Formation enthalpies for transition metal alloys using machine learning

Shashanka Ubaru, Agnieszka Międlar, Yousef Saad, and James R. Chelikowsky

Phys. Rev. B **95**, 214102 — Published 1 June 2017

DOI: 10.1103/PhysRevB.95.214102

Formation enthalpies for transition metal alloys using machine learning

Shashanka Ubaru[1], Agnieszka Międlar[2], Yousef Saad[1], and James R. Chelikowsky[3]
[1]*Department of Computer Science and Engineering, University of Minnesota,*
Twin Cities, MN 55455, USA. [ubaru001,saad]@umn.edu.
[2]*Department of Mathematics, University of Kansas,*
Lawerence, KS 66045-7594, USA. amiedlar@ku.edu.
[3]*Center for Computational Materials, Institute for Computational Engineering and Science,*
and Departments of Physics and Chemical Engineering,
University of Texas, Austin, TX 78712, USA. jrc@utexas.edu.

The enthalpy of formation is an important thermodynamic property. Developing fast and accurate methods for its prediction is of practical interest in a variety of applications. Material informatics techniques based on machine learning have recently been introduced in the literature as an inexpensive means of exploiting materials data, and can be used to examine a variety of thermodynamics properties. We investigate the use of such machine learning tools for predicting the formation enthalpies of binary intermetallic compounds that contain at least one transition metal. We consider certain easily available properties of the constituting elements complemented by some basic properties of the compounds, to predict the formation enthalpies. We show how choosing these properties (input features) based on a literature study (using prior physics knowledge) seems to outperform machine learning based feature selection methods such as sensitivity analysis and LASSO (Least Absolute Shrinkage and Selection Operator) based methods. A nonlinear kernel based support vector regression method is employed to perform the predictions. The predictive ability of our model is illustrated via several experiments on a dataset containing 648 binary alloys. We train and validate the model using the formation enthalpies calculated using a model by Miedema, which is a popular semi-empirical model used for the prediction of formation enthalpies of metal alloys.

I. INTRODUCTION

The thermodynamic data of alloys such as the standard enthalpy of formation ΔH (also known as standard heat of formation) plays an important role in several applications, e.g., in the calculation of phase diagrams and materials design, in the exploration of new materials having high melting points that can be used in advanced coal-fired plants, building heat-exchangers, filters, and turbines, and many more. The heat of formation of an alloy indicates its stability, *i.e.*, a more negative enthalpy of formation implies a more stable alloy. Also, the sign of ΔH is a fundamental property that can serve as an indicator for the stability of a given alloy. Systems with a positive ΔH are only stabilized by entropy considerations. In addition, the formation enthalpies of compounds are also significant for certain high-throughput density functional theory (DFT) calculations[1]. Unfortunately, it is well known that determining such thermodynamic properties via experiments is difficult, especially for compounds with high melting points.

Since the experimental determination of thermodynamic properties of a vast combinations of elements is inefficient, recent research has focused on developing various computational approaches to predict and estimate these properties of interest. In the case of the enthalpies of formation of compounds, several different approaches have been proposed over the years. For example, we note the Hildebrand formula[2] for enthalpy of solutions, a semi-empirical model of alloy cohesion by Miedema *et al.*[3], and a modified embedded atom model for random alloys[4]. Popular among these, particularly for binary metal alloys, is Miedema's model.

In a series of papers[3,5–7], Miedema and his coauthors developed a semi-empirical method for predicting the heat of formation of binary intermetallic compounds that contain at least one transition metal. They showed that the formation enthalpies of such binary alloys can, in general, be described in terms of a simple atomic model, that depends only on two parameters of the constituent atoms. Their model has been very successful in predicting correctly the signs for the heats of formation. However, it is less quantitative for predicting the magnitude of the enthalpy change and requires certain experimental information.

With the advent of density functional theory and its concurrent implementations for realistic computations[8,9], using first principles or *ab initio* calculations for predicting and understanding material properties has become popular[10–12]. One can compute accurate values for the formation enthalpies of compounds using such calculations. Also, some comparative studies between the Miedema model predictions and the *ab initio* calculations for transition-metal compound formation now exist[13,14]. However, a major drawback of DFT calculations is the relative high computational cost, especially for a quick screening of a large database, and the need for certain prior information such as a known crystal structure.

In recent years, as a result of the Material Genome Initiative[15], machine learning (ML) techniques have emerged among other 'material informatics' methods, for exploiting materials data. A popular approach in the literature is to apply tools from machine learning on certain DFT calculations to accelerate prediction of various properties of compounds[16–22]. Ideas from machine learning have been coupled with databases of *ab initio* calculations to estimate molecular electronic properties in chemical compound space, including the enthalpy of formation of compounds[23,24]. However, these methods still have the major disadvantage of requiring results from many DFT calculations, which may not be possible for alloys without given crystal structures, *i.e.*, amorphous or noncrystalline alloys. Recently, a machine learning approach to predict the density functional theory total energies has been implemented and these predictions are used to compute the enthalpies of formation of metal-nonmetal compounds[1].

Our paper presents an alternative machine learning approach to predict the formation enthalpies of binary metal alloys. The method we propose differs from previous ML techniques in that it uses readily available properties of the constituting elements (elemental properties), complemented by some basic properties of the compounds that are available in popular databases (e.g., Materials Project[25]), to predict the formation enthalpies.

A large set of (publicly available) elemental properties is considered and three different methods are explored to select (a smaller set of) appropriate elemental properties for enthalpy prediction from this large set. The three sets of elemental properties used are: (i) properties selected based on a literature study, (ii) properties obtained through sensitivity analysis. (iii) properties selected by a modified LASSO (Least Absolute Shrinkage and Selection Operator) method[26–28]. The first set can be viewed as a set selected based on prior physics knowledge, while the latter two are based on machine learning methods (do not take into account any physics knowledge), these methods are defined in Section III A. Our results indicate that features (elemental properties) selected based on the prior physics knowledge perform better in predicting enthalpies than those obtained through machine learning techniques.

A well-known method exploited in machine learning and known as "Support Vector Regression" is employed for the formation enthalpy predictions. The approach proposed in this work is fast and *does not require DFT calculations*, since the model takes available properties of elements and compounds as input, and is trained and validated against (or reproduces) the formation enthalpies calculated using Miedema's model, which are also easily available for many binary alloys. Since the Miedema's model is itself not very accurate, the proposed machine learning approach cannot give highly accurate formation enthalpies. However, the presented method is an extremely inexpensive technique aimed at predicting formation enthalpies of new compounds (as accurately as Miedema's model) without any empirical information. Such enthalpy predictions suffices in many applications such as new material discovery, stability analysis and melting point predictions. In applications where accurate formation enthalpies are required, these predictions can be coupled with simple DFT calculations (which are less expensive than full DFT calculations taking elemental properties as an input) to obtain accurate enthalpies. This is a popular approach used to improve Miedema's model predictions[13,14].

Section II briefly describes the Miedema's model for prediction of enthalpy of formation. The two key components of our approach, the feature selection method and the machine learning model are discussed in Section III. Experimental results with accompanied analysis, discussion and final conclusions are presented in Section IV. Appendix provides some additional details.

