 0.011 Å) by including information from the similarity measurement (Fig. 4*b*).

3.3. Experiment 3: mining the experimentally observed Curie temperature data of rare earth–transition metal alloys

In this experiment, we collected experimental data related to 101 binary alloys consisting of transition and rare earth metals from the NIMS AtomWork database (Villars *et al.*, 2004; Xu *et al.*, 2011), which included the crystal structures of the alloys and their observed Curie temperatures T_{C}.

To represent the structural and physical properties of each binary alloy, we used a combination of 21 variables divided into three categories, as summarized in Table 4. The first and second categories contained predicting variables describing the atomic properties of the transition metal elements (T) and rare earth elements (R), respectively. The properties were as follows: (i) atomic number (Z_{R}, Z_{T}); (ii) covalent radius (r_{covR}, r_{covT}); (iii) first ionization ($\mathrm{IP_R}$, $\mathrm{IP_T}$); and (iv) electronegativity (χ_{R}, χ_{T}). In addition, predicting variables related to the magnetic properties were included: (v) total spin quantum number (S_{3d}, S_{4f}); (vi) total orbital angular momentum quantum number (L_{3d}, L_{4f}); and (vii) total angular momentum (J_{3d}, J_{4f}). For R metallic elements, additional variables $J_{4f}g_j$ and $J_{4f}(1-g_j)$ were added, because of the strong spin-orbit coupling effect. As in the two previous experiments, a third category variable was chosen which contained values calculated from the crystal structures of the alloys reported in the AtomWork database. The designed predicting variables included the transition (C_{T}) and rare earth (C_{R}) metal concentrations. Note that if we use the atomic percentage for the concentration, the two quantities are not independent. Therefore, in this work, we measured the concentrations in units of atoms Å^{-3}; this unit is more informative than the atomic percentage as it contains information on the constituent atomic size. As a consequence, (C_{T}) and (C_{R}) were not completely dependent on each other. Other additional structure variables were also added: the mean radius of the unit cell between two rare earth elements r_{RR}, between two transition metal elements r_{TT}, and between transition and rare earth elements r_{TR}. We set the experimentally observed T_{C} as the target variable.

Table 4
The designed predicting variables describing the intrinsic properties of the constituent elements and the structural properties in the T_{C} value prediction for the rare earth–transition metal alloys problem.

Category	Predicting variables
Atomic properties of transition metals	Z_{T}, r_{covT}, $\mathrm{IP_T}$, χ_{T}, S_{3d}, L_{3d}, J_{3d}
Atomic properties of rare earth metals	Z_{R}, r_{covR}, $\mathrm{IP_R}$, χ_{R}, S_{4f}, L_{4f}, J_{4f}, $J_{4f}g_j$, $J_{4f}(1-g_j)$
Structural information	C_{T}, C_{R}, r_{TT}, r_{TR}, r_{RR}

A kernel regression-based variable selection analysis was performed for these data using leave-one-out cross-validation. Among all the examined variable combinations, ($2^{21}-1=2\,097\,151$), we found 84 870 combinations for which the corresponding GKR models exhibited R^2 scores exceeding 0.90 (Fig. 2). Among these, there were 59 variable combinations yielding GKR models associated with R^2 scores exceeding 0.95. These predicting variable combinations were selected for the next analysis step. The highest PA in this experiment was 0.968 (MAE: 42.74 K), obtained using the combination $\{C_{\mathrm{R}}, Z_{\mathrm{R}}, Z_{\mathrm{T}}, \chi_{\mathrm{T}}, r_{\mathrm{covT}}, L_{3d}, J_{3d}\}$. We could obtain a better PA with an R^2 score of 0.974 (MAE: 37.87 K) by applying ensemble averaging to the GKR models, which were

