Predicting density functional theory total energies and enthalpies of formation of metal-nonmetal compounds by linear regression

Ann M. Deml, Ryan O'Hayre, Chris Wolverton, and Vladan Stevanović

Phys. Rev. B **93**, 085142 — Published 29 February 2016

DOI: 10.1103/PhysRevB.93.085142

Predicting DFT total energies and enthalpies of formation of metal-nonmetal compounds by linear regression

Ann M. Deml,[1,2] Ryan O'Hayre,[1] Chris Wolverton,[3] Vladan Stevanović[1,2,*]
[1]*Colorado School of Mines, Golden, Colorado, USA*
[2]*National Renewable Energy Laboratory, Golden, Colorado, USA*
[3]*Northwestern University, Evanston, Illinois, USA*

The availability of quantitatively accurate total energies (E_{tot}) of atoms, molecules, and solids, enabled by the development of density functional theory (DFT), has transformed solid state physics, quantum chemistry, and materials science by allowing direct calculations of measureable quantities such as enthalpies of formation (ΔH_f). Still, the ability to compute E_{tot} and ΔH_f values does not necessarily provide insights into the physical mechanisms behind their magnitudes or chemical trends. Here, we examine a large set of calculated E_{tot} and ΔH_f values obtained from the DFT+U-based FERE approach (*Phys. Rev. B* **85**, 115104 (2012)) to probe relationships between the $E_{tot}/\Delta H_f$ of metal-nonmetal compounds in their ground-state crystal structures and properties describing the compound compositions and their elemental constituents. From a stepwise linear regression, we develop a linear model for E_{tot}, and consequently ΔH_f, that reproduces calculated FERE values with a mean absolute error of ~80 meV/atom. The most significant contributions to the model include calculated total energies of the constituent elements in their reference phases (e.g. metallic iron or gas phase O_2), atomic ionization energies and electron affinities, Pauling electronegativity differences, and atomic electric polarizabilities. These contributions are discussed in the context of their connection to the underlying physics. We also demonstrate that our $E_{tot}/\Delta H_f$ model can be directly extended to predict the E_{tot} and ΔH_f of compounds outside the set used to develop the model.

I. INTRODUCTION

The development and implementation of density functional theory (DFT)[1,2] enabled direct and quantitatively accurate calculations of the total energies and electronic structures of many electron systems such as atoms, molecules, and solids. As a result, directly measurable quantities derived from total energies and/or electronic structures have also become directly accessible to calculations. These quantities include atomic and crystal structures, bulk moduli, enthalpies of formation, optical properties, phonon dispersions, and others.[3] Consequently, first principles, or *ab initio,* calculations are now widely used to predict and understand material properties. Furthermore, due to rapid increases in computer power, high-throughput DFT calculations have emerged as means to address the limited availability of measured physical properties across large chemical spaces of both known and hypothetical materials.[4–9] The increased availability of data, combined with modern data mining and machine learning techniques, has enabled the construction of predictive models that not only provide insights into the composition dependence of materials properties but that can also replace DFT calculations and further accelerate data generation.[10–14]

For both fundamental and practical value, the enthalpies of formation (ΔH_f) of compounds are of significant interest for first principle high-throughput calculations. ΔH_f measure the change in enthalpy upon forming a compound from its constituent elements in their standard states; therefore, it can be directly derived from the calculated total energies of a compound and its constituent elements. The fundamental value of ΔH_f lies in providing an energy scale that measures the strength of chemical bonding in a compound relative to the strength of bonding in its constituent elements. The practical value of ΔH_f is in providing enthalpies of chemical reactions that can be used in predicting: the stability of materials with respect to decomposition into competing phases,[15,13] the existence of new (previously unreported) materials,[5,16,10] material growth conditions,[5,15] Li-ion battery voltages,[17,18] etc. Furthermore, ΔH_f has recently been shown to be a key descriptor in modeling the formation energies of oxygen vacancies in metal oxides.[19,20] Nevertheless, experimental ΔH_f are available for only a fraction of known metal-nonmetal compounds, most of which are binary compounds (one metal) along with a small number of ternary (two metals) and higher order compounds.[21,22]