II. STANDARD ENTHALPY OF FORMATION

The standard enthalpy of formation $\Delta H \left[\frac{\text{kJ}}{\text{mol}}\right]$ of a compound, also known as the standard heat of formation, measures the change of enthalpy during the formation of 1 mole of the compound from the individual constituting elements. Formation enthalpies play a fundamental role in predicting the thermodynamical stability of new materials. For example, they are crucial in evaluating the performance of Li-ion batteries[29,30], in designing materials for chemical hydrogen storage[31] and in modeling the formation energies of metal oxides[32,33].

Although the advent of DFT made calculations of enthalpies of formation possible[34], the calculated values of ΔH are available only for a limited number of compounds[7,35]. As such, we focus on the Miedema model for predicting the enthalpy of formation given by

$$\Delta H \propto f(c)\big(-P(\Delta\phi^*)^2 + Q(\Delta n_{ws}^{1/3})^2\big), \qquad (1)$$

where $\Delta\phi^*$ denotes the *difference in the work functions* of the two metals, $\Delta n_{ws}^{1/3}$ the *difference in electronic densities* at the boundary of the Wigner-Seitz cell of the pure metals, $f(c)$ is an unknown function of concentration, and P, Q are empirical constants. Miedema's model assumes that the formation enthalpy depends on the two parameters, ϕ^* and $n_{ws}^{1/3}$. The first parameter arises from the charge transfer between neighboring cells which is proportional to $\Delta\phi^*$, and accounts for attractive forces within the compound. The second parameter arises from a surface tension term, proportional to $\Delta n_{ws}^{1/3}$, which accounts for repulsive forces. Note that a slightly modified formulation of the formation enthalpy is needed for alloys involving a transition metal and one of the polyvalent non-transition metals, namely,

$$\Delta H \propto f(c)\big(-P(\Delta\phi^*)^2 + Q(\Delta n_{ws})^2 - R\big), \qquad (2)$$

with R being a constant.

The work function ϕ^* characterizes the electronegativity parameter or the chemical potential for electronic charge. Since the work function ϕ^* can be hard to compute, it is replaced by an experimental work function ϕ[7]. A problem with this substitution is that various experimental values have been reported for the work function, and it is not known how to select the best one. Also, the work function can depend on the nature of the surface structure of the elemental crystal.

Obtaining values for n_{ws} can also be problematic, depending on the anisotropy of the elemental bonding. In Miedema's model, this value is approximately calculated as[5]

$$(n_{ws})^2 = \frac{B}{V},$$

where B is the *experimental bulk modulus* and V is the *molar volume* of pure metals. The computation of the above ratio may be an issue owing to inaccurate or missing experimental data. In many cases, the above equation is used to predict the bulk modulus of elements[36]. The constants P and Q depend on the type of metals that are present in the alloy[37]; their values are not universal[7]. Thus, using Miedema's model for predicting the formation enthalpies of new compounds not only require certain experimental results, but may also yield unreliable results due to the variations in these constants.

Here, we present a non-empirical method to rapidly predict the formation enthalpies of binary transition metal alloys (including their signs) using machine learning techniques.

III. MACHINE LEARNING FOR PREDICTION

In this work, we apply well-known supervised regression techniques to predict properties of compounds that are hard and expensive to compute otherwise, using easily available physical, chemical and structural properties of the compounds, known as *features* in machine learning or *descriptors* in material science. In many cases, the atomic and elemental properties of the constituting atoms of the compounds are included as input features. The performance of these machine learning predictions depends primarily on the following two aspects: the *feature selection* and the *machine learning model* used.

A. Features Selection

A quintessential step for successful predictions is identifying the key characteristics of the constituting elements (elemental features), that dictate or affect the properties of the compounds that we wish to predict. In this paper, we consider three different approaches for feature selection.

a. Literature study: In order to identify a good set of elemental features that influence the formation enthalpies of compounds, let us first look at Miedema's model[3]. It has been known for a long time[38] that the work function ϕ is correlated to the ionization energy, the electron affinity and the electronegativity of constituting elements. While ionization energy and electron affinity are properties of isolated atoms, the electronegativity provides information about the attraction the given atom has for electrons in an ionic (or partially ionic) bond formed with another atom. For pure metals, the theoretical electron density values n_{ws} depends on bulk modulus B and molar volume V. Thus, Miedema's model suggests that the following features of the constituting elements are crucial for the prediction of the formation enthalpies: *ionization energy*, *electronegativity*, *electron density* and *molar volume.*

Another model which helps to identify the elemental features that affect the formation enthalpies, is the Hildebrand formula for the enthalpy of solution of two liquids[2]. This formula depends on two properties of the constituting liquids, namely, the *enthalpy of vaporization* of the liquids and their *molar volumes*. The development of Miedema's model was influenced by this formula[6]. The formation enthalpies describe the cohesion in the metal alloys[7]. The modified embedded atom method by Ouyang *et al.*[4] uses the *cohesive energy*, *formation energy* and *atomic volumes* of pure elements to describe the work function ϕ in Miedema's model. From the above studies, we expect that the following seven elemental properties are likely to be the most influential features in predicting the formation enthalpy: *ionization energy*, *electron affinity*, *electronegativity*, *electron density*, *enthalpy of vaporization*, *cohesive energy* and *molar volume*.

b. Sensitivity method: A machine learning approach to identify the elemental features that provide good property predictions is to use the *'sensitivity method'* described by Saad *et al.*[39]. To verify the impact of elemental features on the enthalpy prediction accuracy, we find the sensitivity of each of the available properties of the constituting atoms (we collected $d' = 49$ properties of each element, see Appendix D, and hence obtained $d = 2d' = 98$ features in total after concatenation to represent the binary alloys).

The sensitivity method applied to our model can be described as follows: Let $X \in \mathbb{R}^{n\times d}$ be a matrix that contains the known properties (the input features/descriptors) of the individual compounds as columns (since X is a concatenation of the d' properties of the two elements forming a compound, the number of columns is $d = 2d'$). First, for a considered feature k, we perturb the values of this feature for both elements of each compound, i.e., the vectors $X(:,k)$ and $X(:,k+d')$ are perturbed respectively by $\varepsilon \approx c10^{-8}\|[X(:,k); X(:,k+d')]\|$, where c is a random number.

Second, we calculate a new coefficient vector a_ε, using the least squares solution $a_\varepsilon = (X^\top X)^{-1}X^\top v$, where v is a vector containing the actual formation enthalpies of the compounds. Next, the norm of the difference between the original (obtained without perturbing columns of X) and the perturbed coefficient vector $\|a_\varepsilon - a\|$ is computed.