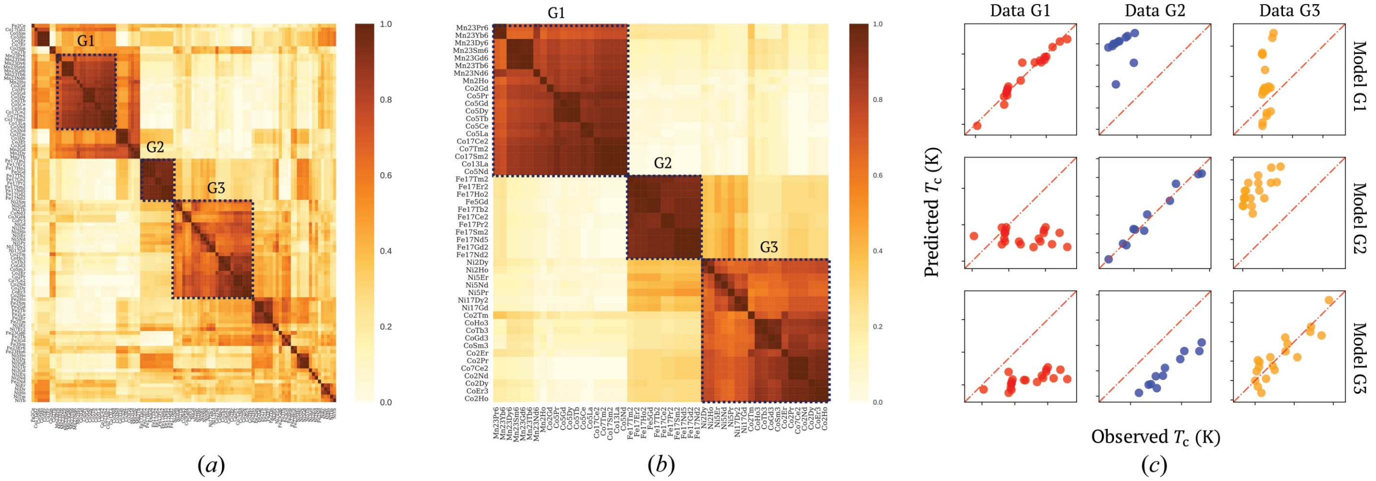


Figure 6
(*a*) The similarity matrix between the rare earth–transition metal alloys yielded by the regression-based committee voting machine. (*b*) Enlarged views of highly similar elements in the G1, G2 and G3 regions of the similarity matrix shown with dashed lines in panel (*a*). (*c*) Confusion matrices measuring linear similarities among alloys in each group as well as dissimilarities between models generated for alloys in different groups.

derived from the selected 59 variable combinations. We considered these variable combinations as candidates for the next step of the analysis.

In the regression-based clustering analysis, 59 variable combinations with 1000 initial randomized states were used to search for the most probable clustering results to construct the committee machine to vote for the similarity between the alloys. The obtained affinity matrix for all the alloys is shown in Fig. 6(*a*). An enlarged view of the three groups of alloys having high similarity (denoted G1, G2 and G3) is shown in Fig. 6(*b*). Further investigation revealed that G1 includes Mn- and Co-based alloys with high T_C, *e.g.* $Mn_{23}Pr_6$ (448 K), $Mn_{23}Sm_6$ (450 K), Co_5Pr (931 K) and Co_5Nd (910 K). Other low-T_C Co-based alloys, *e.g.* Co_2Pr (45 K) and Co_2Nd (108 K), are counted as having higher similarity to the Ni-based alloys in G3, *e.g.* Ni_5Nd (7 K) and Ni_2Ho (16 K). In contrast, G2 includes all the Fe-based $Fe_{17}RE_2$ alloys, where RE represents different rare earth metals. To confirm the value of our similarity measure, Fig. 6(*c*) shows the linearities between the observed and predicting variables for these groups, as well as the dissimilarities among these groups.

In the next analysis step, we utilized the obtained similarity measure to predict T_C for a new material using the same strategy as in the two previous experiments. The nonlinear model obtained using ensemble averaging of the best nonlinear regression models and having an R^2 score of 0.974 (MAE: 37.87 K) could be improved significantly to attain an R^2 score of 0.991 (MAE: 24.16 K) utilizing the information from the similarity measurement (Fig. 4*c* and Table 2). The obtained results provide significant evidence to support our hypothesis that the similarity voted for by the committee machine indicates the similarity in the actuating mechanisms of the T_C of the binary alloys.