Toward the aim of achieving accurate and high-throughput calculations of ΔH_f, approaches have been developed over the last decade or so to correct the apparent inability of the standard approximation to DFT (more precisely DFT+U) to provide quantitatively accurate ΔH_f values.[15,23–26] As a result of these efforts, a large number of calculated ΔH_f are now available in a several open online databases.[8,7,9,27,28] Calculations of ΔH_f do, however, pose

two primary limitations: (1) computational approaches rely critically on the availability of structural information, and (2) calculated ΔH_f values themselves do not provide insights into the dependence of ΔH_f on other materials properties, e.g. composition and/or structure. These limitations continue to motivate the development of alternative methods to predict ΔH_f and expand understanding of its chemical trends.

Historically, numerous efforts have aimed to develop empirical methods for estimating ΔH_f. For example, the semi-empirical Miedema model, which was developed to predict the ΔH_f of metal alloys from intrinsic properties of their constituent elements,[29,30] has been shown to perform well for metal alloys as well as metallic hydrides.[31,32] Similarly, the CALPHAD approach has been quite successful in calculating the ΔH_f of complex metal alloy systems from thermodynamic properties of their binary constituents.[33,34] With regard to metal-nonmetal compounds, the ΔH_f of binary transition metal-nonmetal compounds have been shown to exhibit a quadratic dependence on composition ($\Delta H_f(MX_z) = az + bz^2$)[35] while the ΔH_f of metal-oxyhalides exhibit a linear correlation with the ΔH_f of their constituent oxides and halides.[36] Main group metal-nonmetal compounds have also been shown to exhibit a linear correlation between the ΔH_f of compounds with different nonmetals (e.g. MCl and MBr) when referenced to the ΔH_f of a third series of compounds (e.g. M_2O).[37]

More recently, Meredig *et al.* successfully employed machine learning techniques to predict ΔH_f across large compositional spaces with a mean absolute error (MAE) ~160 meV/atom.[13] Additionally, a simple heuristic approach that estimates the formation energy of a ternary system from the formation energies of its binary constituents has been demonstrated to achieve predictive accuracy across an extremely wide range of ternary chemistries including metal alloys and metal-nonmetal compounds.[13]

Motivated to develop further insight into the physical mechanisms that influence the magnitude and chemical trends in ΔH_f, we employ an alternative approach to predict the ΔH_f of metal-nonmetal compounds. In this paper, instead of modeling ΔH_f directly, we use stepwise linear regression to probe the relationships between the E_{tot} of metal-nonmetal compounds and the physical and chemical properties describing compound compositions and their elemental constituents. Predicted ΔH_f values are subsequently derived from the predicted E_{tot} according to

$$\Delta H_f = E_{tot} - \sum_i c_i \, \mu_i^{FERE} \quad (1)$$

where c_i and μ_i^{FERE} are the stoichiometric coefficients of the constituent elements and the Fitted Elemental-phase Reference Energies[15] (FERE) of the elements in their standard state reference phases, respectively. The FERE method significantly improves the accuracy of ΔH_f derived from DFT+U compared to experiment.[15] Directly modeling the E_{tot} of compounds instead of their ΔH_f avoids the DFT+U problem of properly comparing the energies of compounds and elements in order to obtain accurate ΔH_f. As a result, this approach also allows treatment of compounds composed of elements for which FERE values are not available.

The chosen stepwise regression approach has the advantage of allowing consideration of numerous candidate descriptors while the relative simplicity of linear functionals has the potential to facilitate model interpretation, particularly in comparison to traditional machine learning approaches such as ensemble decision trees, artificial neural networks, and Bayesian networks. Our approach parallels a recent methodology proposed for the systematic selection of physically meaningful descriptors to model material properties.[38]

From the calculated E_{tot} of metal-nonmetal compounds in their ground-state crystal structures, we develop a model of E_{tot} (and, therefore, ΔH_f) with a MAE~80 meV/atom (**Fig. 1**). The model inputs are solely composition dependent and include terms describing the compound composition and the physical and chemical properties of its elemental constituents. Within the considered approach and set of descriptors, contributions from up to 82 terms are required to accurately predict E_{tot} indicating that no simple combination of the considered descriptors describes the underlying physics influencing E_{tot} sufficiently well within a linear functional. The most significant contributions include terms describing calculated total energies of the constituent elements in their reference phases, atomic ionization energies and electron affinities of the metal and nonmetal species respectively, Pauling electronegativity differences between metal and nonmetal species, and atomic electric polarizabilities.