Finally, the ratio $\frac{\|a_\varepsilon - a\|}{\varepsilon}$ is assigned as the sensitivity measure of the k-th feature. The top seven most sensitive features for the prediction of formation enthalpies are : the *electrochemical equivalent weight*[40], *first oxidation state*, *group number*, *effective nuclear charge* (Slater's rule), *metal radius*, *electronegativity* and *distance core electron*.

c. LASSO method: Another alternative method used recently in the literature[27,28] for feature selection is the so called *compressed sensing approach*, which is a LASSO[26] type method. Given a large feature matrix $X \in \mathbb{R}^{n\times d}$, and the output vector v (property to be predicted), the LASSO method yields a sparse relation between X and v by solving the convex optimization problem

$$\arg\min_{\beta\in\mathbb{R}^d} \|v - X\beta\|_2^2 + \lambda\|\beta\|_1, \tag{3}$$

where the ℓ_1-norm $\Big(\|\beta\|_1 = \sum_i \beta(i)\Big)$ promotes the sparsity in vector β. Thus, the sparsity of vector β helps us to select the descriptors (columns of X) that best describe v in the least squares sense. However, recall that the matrix X is formed by simply concatenating the properties of the two constituting elements. Using the LASSO method directly will not guarantee selection of the same set of properties for the two elements. That is, the vector β need not have same nonzero coordinates in the first $d' = 49$ coordinates ($\beta(1:49)$) and last d' coordinates ($\beta(50:98)$). We indeed obtained different sets of features being selected for the two elements when the LASSO method was used directly in our experiments. In order to overcome this issue, we propose the following *modified LASSO problem* obtained by splitting vector β as $\beta = [\beta_1; \beta_2]$,

$$\min_{\beta\in\mathbb{R}^d} \|v - X\beta\|_2^2 + \mu\|\beta_1 - \beta_2]\|_2^2 + \lambda\|\beta\|_1$$
$$\text{or} \quad \min_{\beta\in\mathbb{R}^d} \|v - X\beta\|_2^2 + \mu\|J\beta\|_2^2 + \lambda\|\beta\|_1,$$

where $J = [I, -I]$ with the identity matrix I. We include the additional term $\mu\|J\beta\|_2$ to ensure that the two halves of the vector β are close (equal), such that the same set of properties is selected for the two elements (from the first 49 and the last 49 features). This modified LASSO problem is still a convex optimization problem and therefore can be easily solved using any of the available optimization packages, e.g. the CVX package[41,42]. The parameters λ and μ were adjusted such that the modified LASSO selects exactly seven properties from both elements, i.e., both β_1 and β_2 have exactly seven nonzero entries. The following seven properties were selected by the LASSO method for the two elements: *atomic weight, density, energy ionization first, temperature boiling, temperature melting, electronegativity* and *bulk modulus.* The modified LASSO method for property selection is also robust, i.e., changing slightly the parameters λ and μ does not give different set of features.

Since the feature matrix $X \in \mathbb{R}^{n\times d}$ consists of two subsets (first 49 and the remaining 49 features) corresponding to the two constituting elements, we can assume that the d features are divided into two groups and use either the group LASSO[43] or the sparse group LASSO[44] methods to select appropriate features from these two groups. However, these methods will not guarantee the selection of the same set of properties from the two groups (for the two constituting elements). Consequently, we will still have to include the additional constraint term $\mu\|J\beta\|_2^2$ proposed above in the optimization objective.

In the presence of compound features, the *Pearson product-moment correlation coefficient* (r)[45] can be used to determine the correlation between two given properties of a compound. That is, given the values of properties $x = \{x_1, \dots, x_n\}$ and $y = \{y_1, \dots, y_n\}$ for each of the n compounds in the dataset, the *Pearson correlation coefficient* is:

$$r = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{4}$$

with $\bar{x} := \frac{1}{n}\sum\limits_{i=1}^{n} x_i$ (analogously for $\bar{y}$). Table II lists the Pearson correlation coefficients between the standard enthalpy of formation and several other compound properties, for the binary alloys in our database.

B. Machine Learning Model

In this work, we use a supervised learning regression method to predict the formation enthalpies of binary metal alloys.

Given n compounds and d specific features (descriptors) we build a matrix $X \in \mathbb{R}^{n\times d}$ that stores the features of each compound as a column of X. We assume that a certain property being studied, e.g., enthalpy of formation, is known for each of the n compounds. We are now presented with a new compound, which is not among the n ones already studied, and whose same d features, as those of the data, are known and stored in a vector $z \in \mathbb{R}^d$. Regression methods attempt to answer the question: “What is our best guess of the enthalpy of formation for this new compound?” Regression methods use X to build a mapping that will yield the desired property from z. In the simplest case of linear regression, this mapping is just a linear combination of the values of the features, and the coefficients of the linear combination are extracted by solving a least squares problem that involves X and the right-hand side of the properties of the n compounds.

Linear regression is often too simple model, and is rarely used to predict complex physical properties. A common and efficient regression technique used for real world data applications is the *support vector regression* or SVR[4647]. In SVR, the idea of *support vector machines* (SVM) developed by Vapnik and Chervonenkis[48] is extended to handle regression problems[49]. SVR is a nonlinear regression technique that employs kernels to implicitly map the inputs into high-dimensional (nonlinear) feature spaces. The details of the SVR method are given in Appendix A.

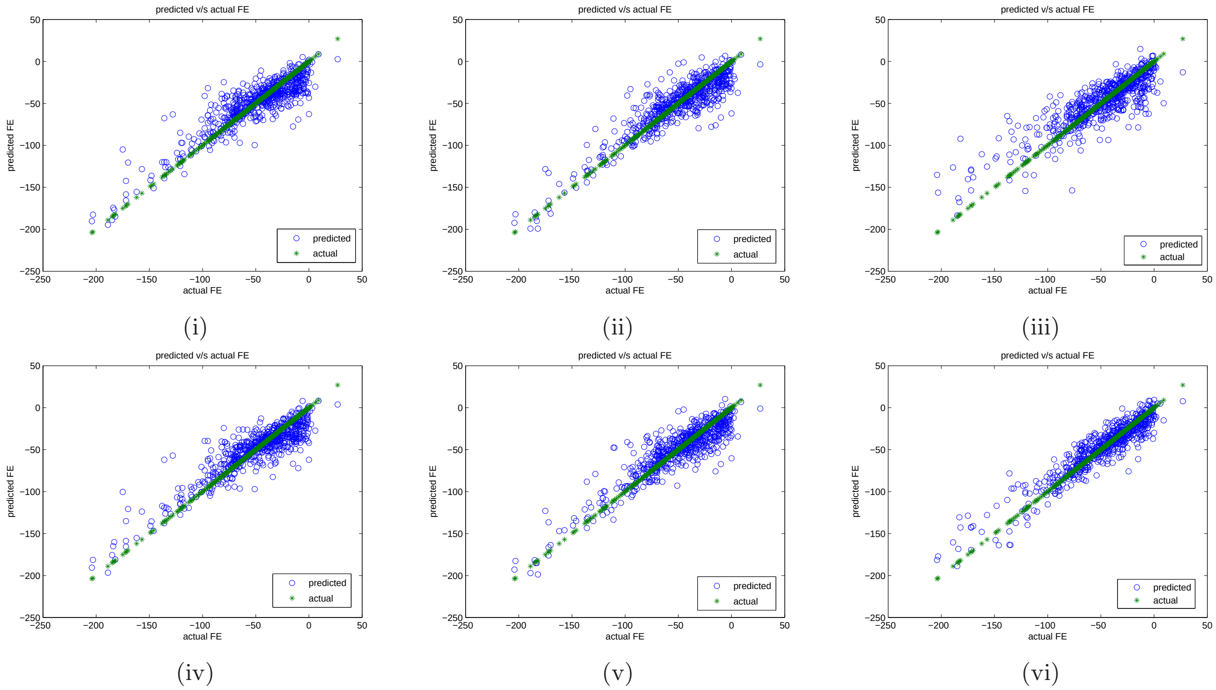
Since the relation between the elemental properties and the desired thermodynamic property of the compound is typically highly nonlinear, in this work, we consider a nonlinear kernel based regression method. A variety of support vector machine methods for regression have been developed in the literature, see e.g.,[47,49–53]. Among these, the most suitable SVR variant for our purposes, is the ε-SVR method with RBF or Gaussian kernels given by

$$k(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|^2\right),$$

see Appendix A for details. For our experiments, we consider the ε-SVR method implemented in the **libSVM** MATLAB library[54]. For the optimal γ value in the kernel, we sweep from 0.1 to 1 with increments of 0.1 and choose the value that yields the best results (smallest error). In Appendix A, we also provide a justification for this choice of the regression method by comparing the prediction performance of SVR against several other popular regression techniques.