4. Conclusions

In this work, we have proposed a method to measure the similarities between materials, focusing on specific physical properties, to describe and interpret the actual mechanism underlying a physical phenomenon in a given problem. The proposed method consists of three steps: variable evaluation based on nonlinear regression, regression-based clustering, and similarity measurement with a committee machine constructed from the clustering result. Three data sets of well characterized crystalline materials represented by key atomic predicting variables were used as test beds. The formation energy, lattice parameter and Curie temperature were considered as target physical properties of the examined materials. Our experiments show that rational and meaningful group structures can be obtained with the help of the proposed approach. The similarity measure information helped significantly increase the prediction accuracy for the material physical properties. Through use of ensemble top kernel ridge prediction models, the R^2 score increased from 0.972 to 0.982 for the formation energy prediction problem, and from 0.974 to 0.991 for the Curie temperature prediction problem after utilizing the similarity information. However, no significant improvement in the the R^2 score was observed for the lattice constant prediction problem. Thus, our results indicate that our proposed data analysis flow can systematically facilitate further understanding of a given phenomenon by identifying similarities among materials in the problem data set.

Funding information

This work was partly supported by PRESTO and by the Materials Research by Information Integration Initiative (MI^{2}I) project of the Support Program for Start-Up Innovation Hub, from the Japan Science and Technology Agency (JST), and by JSPS KAKENHI Grant-in-Aid for Young Scientists (B) (grant No. JP17K14803), Japan.

References

Almuallim, H. & Dietterich, T. G. (1991). *The Ninth National Conference on Artificial Intelligence*, pp. 547–552. Menlo Park: AAAI Press.

Behler, J. & Parrinello, M. (2007). *Phys. Rev. Lett.* **98**, 146401.

Biesiada, J. & Duch, W. (2007). *Computer Recognition Systems 2. Advances in Soft Computing*, Vol. 45. Heidelberg: Springer.

Blei, D. M. (2012). *Commun. ACM*, **55**, 77–84.

Blum, A. L. & Langley, P. (1997). *Artif. Intell.* **97**, 245–271.

Botu, V. & Ramprasad, R. (2015). *Int. J. Quantum Chem.* **115**, 1074–1083.

Dietterich, T. G. (2000). *Proceedings of the First International Workshop on Multiple Classifier Systems*, 21–23 June 2000, Cagliari, Italy. *Lecture Notes in Computer Science*, Vol. 1857, edited by J. Kittler and F. Roli, pp. 1–15. Heidelberg: Springer.

Duangsoithong, R. & Windeatt, T. (2009). *Machine Learning and Data Mining in Pattern Recognition*, edited by Petra Perner, pp. 206–220. Heidelberg: Springer.

Einbeck, J., Evers, L. & Bailer-Jones, C. (2008). *Principal Manifolds for Data Visualization and Dimension Reduction. Lecture Notes in Computational Science and Engineering*, Vol. 58, edited by A. N. Gorban, B. Kégl, D. C. Wunsch and A. Zinovyev, pp. 178–201. Heidelberg: Springer.

Everitt, S., Landau, S., Leese, M. D. & Stahl (2011). Editors. *Cluster Analysis*, 5th ed., ch. 4, *Hierarchical Clustering. Wiley Series in Probability and Statistics*. Chichester: Wiley.

Fernandez, M., Boyd, P. G., Daff, T. D., Aghaji, M. Z. & Woo, T. K. (2014). *J. Phys. Chem. Lett.* **5**, 3056–3060.

Fukunaga, K. & Olsen, R. (1971). *IEEE Trans. Comput.* **C-20**, 1615–1616.

Ghiringhelli, M., Vybiral, J., Levchenko, V., Draxl, C. & Scheffler, M. (2015). *Phys. Rev. Lett.* **114**, 105503.

Goldsmith, B. R., Boley, M., Vreeken, J., Scheffler, M. & Ghiringhelli, M. (2017). *New J. Phys.* **19**, 013031.

Hastie, T., Tibshirani, R. & Friedman, J. H. (2009). Editors. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

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

Jain, A., Shin, Y. & Persson, A. (2016). *Nat. Rev. Mater.* **1**, 15004.

Jones, R. O. (2015). *Rev. Mod. Phys.* **87**, 897–923.

Jones, R. O. & Gunnarsson, O. (1989). *Rev. Mod. Phys.* **61**, 689–746.

Kanungo, T., Mount, M., Netanyahu, S., Piatko, D., Silverman, R. & Wu, A. Y. (2002). *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 881–892.

Kohavi, R. (1995). *IJCAI'95 – Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 20–25 August 1995, Montreal, Canada, Vol. 2, pp. 1137–1143. San Francisco: Morgan Kaufmann Publishers.