We also explore the applicability of our model to elements not originally included in training the model and to experimentally unreported compounds which may be less stable than those included in the training set. We find that the E_{tot}, and consequently ΔH_f, of compounds with elements such as Mo and Pb that were not included in the original training set are accurately predicted via the addition of a simple, fitted correction for each element. The E_{tot} and ΔH_f of 95 experimentally unreported metal-chalcogenides of the form A_2BX_4 (X=O, S, Se, Te)[5] are also accurately predicted with a MAE=72 meV/atom.

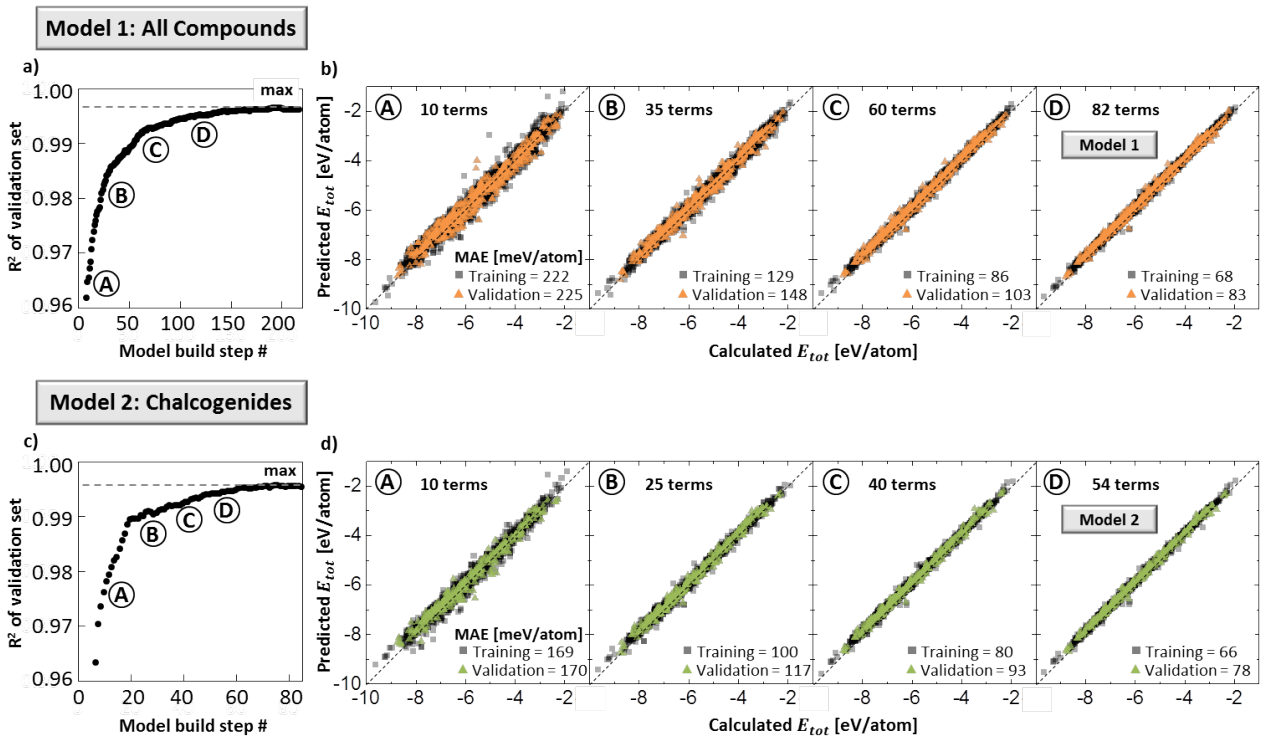

FIG. 1. Stepwise linear regression sequentially adds and removes terms from the total energy (E_{tot}) model according to their effective significance probability. The resultant increase in the R^2 of the validation set is shown for a) Model 1: All Compounds and c) Model 2: Chalcogenides. The best models (D) were selected based on diminishing increases in the R^2 of the validation set. b,d) Intermediate models A-C and final models D represent specific steps during the build and show increasing accuracy with the number of terms in the model.