IV. RESULTS AND DISCUSSION

Here, we present our results for the prediction of formation enthalpy for transition metal alloys using the support vector regression (SVR) model. The results obtained using other regression methods are reported in the Appendix A. We found that, SVR outperforms the other regression approaches.



(i) (ii) (iii)

(iv) (v) (vi)

FIG. 1. Predictions of enthalpies of formation obtained using (i) the elemental properties from the literature, (ii) the elemental properties from the sensitivity analysis, (iii) the elemental properties from the modified LASSO, (iv) the elemental properties from the literature and the compound properties, (v) the elemental properties from the sensitivity analysis and the compound properties, and (vi) the elemental properties from the modified LASSO with the compound properties.

TABLE I. Relative errors in standard enthalpy of formation predictions for different feature sets. MAE: Mean Absolute Error; RMSE: Root Mean Square Error; MRRE: Mean-Regularized Relative Error; NRE: Net Relative Error. See Appendix A 2 for details.

Feature Set	MAE	RMSE	MRRE	NRE	R^2
Literature	1.3809	5.5598	0.0157	0.0286	0.9563
Sensitivity	1.7657	5.7145	0.0195	0.0365	0.9468
LASSO	4.6838	9.0660	0.1049	0.2004	0.6858
Literature+compound	1.3682	5.4965	0.0156	0.0283	0.9556
Sensitivity+compound	1.6422	5.5695	0.0181	0.0340	0.9508
LASSO+compound	2.2960	6.9060	0.0580	0.1096	0.8704

Standard Enthalpy of Formation for Transition Metal Alloys

To illustrate the use of machine learning tools for the prediction of the enthalpy of formation for binary metal alloys, we considered 648 transition metal alloys whose formation enthalpies are available[7]. These formation enthalpies are computed using the Miedema *et al.* model. Details about these compounds are given in Appendix C.

Previously, we discussed feature selection. Collecting such features/properties of the constituting elements is the first step of the prediction. We acquired 49 different chemical properties of all the elements from the Database on Properties of Chemical Elements[55], see Appendix D for more details. Next, six different physical properties of the 648 compounds (compound features) were collected from the Materials Project database[56][57].

These six properties were: *band gap, number of atoms per unit cell (nsite), volume, magnetic moment, density* and *energy-per-atom* (energy normalized to per atom in the unit cell), see[57] (The Materials API). We also collected six different crystal properties of these 648 compounds from the same database, namely the three *unit cell dimensions* a, b, c and the three *unit cell angles* α, β, γ. Various experiments were conducted using these data features. Figure 1 and Table I present the results obtained from these experiments for the prediction of the formation enthalpies of these 648 transition metal alloys using the support vector regression method and various feature sets.

As mentioned in Section III A, we considered three approaches to select the appropriate elemental features (feature selection) that affect the formation enthalpies of the metal alloys the most. The first set of features was based on the literature study, and we refer to this set of features as the 'literature set'. In this set, we considered 7 elemental properties of the two constituting elements of the binary alloys as the input features (14 values in total), namely, *ionization energy*, *electronegativity*, *electron density*, *enthalpy of vaporization*, *cohesive energy*, *electrochemical equivalent weight* and *molar volume*. The order of concatenation of features is done based on the atomic number. The features of the element with smaller atomic number are chosen as first 7 columns of the feature matrix. Concatenating the elemental features does not incorporate the stoichiometric information (the ratios of the individual elements in the compound). We feed this information to the regression model as two new features. That is, we include two additional features as inputs, whose values are the ratios of the first and the second element of the compound, respectively. For example, for compound ScGe, the values of these two features will be $[0.5, 0.5]$, and for $\texttt{ScGe}_2$, their values will be $[0.33, 0.67]$. Thus, we consider 16 features in total. In practice, we need to choose only one of these two stoichiometric features since the other seems redundant.

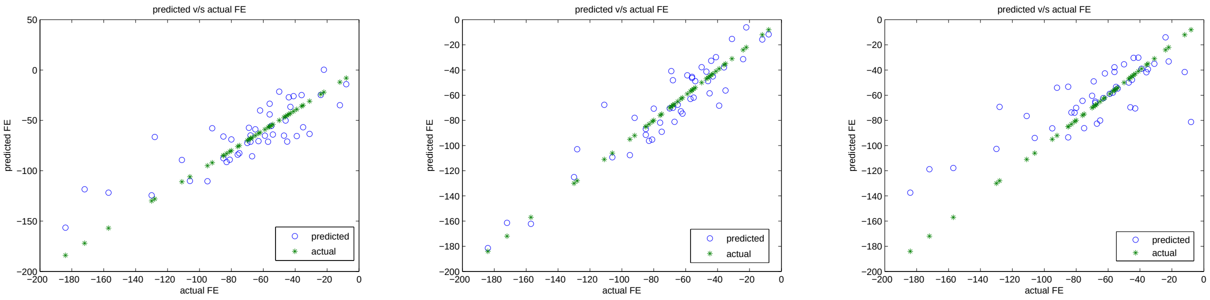
Figure 1(i) presents the formation enthalpies predicted by the SVR model against the actual formation enthalpies (obtained from[7]) using the literature set of elemental properties as input features. The errors obtained for this experiment are listed in Table I. Details about different error measures can be found in the Appendix A. We used a 10 fold cross-validation method to predict the formation enthalpies of the 648 compounds. That is, we repeated the experiments 10 times with 10% of the dataset (around 65 compounds) chosen at random without replacement from the 648 compounds used as test data. Hence, after the 10 trials we have all 648 alloys' formation enthalpies predicted once by the model. These predicted values of the test data are those presented in the figure. Note that we do not present the formation enthalpies predicted for the training data since these predictions are typically good. A good prediction for training data does not indicate that the model has a good prediction ability, since such a model might perform poorly for a given test dataset.

TABLE II. Pearson's correlation coefficients between the compound features and the formation enthalpy vs. sensitivities of compound features.

Name of the feature	Correlation r	Sensitivity
magnetic moment	0.2010	0.5842
energy-per-atom	0.1558	0.3961
density	0.1025	0.2108
n-cell-length-c	0.0885	0.8734
nsite	0.0644	0.6565
n-cell-length-a	0.0615	1.7345
band gap	0.0350	8.5279

The second set of features considered was based on the sensitivity method[39] discussed in Section III A, and we denote this set as 'sensitivity set'. In this set, we considered 7 features selected from the 49 elemental properties using the sensitivity method. These features are expected to significantly influence the prediction accuracy. According to the sensitivity model, the seven most effective features in predicting the formation enthalpy are: *electrochemical weight equivalent*, *oxidation state first*, *group number*, *nuclear charge effective*, *radii metal*, *electronegativity* and *distance core electron*. Figure 1(ii) presents the formation enthalpies predicted by the SVR model, using the sensitivity set of elemental features (16 in total, including the two stoichiometric features), against the actual formation enthalpies.