Kohavi, R. & John, H. (1997). *Artif. Intell.* **97**, 273–324.

Kohn, W. & Sham, L. J. (1965). *Phys. Rev.* **140**, A1133–A1138.

Kusne, G., Keller, D., Anderson, A., Zaban, A. I. & Takeuchi, I. (2015). *Nanotechnology*, **26**, 444002.

Kvalseth, T. O. (1985). *Am. Stat.* **39**, 279–285.

Landauer, T. K., Foltz, P. W. & Laham, D. (1998). *Discourse Process.* **25**, 259–284.

Le, T. V., Epa, V. C., Burden, F. R. & Winkler, A. (2012). *Chem. Rev.* **112**, 2889–2919.

Liu, H. & Yu, L. (2005). *IEEE Trans. Knowl. Data Eng.* **17**, 491–502.

Liu, Y., Zhao, T., Ju, W. & Shi, S. (2017). *J. Materiomics*, **3**, 159–177.

Lloyd, S. P. (1982). *IEEE Trans. Inf. Theory*, **28**, 129–137.

Lu, W., Xiao, R., Yang, J., Li, H. & Zhang, W. (2017). *J. Materiomics*, **3**, 191–201.

Lum, Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J. & Carlsson, G. (2013). *Sci. Rep.* **3**, 1236.

MacQueen, J. (1967). *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, *Statistics*, pp. 281–297. University of California Press.

Murphy, K. P. (2012). Editor. *Machine Learning: A Probabilistic Perspective*. MIT Press.

Opitz, D. & Maclin, R. (1999). *JAIR*, **11**, 169–198.

Picard, R. R. & Cook, D. (1984). *J. Am. Stat. Assoc.* **79**, 575–583.

Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. & Ramprasad, R. (2013). *Sci. Rep.* **3**, 2810.

Rajan, K. (2015). *Annu. Rev. Mater. Res.* **45**, 153–169.

Rupp, M. (2015). *Int. J. Quantum Chem.* **115**, 1058–1073.

Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. (2013). *JOM*, **65**, 1501–1509.

Settles, B. (2010). Computer Sciences Technical Report No. 1648. University of Wisconsin-Madison, USA.

Seung, H. S., Opper, M. & Sompolinsky, H. (1992). *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 27–29 July 1992, Pittsburgh, Pennsylvania, USA, pp. 287–294. New York: ACM.

Smith, J. S., Isayev, O. & Roitberg, A. E. (2017). *Chem. Sci.* **8**, 3192–3203.

Snyder, J. C., Rupp, M., Hansen, K., Müller, K. & Burke, K. (2012). *Phys. Rev. Lett.* **108**, 253002.

Srinivasan, S., Broderick, R., Zhang, R., Mishra, A., Sinnott, B., Saxena, K., LeBeau, M. & Rajan, K. (2015). *Sci. Rep.* **5**, 17960.

Stone, M. (1974). *J. R. Stat. Soc. Ser. B (Methodological)*, **36**, 111–147.

Sumpter, B. G., Vasudevan, R. K., Potok, T. & Kalinin, S. V. (2015). *NPJ Comput. Mater.* **1**, 15008.

Takahashi, K., Takahashi, L., Baran, J. D. & Tanaka, Y. (2017). *J. Chem. Phys.* **146**, 011002.

Tibshirani, R. (1996). *J. R. Stat. Soc. Ser. B (Methodological)*, **58**, 267–288.

Tresp, V. (2001). *Neural Comput.* **12**, 2000.

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.

Vidal, R., Ma, Y. & Sastry, S. (2015). *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 1945–1959.

Villars, P., Berndt, M., Brandenburg, K., Cenzual, K., Daams, J., Hulliger, F., Massalski, T., Okamoto, H., Osaki, K., Prince, A., Putz, H. & S. Iwata (2004). *J. Alloys Compd.* **367**, 293–297.

Xu, Y., Yamazaki, M. & Villars, P. (2011). *Jpn. J. Appl. Phys.* **50**, 11RH02.

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S. & Stoica, I. (2016). *Commun. ACM*, **59**, 56–65.

Zhang, C. & Ma, Y. (2012). *Ensemble Machine Learning: Methods and Applications.* Heidelberg: Springer.