II. MODELING TOTAL ENERGIES OF COMPOUNDS BY LINEAR REGRESSION

To probe the relationships between the E_{tot} of metal-nonmetal compounds and the physical and chemical properties describing the compounds (mainly composition) and their elemental constituents, we use a stepwise linear regression approach implemented in JMP.[39] We analyze the calculated E_{tot} of ~2,000 stoichiometric and fully ordered compounds in their ground-state crystal structures. This data set includes 12 % binary, 67 % ternary, 20 % quaternary, and <1 % higher order chemistries reported in the Inorganic Crystal Structure Database (ICSD)[40] with constituent elements spanning a relatively large portion of the periodic table, as shown in **Fig. 2**. The percentage of compounds containing each element are also provided in **Fig. 2**. We employ a stepwise linear regression approach to model E_{tot} (1) in order to enable consideration of numerous candidate descriptors and (2) to take advantage of the relative simplicity of linear functionals which thereby offer the potential to facilitate model interpretation compared to traditional machine learning models.

From the ICSD, we considered all stoichiometric and fully ordered ionic compounds with cations from groups 1-14 of the periodic table and anions from groups 15-17 (**Fig. 2**). In addition, we constrained our dataset to the ~4,200 compounds with a single anion specie and with integer formal charges on each of the elements. Our E_{tot} analysis considers only the lowest energy magnetic configuration and the ground state crystal structure as determined by DFT+U for each unique metal-nonmetal composition resulting in a set of 2,227 compounds. Furthermore, charge dependent candidate descriptors are not available for all elements in all charge states (e.g. spin-orbit coupling constants were not available for Mn^{5+} and Mn^{6+}), which resulted in a final data set of 2,046 compounds. It is relevant to note that in comparison to other metal-nonmetal compounds, metal oxides have been most extensively studied and represent a significant fraction of ICSD entries. Consequently, metal oxides constitute 45 % of the compounds included in our study. All nonmagnetic and magnetic compounds containing N, O, P, or S anions were included in our dataset. However, due to limited computational resources, only compounds containing main group metals and/or the transition metals shown in blue in Fig. 1 were included for compositions with F, Cl, As, Se, Sb, Te, or Bi anions. The omitted transition metal compounds constitute a small fraction of the total number of compounds in our dataset.

Spin polarized DFT+U[41] calculations of E_{tot} were performed following the procedure used to develop the FERE values as described in Ref. [15]. A plane wave basis set, the PBE exchange-correlation functional,[42] and the