The third set of features was selected based on the modified LASSO method described in Section II. We call this set a 'LASSO set'. The parameters μ and λ were adjusted such that the same seven features were selected for both the elements (same nonzero entries in β_1 and β_2). The selected features were: *atomic weight, density, energy ionization first, temperature boiling, temperature melting, electronegativity* and *bulk modulus*. Figure 1(iii) presents the formation enthalpies predicted by the SVR model, using the LASSO set of elemental features (16 in total),



(i) (ii) (iii)

FIG. 2. Predictions of enthalpies of formation of $\mathtt{Sc}$ binary alloys obtained using (i) the elemental properties from the literature, (ii) the elemental properties from sensitivity analysis, (iii) elemental properties from the modified LASSO method.

TABLE III. Predicted and actual formation enthalpies (FE) of $\mathtt{Sc}$ binary alloys using the 14(+2) elemental properties (literature set) and compound properties as input features. Values in $\left[\frac{\mathrm{kJ}}{\mathrm{mol}}\right]$.

Chemical formula	Actual FE	Predicted FE	Chemical formula	Actual FE	Predicted FE
	best			worst	
$\mathtt{Sc_5Ge_3}$	-75	-75	$\mathtt{ScBe_5}$	-31	-58
$\mathtt{Sc_3Ga_2}$	-63	-63	$\mathtt{Sc_3In}$	-39	-69
$\mathtt{ScCd}$	-55	-56	$\mathtt{ScN}$	-184	-152
$\mathtt{Sc_5Sn_3}$	-76	-78	$\mathtt{ScIr}$	-92	-60
$\mathtt{ScAl}$	-68	-66	$\mathtt{Sc_3P_2}$	-157	-110
$\mathtt{ScGe}$	-85	-88	$\mathtt{ScP}$	-172	-124

against the actual formation enthalpies.

For the following numerical experiments, we considered the compound features (6 physical and 6 crystal properties of the alloys) along with the elemental and stoichiometric information as the input features for the SVR model. Figure 1(iv) presents predicted versus actual formation enthalpies obtained using collectively the literature and the compound feature sets ($16+12=28$ in total). Similarly, Figure 1(v) presents the results obtained when the sensitivity and the compound feature sets were used in the SVR model. The results obtained when the LASSO and the compound feature sets were used together are presented in Figure 1(vi). The various error measures obtained for each of these six experiments are listed in Table I.

To complete our investigations, we also tried to extend the size of our input feature sets by considering some (six) prototypical functions of the features, namely, $x, x^2, x^3, \sqrt{|x|}, \log(1+|x|)$, and e^x, where x represents the given feature. That is, we expanded the 14 initial features to $14 \times 6 = 84$ features using the above functions. This heuristic was previously used in some material informatics literature[27,28,58]. However, we did not observe any significant improvements in the resulted predictions after applying such heuristics since we are already using a nonlinear kernel. In article[58] these nonlinear functions are coupled with a nonlinear kernel method. Results when such nonlinear functions of the features were used with the LASSO feature selection method are discussed in the Appendix.

In order to understand the influence of the compound features on the formation enthalpy, we additionally computed Pearson's correlation coefficients r, defined in (4), between the 12 aforementioned compound properties and the formation enthalpies of the compounds. Table II contains the top seven most correlated features along with the calculated Pearson's correlation values. For the sake of completeness, we also present the associated sensitivities. An interesting observation here is that, the sensitivity value obtained for unit cell length a is almost twice of the sensitivity of unit cell length c. This makes sense since the volume of a compound $V \propto a^2c$, and volume is an important property that influences the formation enthalpy of a compound. This shows how some of the physical interactions are captured by ML methods.

The aforementioned experimental results lead to the following observations. Firstly, we note that the three feature selection methods select three different sets of features with little overlap. This shows that: a) there are multiple sets of elemental features that are likely to influence the formation enthalpy of the alloys; b) the machine learning features are not the same as those selected based on a-priori knowledge of underlying physics; c) the two machine learning feature sets also differ. Our main observation is that predictions based on the literature set (based on prior knowledge) are better than the ones obtained using the machine learning sets. Clearly, the fourth feature set (literature+compound) yields the best results amongst all the experiments. This shows that coupling actual knowledge of relevant physics (domain knowledge) with machine learning provides improved performance. This is likely because the machine learning methods attempt to find a linear relation between the features and the target property. However, the actual relation between the different properties of a compound will typically be highly nonlinear. Hence, we observe that coupling prior physics (domain) knowledge with machine learning methods tend to give better results than using pure machine learning features. We also observe that the different machine learning methods do not yield same results (do not agree with each other). The ranking based on sensitivities does not match the one based on Pearson's correlation coefficients. Moreover, the features selected by the sensitivity method differs from the ones selected by the modified LASSO.

SVR Model's Predictive Ability

One of the primary goals of developing new techniques for predicting properties of compounds is the hope to identify compounds with desired properties or to predict some unknown properties of existing compounds. In this experiment, we examine such predictive ability of our SVR based model by predicting the formation enthalpies of several new compounds. Let us assume that all compounds containing the element `Sc` (scandium) are unknown to our SVR model, i.e., we set aside all compounds containing `Sc` as a test dataset and put all other compounds into the training set. Element `Sc` was chosen since we have 45 binary alloys containing `Sc` in our initial dataset (which is a good number of instances for testing), and also because the values of the formation enthalpy of these compounds lie across a wide range $[-181, -6]$, making it a good test set. Once the model is trained on the remaining 603 compounds, we predicted the formation enthalpies (FE) of the 45 `Sc` binary alloys. Similar experiment results with other elements are presented in Appendix B.

The corresponding results are presented in Figure 2 and Figure 3 of Appendix B. The plots display predicted FE values for `Sc` binary alloys using for Figure 2(i) the elemental properties (literature set). The results obtained using the elemental properties (sensitivity analysis), and the elemental properties (modified LASSO method) are presented in Figure 2(ii)–(iii), respectively. The results obtained when the compound properties were coupled with these sets of elemental properties are presented in Appendix B. In all six test cases the features accounting for stoichiometric values were also included.

Table III (and tables V- VI in Appendix B) list the compounds' chemical formula, the predicted and the actual formation enthalpy values of the top six closest (best) predictions (left side) and the bottom six farthest (worst) predictions (right side) for the case of `Sc` binary alloys using the 14(+2) elemental properties (literature set) and the compound properties. Tables for the elemental properties obtained with sensitivity analysis and the modified LASSO method are given in Appendix B. When using 14(+2) elemental properties (literature set) and the compound properties as input features we have 56% of the 45 enthalpy predictions within the mean-regularized relative error of 0.1 (10% relative error) and 71% within 0.15 (15% relative error), whereas for the elemental properties (sensitivity analysis) and the compound properties we have 73% of the 45 enthalpy predictions within the mean-regularized relative error of 0.1 (10% relative error) and 89% within 0.15 (15% relative error). For the sake of completeness, analogous statistics for other 3d-, 4d-, 5d-transition, actinide and noble metals are presented in Table VII of Appendix B.

We observe that the values of formation enthalpies predicted by the SVR model are very close to the values obtained using Miedema's model and the "worst" predictions in Tables III–V include some alloys of `Sc` with heavy elements, i.e., `Bi`, `Pd` and `Ir`. This experiment illustrates the ability of our SVR model to predict formation enthalpies of the new compounds.

Acknowledgements

Authors would like to thank Peter Solfest and Zahra Forootaninia for their help in the experiments and acknowledge support provided by the Scientific Discovery through Advanced Computing (SciDAC) program funded by U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research and Basic Energy Sciences under award number DE-SC0008877.

[1] A. Deml, R. O'Hayre, C. Wolverton, and V. Stevanović, Phys. Rev. B **93**, 085142 (2016).

[2] R. Scott and R. Hildebrand, New York: Reinhold Publ. Corp (1950).

[3] A. Miedema, F. de Boer, and P. de Chatel, J. Phys. F: Metal Phys. **3**, 1558 (1973).

[4] Y. Ouyang, B. Zhang, S. Liao, Z. Jin, and H. Chen, Trans. Nonferrous Met. Soc. China **8**, 60 (1998).

[5] A. Miedema, F. De Boer, and R. Boom, CALPHAD **1**, 341 (1977).

[6] A. Miedema, P. De Chatel, and F. De Boer, Physica B+C **100**, 1 (1980).

[7] F. De Boer, W. Mattens, R. Boom, A. Miedema, and A. Niessen, *Cohesion in metals: Transition Metal Alloys (Cohesion and Strucure)* (North-Holland, Amsterdam, 1988).

[8] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).

[9] W. Kohn and L. Sham, Phys. Rev. **140**, A1133 (1965).

[10] J. Ihm, A. Zunger, and M. Cohen, J. Phys. C **12, 4409** (1979), erratum: J. Phys. C 13, 3095 (1980).

[11] M. Yin and M. Cohen, Phys. Rev. Lett. **45, 1004** (1980).

[12] A. Becker, Phys. Rev. A **38, 3098** (1988).

[13] A. Williams, C. Gelatt, and V. Moruzzi, Phys. Rev. Lett. **44**, 764 (1980).

[14] J. Chelikowsky, Phys. Rev. B **25**, 6506 (1982).

[15] `https://www.mgi.gov/`.

[16] A. Oganov and C. Glass, J. Chem. Phys. **124**, 244704 (2006).

[17] Z.-K. Liu, L.-Q. Chen, and K. Rajan, JOM **58**, 42 (2006).

[18] G. Hautier, C. Fischer, A. Jain, T. Mueller, and G. Ceder, Chem. Mater. **22**, 3762 (2010).

[19] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, Scientific Reports **3** (2013).

[20] A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka, Phys. Rev. B **89**, 054303 (2014).

[21] P. Dey, J. Bible, S. Datta, S. Broderick, J. Jasinski, M. Sunkara, M. Menon, and K. Rajan, Comput. Mater. Sci. **83**, 185 (2014).

[22] J. Lee, A. Seko, K. Shitara, and I. Tanaka, (2015).

[23] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. von Lilienfeld, New J. Phys. **15**, 095003 (2013).

[24] A. Teixeira, J. Leal, and A. Falcao, J. Cheminform. **5** (2013).

[25] `https://materialsproject.org/`.

[26] R. Tibshirani, Journal of the Royal Statistical Society. Series B (Methodological) , 267 (1996).

[27] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Physical review letters **114**, 105503 (2015).

[28] C. Kim, G. Pilania, and R. Ramprasad, Chemistry of Materials **28**, 1304 (2016).

[29] V. Chevrier, S. Ong, R. Armiento, M. Chan, and G. Ceder, Phys. Rev. B **82** (2010).

[30] J. Bhattacharya and C. Wolverton, J. Electrochem. Soc. **161**, A1440 (2014).

[31] J. Yang, A. Sudik, C. Wolverton, and D. Siegel, Chem. Soc. Rev. **39**, 656 (2010).

[32] A. Deml, V. Stevanović, C. Muhich, C. Musgrave, and R. O'Hayre, Energy Environ. Sci. , 1996 (2014).

[33] A. Deml, A. Holder, R. O'Hayre, C. Musgrave, and V. Stevanović, J. Phys. Chem. Lett. **6**, 1948 (2015).

[34] R. Martin, *Electronic Structure: Basic Theory and Practical Methods, 1st ed.* (Cambridge University Press, Cambridge, United Kingdom, 2008).

[35] O. Kubaschewski, C. Alcock, and P. Spencer, *Materials Thermochemistry, 6th ed.* (Pergamon Press, New York, 1993).

[36] C. Li, Y. Chin, and P. Wu, Intermetallics **12**, 103 (2004).

[37] P. Ray, M. Akinc, and M. Kramer, in *Proceeding of 22nd Annual Conference on Fossil Energy Materials* (2008).

[38] F. Rother and H. Bomke, Z. Phys. **86**, 231 (1933).

[39] Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. Chelikowsky, and W. Andreoni, Phys. Rev. B **85**, 104104 (2012).

[40] The electrochemical equivalent weight of an element is the ratio between its atomic weight and its principal valence number.

[41] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," `http://cvxr.com/cvx` (2014).

[42] M. Grant and S. Boyd, in *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, edited by V. Blondel, S. Boyd, and H. Kimura (Springer-Verlag Limited, 2008) pp. 95–110, `http://stanford.edu/~boyd/graph_dcp.html`.

[43] M. Yuan and Y. Lin, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**, 49 (2006).

[44] J. Friedman, T. Hastie, and R. Tibshirani, arXiv preprint arXiv:1001.0736 (2010).

[45] K. Pearson, in *Proc. R. Soc. Lond.*, Vol. 58 (1895) pp. 240–242.

[46] SVR was primarily developed at AT&T Bell Laboratories by Vapnik and co-workers for industrial purposes. Hence, SVR has been particularly effective for practical data applications[50].

[47] V. Vapnik, *The Nature of Statistical Learning Theory* (Springer Science & Business Media, New York, 2013).

[48] V. Vapnik and A. Chervonenkis, Autom. Remote Control **25**, 103 (1964).

[49] H. Drucker, C. Burges, L. Kaufman, A. Smola, and V. Vapnik, in *Advances in Neural Information Processing Systems* (1997) pp. 155–161.

[50] A. Smola and B. Schölkopf, Stat. Comput. **14**, 199 (2004).

[51] B. Schölkopf and C. Burges, *Advances in Kernel Methods: Support Vector Learning* (MIT Press, 1999).

[52] C. Cortes and V. Vapnik, Machine Learning **20**, 273 (1995).

[53] A. Smola, *Regression estimation with support vector learning machines*, Master's thesis, Technische Universität München (1996).

[54] C.-C. Chang and C.-J. Lin, ACM Transactions on Intelligent Systems and Technology **2**, 27:1 (2011), software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[55] `http://phases.imet-db.ru/elements/main.aspx`.

[56] `https://materialsproject.org/`.

[57] A. Jain, S. Ong, G. Hautier, W. Chen, W. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. Persson, APL Materials **1**, 011002 (2013).

[58] G. Pilania, A. Mannodi-Kanakkithodi, B. Uberuaga, R. Ramprasad, J. Gubernatis, and T. Lookman, Scientific Reports **6**, 19375 (2016).

[59] B. Boser, I. Guyon, and V. Vapnik, in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory* (ACM, 1992) pp. 144–152.

[60] B. Schölkopf, A. Smola, and K.-R. Müller, Neural Comput. **10**, 1299 (1998).