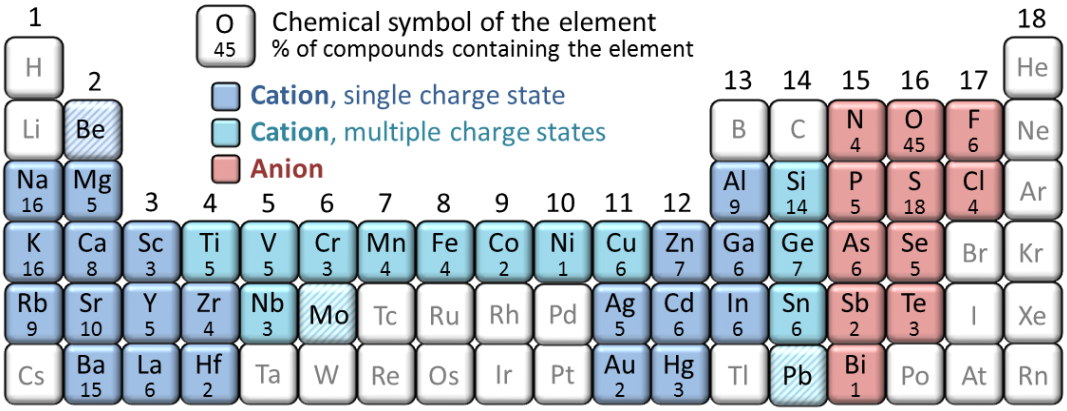

FIG. 2. We analyze the total energies of ~2,000 metal-nonmetal compounds which include binary, ternary, and higher order chemistries with constituent elements spanning the colored portion of the periodic table. Constituent elements include metals, or cations, with a single charge state (blue) and with multiple charge states (aqua) and nonmetals, or anions (red). The percent of compounds containing each element is provided below the chemical symbol. Be, Mo, and Pb (hashed) were used to demonstrate direct extension of the E_{tot} and ΔH_f models to elements outside the set used to develop the models.

projector augmented wave (PAW) method[43] were used as implemented in the VASP computer code.[44] The E_{tot} of compounds and elements in VASP are referenced to the sum of the total energies of the isolated constituent "pseudo" atoms in the reference configurations used to generate their pseudopotentials (i.e. without spin polarization or nonzero U values). We use the same E_{tot} definition throughout this work.

A constant U=3 eV value was used for all transition metals except Cu and Ag for which we use U=5 eV in accordance with the parameters used to develop the FERE method.[15] For all nontransition metals, the Hubbard U parameter was set to zero. A Monkhorst-Pack k-point sampling[45] was applied such that benchmarked total energies converged within 3 meV/atom with respect to the number of k points. A plane wave energy cutoff was set to 340 eV corresponding to a value ~20% greater than the highest cutoff energy suggested by the pseudopotentials specified in Ref. [15] (282 eV for the soft oxygen pseudopotential). Full volume, cell shape, and atomic position relaxations were performed starting from structures reported in the ICSD.[40] Spin degrees of freedom were treated explicitly, and a limited search for the lowest energy magnetic configuration was performed as described in Ref. [15]. The calculated results including structures, magnetic configurations, E_{tot}, and ΔH_f are available online at materials.nrel.gov.

As candidate descriptors, we consider properties describing the compound composition and the physical and chemical properties of its elemental constituents as an approximate chemical description of the compound (**Table 1**). For those elemental properties marked with an asterisk (*), candidate descriptors include the maximum, minimum, range, standard deviation, and stoichiometric weighted mean of the elemental property resulting in a total of 124 main descriptors. The square root and inverse of each term (an additional 248 descriptors) as well as the products of the primary and stoichiometric weighted mean values (4,692 descriptors) are also included for a total of ~5,000 candidate descriptors. It is pertinent to note that, similar to previous work by Meredig *et al*.,[13] these model inputs are solely composition dependent (i.e. structure independent) which enables the prediction of E_{tot}, ΔH_f, and related properties such as thermodynamic stability, without a known ground state structure or costly structure search. In addition, the considered candidate descriptors are versatile and applicable to compounds with any number of elemental constituents thereby enabling prediction of E_{tot} for binaries, ternaries, quaternaries, and even higher order chemistries.

Stepwise linear regression sequentially adds and removes terms (candidate descriptors) from a linear model according to their effective significance probabilities. To verify the true predictive power of our models, we randomly divide each dataset into three groups. We use 70 % of the dataset to train the model and 15 % for preliminary assessment (validation) of the model's predictive ability and selection of the best model. We completely withhold the remaining 15 % for an independent assessment (testing) of the model's predictive ability. We also withhold all compounds containing Mo or Pb for later testing of the model's ability to predict the E_{tot} and ΔH_f of compounds containing elements outside the set used to develop the model. Ten-fold cross validation was also used within the training set. The iterative addition and removal of terms was restricted according to effective significance probabilities where the smaller the *p*-value, the larger the significance of the term. Only terms with a p-value (significance level) ≤ 0.25, those with at least moderate significance, were allowed to enter the model. Likewise, only terms with a p-value ≥ 0.25, those with a loss of significance due to the addition of other terms, were allowed to be removed.