[61] "Database on Properties of Chemical Elements," (2014), russian Academy of Sciences, A. A. Baikov Institute of Metallurgy and Materials Science. The last database update: 21 February 2014, `http://phases.imet-db.ru/elements/main.aspx`.

Appendix A: Supervised Learning Regression Methods

In this section, we provide additional details of the SVR method used in this paper. We also compare the prediction performance of SVR against other popular regression methods.

1. Support Vector Regression

As mentioned in Section III B, in this work, we employ the nonlinear kernel ε-SVR method with RBF kernels. Initially, the support vector machines (SVM) were combined with the kernels to obtain nonlinear classifications[59]. This idea was later extended to the regression problem by introducing an alternative loss function[49,53]. We consider the ε- insensitive loss function, or ε-SVR[50], which is a popular SVR method. Here, the objective is to compute a function $f(x)$ that has deviations at most ε away from the target training points v_i. In the linear case, the function f is given by,

$$f(x) = \langle w, x\rangle + b, \tag{A1}$$

where $w \in \mathbb{R}^d$ are the weights. In the simplest case, the ε-SVR can be written as a convex optimization problem:

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\|w\|^2 \\ &\text{subject to} && |v_i - \langle w, x\rangle - b| \le \varepsilon \end{aligned} \tag{A2}$$

where x_i are the rows of the feature matrix X and v_i are components of target vector v. The standardized version of SVR also includes slack variables[52]. The above optimization problem is usually solved using its Lagrangian dual and quadratic programming or interior point methods, see[47,50] for details.

As discussed earlier, the relation between the elemental features and the predicted property is expected to be highly nonlinear. The above SVR can be made nonlinear by using implicit mapping and nonlinear kernels[50]. The SVR algorithm only depends on the inner (dot) products between the feature vectors x_i. Hence, we can define the mapping to the kernel space implicitly by simply replacing the dot products as $\langle x_i, x_j\rangle \to k(x_i, x_j)$. Then, the function f becomes,

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) k(x_i, x) + b,$$

where α_i, α_i^* are the Lagrange multipliers from the dual problem.

Only certain types of kernels $k(\cdot,\cdot)$ that satisfy the Mercer's condition[60] can be used (called admissible SV kernels). Many admissible SV kernels have been proposed in the literature[47,50,51]. In this work, we consider the popular RBF (Radial Basis Function) or Gaussian kernels given by

$$k(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|^2\right),$$

where γ is the scaling factor.

TABLE IV. Comparison of different regression techniques.

Method name	MAE	RMSE	MRRE	NRE	R^2
LR	23.5445	32.1228	0.2621	0.4885	0.1792
LR ($L1$ reg.)	23.5575	32.0789	0.2626	0.4892	0.1979
SVR	**6.6816**	**11.4038**	**0.0677**	**0.1398**	**0.8752**
RR.	22.8695	31.6553	0.2457	0.4745	0.2174
PLS	23.3401	31.3801	0.2593	0.4843	0.2631
KRR (Laplacian)	8.2970	15.3008	0.1395	0.1889	0.7638
KRR (Gaussian)	10.4969	18.0065	0.1841	0.3241	0.6888

2. Comparison of Machine Learning Methods

The primary reason for choosing SVR in this work is because SVR outperforms other popular regression methods in predicting the formation enthalpies of compounds. In this section, we present the following experiment, which justifies our choice of the SVR method. First, we consider five popular regression techniques for predicting the properties of materials, namely : Support Vector Regression (SVR), as implemented in the libSVM Matlab library[54], using ε-SVR method with radial basis functions; Partial Least Squares (PLS), available as Matlab built-in function; Linear Regression (LR), Linear Regression with L1-Regularization (LR-reg), Robust Regression (RR), Kernel Ridge Regression (KRR) with Laplacian and Gaussian Kernels.

Table IV presents the performance of each of these five regression techniques in predicting the formation enthalpies of the 648 compounds in our dataset, using five different evaluation measures. In almost all of our experiments, Support Vector Regression (SVR) method performed significantly better compared to the other methods. Therefore, due to the consistently superior performance of SVR, results from other regression methods are not reported in the extensive experimental results.

The performance evaluation of analyzed regression techniques, and the various input features were assessed using the following five error measures:

1. Mean Absolute Error (MAE)

$$\frac{1}{n}\sum_{i=1}^{n}|v_i - \widehat{v}_i|,$$

where v is the vector of actual formation enthalpies and $\widehat{v}$ is the vector of predicted formation enthalpies.

2. Root Mean Square Error (RMSE)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_i - \widehat{v}_i)^2}.$$

3. Mean-Regularized Relative Error (MRRE)

$$\frac{1}{n}\sum_{i=1}^{n}\frac{|\widehat{v}_i - v_i|}{|\bar{v}| + |v_i|},$$

where $\bar{v} = \sum_{i=1}^{n} v_i/n$ is the mean value of v. We use regularized relative error since some v_i can be zero.

4. Net Relative Error (NRE)

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{i=1}^{n}|v_i - \widehat{v}_i|}{n\cdot|\bar{v}|}.$$

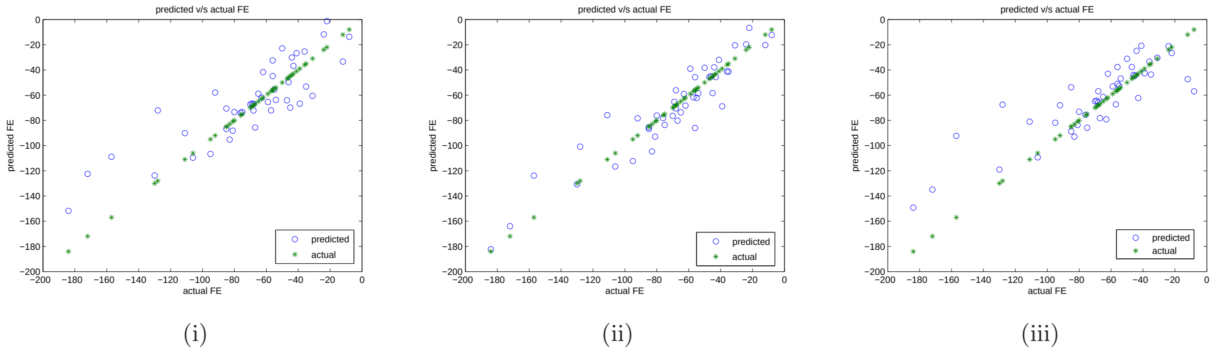
5. R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(v_i - \widehat{v}_i)^2}{\sum_{i=1}^{n}(v_i - \bar{v})^2} = \frac{\sum_{i=1}^{n}(\hat{v}_i - \bar{v})^2}{\sum_{i=1}^{n}(v_i - \bar{v})^2}.$$

In the linear case, this measure is equivalent to the ratio between the explained sum of squares (also called regression sum of squares) and the total sum of squares (proportional to the variance).

Note that except for R^2, the error measure closer to zero indicates better performance. For R^2 the desired value is 1, indicating a perfect prediction.

d. Feature selection with nonlinear functions: Since the relations between the elemental features and the formation enthalpy are likely to be nonlinear, in the LASSO feature selection method, we also included nonlinear functions of the features. Along with the 98 features (from 49 elemental properties), we used the six prototypical functions of the features, namely, $x, x^2, x^3, \sqrt{|x|}, \log(1+|x|)$, and e^x, where x represents the given feature. The idea of combining LASSO with such functions of features were used in previous literature, for example[27,28,58]. Although the modified LASSO method chose a few nonlinear functions of the features are best representatives, the same set of features and functions were not selected for the two elements (we tired tuning the parameters λ and μ with a range of values, but in vein). More importantly, the FE predictions (using both LASSO and RBF SVR methods) we obtained from these (nonlinear) features were quite poor. Hence, we have not reported these results.



(i)

(ii)

(iii)

FIG. 3. Predictions of enthalpies of formation of Sc binary alloys obtained using (i) elemental properties from the literature, (ii) the elemental properties from the sensitivity analysis, (iii) elemental properties from the modified LASSO method and the compound properties.

TABLE V. Predicted and actual formation enthalpies (FE) of Sc binary alloys using the elemental properties (sensitivity analysis) and the compound properties as input features. Values in $\left[\frac{\mathrm{kJ}}{\mathrm{mol}}\right]$.

Chemical formula	Actual FE	Predicted FE	Chemical formula	Actual FE	Predicted FE
	best			worst	
$ScPd_3$	-85	-86	$ScAl_2$	-59	-39
ScAs	-130	-131	ScBi	-83	-105
$ScZn_2$	-46	-45	ScPd	-128	-101
$ScAl_3$	-47	-46	Sc_3In	-39	-69
ScN	-184	-182	Sc_2C	-56	-86
ScGe	-85	-87	Sc_3P_2	-157	-124

TABLE VI. Predicted and actual formation enthalpies (FE) of Sc binary alloys using the elemental properties (LASSO analysis) and the compound properties as input features. Values in $\left[\frac{\mathrm{kJ}}{\mathrm{mol}}\right]$.

Chemical formula	Actual FE	Predicted FE	Chemical formula	Actual FE	Predicted FE
	best			worst	
Sc5Sn3	-76	-76	ScPd3	-85	-53
ScGa3	-45	-44	ScN	-184	-149
ScBe5	-31	-30	ScMn2	-12	-47
ScZn2	-46	-44	ScP	-172	-135
ScGa	-68	-66	ScMg	-8	-56
ScCd	-55	-53	ScPd	-128	-67

Appendix B: Prediction Ability Results

In this section, we present additional results related to the prediction ability experiments. Figure 3 plots the predictions of enthalpies of formation of Sc binary alloys when the compound properties were coupled with three sets of elemental properties. Tables V and VI list the best and worst predictions for the case of Sc binary alloys using the 14(+2) elemental properties (sensitivity and LASSO sets) and the compound properties. Table VII lists the percentage of compounds in the dataset containing a particular element whose enthalpy prediction is within the mean-regularized relative error of 0.1 (10% relative error) when predicted using our SVR model with literature feature set.

TABLE VII. Percent of compounds in the dataset containing a particular element whose enthalpy prediction is within the mean-regularized relative error of 0.1 (10% relative error) when using elemental properties (sensitivity analysis) and the compound properties as input features.

TABLE VIII. 3d-, 4d-, 5d-transition metals

Sc (73% of 45)	Ti (34% of 35)	V (28% of 32)	Cr (17% of 18)	Mn (6% of 32)	Fe (34% of 35)	Co (50% of 36)	Ni (11% of 38)
Y (58% of 38)	Zr (64% of 42)	Nb (56% of 36)	Mo (21% of 34)	Tc (22% of 9)	Ru (38% of 26)	Rh (70% of 33)	Pd (40% of 45)
La (48% of 33)	Hf (70% of 47)	Ta (65% of 34)	W (11% of 19)	Re (28% of 18)	Os (61% of 28)	Ir (59% of 34)	Pt (51% of 45)

TABLE IX. actinide and noble metals

Th (18% of 11)	U (50% of 6)	Pu (100% of 1)

Cu (0%of 7)
Ag (43% of 7)
Au (14% of 7)

TABLE X. List of 3d-, 4d-, 5d-transition metals.

Sc	Ti	V	Cr	Mn	Fe	Co	Ni
Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd
La	Hf	Ta	W	Re	Os	Ir	Pt

TABLE XI. List of non-transition metals.

Li	Be		B	C	N
Na	Mg		Al	Si	P
K	Ca	Zn	Ga	Ge	As
Rb	Sr	Cd	In	Sn	Sb
Cs	Ba	Hg	Tl	Pb	Bi

Appendix C: Dataset of 648 Compounds[7]

We consider the dataset[7] (Chapter III) of binary alloys based on each of the 3d-, 4d- or 5d-transition metals sequentially, according to their position in the periodic table (row-wise) including only one rare-earth metal **La**, see Table X. We then predict the enthalpies of formation for ordered compounds with the following compositions of metal **A** with transition and non-transition metal, respectively: $\mathbf{AX}_5$, $\mathbf{AX}_3$, $\mathbf{AX}_2$, $\mathbf{A}_3\mathbf{X}_5$, $\mathbf{A}_2\mathbf{X}_3$, $\mathbf{AX}$, $\mathbf{A}_3\mathbf{X}_2$, $\mathbf{A}_5\mathbf{X}_3$, $\mathbf{A}_2\mathbf{X}$, $\mathbf{A}_3\mathbf{X}$ and $\mathbf{A}_5\mathbf{X}$. The alloying partner metals **X** are arranged as follows: the 3d-, 4d- and 5d-transition metals as in Table X, the actinide metals **Th**, **U** and **Pu**, and the noble metals **Cu**, **Ag** and **Au**. Due to identical parameters for **Y** and **Gd**, only **Y** is considered. The non-transition partner metals are grouped according to equal valency and similar alloying behavior in the periodic table (column-wise), see Table XI. The values for other compositions can be easily interpolated from the obtained predictions. Out of the 648 binary alloy compounds, there are 416 alloys with unique combinations of elements. We validated our predictions against the enthalpies of formation calculated using Miedema's model and systematized in[7].

Appendix D: Elemental Features

TABLE XII. 49 considered elemental properties from the ‘Database on Properties of Chemical Elements’[61].

Name of the feature	Name of the feature
Atomic electron scattering factor at 0.5	Mendeleev H t-d start left
Atomic environment number (Villars, Daams)	Mendeleev H t-d start right
Atomic number start counting left top, left-right sequence	Mendeleev Pettifor
atomic weight	Mendeleev Pettifor regular
Charge nuclear effective (Clementi)	Mendeleev t-d start left
density	Mendeleev t-d start right
Distance core electron (Schubert)	molar heat capacity
Distance valence electron (Schubert)	moment nuclear magnetic
Electrochemical weight equivalent	nuclear charge effective
Electronegativity (Martynov&Batsanov)	first oxidation state (number)
Electronegativity absolute	periodic number counting bottom right, right-left sequence
Energy cohesive Brewer	periodic number counting left bottom, left-right sequence
energy of ionization first	periodic number counting top right, right-left sequence
enthalpy of melting	quantum number
enthalpy of vaporization	radius covalent
entropy of solid	radius metal (Waber)
group number	radius pseudo-potential (Zunger)
magnetic frequency of nuclei	spin nuclei
magnetic resonance	temperature boiling
mass attenuation coefficient for MoKα	temperature melting
Mendeleev chemists sequence	thermal neutron capture cross section
Mendeleev d-t start left	valence electron number
Mendeleev d-t start right	volume atom (Villars, Daams)
Mendeleev H d-t start left	bulk modulus
Mendeleev H d-t start right	