TABLE 1. Properties describing the compound composition and the physical and chemical properties of its elemental constituents were considered as candidate descriptors for modeling the total energies of the compounds. Candidate descriptors for properties marked with an asterisk (*) included the maximum, minimum, range, standard deviation, and stoichiometric weighted mean of the given elemental property. The square root and inverse of each term were also included along with the products of the primary (those without an asterisk) and stoichiometric weighted mean values. The reference phases of the elements are those corresponding to standard conditions (T=298 K and p=1 atm).

	Candidate descriptors	Description
	atoms/formula unit	Number of atoms per stoichiometric formula unit
	fraction of transition metals	Fraction of transition metal elements with multiple charges states as shown in Fig. 2
	atomic number*	Atomic numbers of the constituent elements
	atomic mass*	Atomic masses of the constituent elements
	row number*	Periods, or row numbers in the periodic table, of the constituent elements
valence electrons	column number*	Groups, or column numbers in the periodic table, of the constituent elements
valence electrons	number of s, p, or d valence electrons	Average number of s, p, or d valence electrons for the neutral atomic species
valence electrons	fraction of s, p, or d valence electrons	Fraction of s, p, or d valence electrons for the neutral atomic species
	atomic radius*,[46]	Measured covalent radii of the constituent elements in their reference phases
	molar volume*,[47]	Molar volumes of the constituent elements in their reference phases
phase change & heat capacity	latent heat of fusion*,[47]	Latent heats of fusion (heating from a solid to a liquid) of the constituent elements in their reference phases
phase change & heat capacity	melting point*,[47]	Melting point temperatures of the constituent elements in their reference phases
phase change & heat capacity	boiling point*,[47]	Boiling point temperatures of the constituent elements in their reference phases
phase change & heat capacity	heat capacity*,[47]	Molar heat capacities of the constituent elements in their reference phases
ionization energies & electron affinity	1st ionization energy*,[47]	1st atomic ionization energies of the constituent elements
ionization energies & electron affinity	cumulative ionization energy*,[47]	Cumulative sums of the 1st, 2nd,… atomic ionization energies of the constituent metals summed to the cation formal charge state
ionization energies & electron affinity	electron affinity,[47]	1st atomic electron affinity of the nonmetal multiplied by the anion formal charge state
electro-negativities	Pauling electronegativity*,[47]	Pauling electronegativities of the constituent elements
electro-negativities	Pauling electronegativity difference*,[47]	Pauling electronegativity differences between first nearest neighbors assuming all metals have equal coordination with the nonmetals
	formal charge*	Integer formal charges of the constituent cations and anions
magnetic properties	crystal field splitting*,[48]	Empirical constants for estimating the charge dependent cation contributions to the crystal field splitting energy for octahedral complexes of the constituent transition metals
magnetic properties	magnetic moment*,[48]	Charge dependent magnetic moments (spin only) of the constituent transition metal cations in high spin, octahedral complexes
magnetic properties	spin-orbit coupling*,[48]	Charge dependent spin-orbit coupling constants for single d-electrons in the constituent transition metal cations
magnetic properties	max total electron spin*	Charge dependent maximum number of unpaired d-electrons for the constituent transition metal cations
	electric polarizability*,[47]	Atomic electric dipole polarizabilities of the constituent elements
elemental-phase reference energies	GGA+U elemental energy*,[15]	GGA+U total energies per atom of the constituent elements in their reference phases
elemental-phase reference energies	Fitted elemental-phase reference energy*,[15]	Fitted total energies per atom of the constituent elements in their reference phases from the FERE approach
elemental-phase reference energies	FERE correction energy*,[15]	Differences between the GGA+U and FERE total energies per atom of the constituent elements in their reference phases

The intermediate models A-C, shown in **Fig. 1**, correspond to individual "steps" during the model build and show that the model accuracy initially increases with the number of terms in the model. From these models, it is evident that within the chosen approach and set of descriptors a relatively large number of terms (30-60) are required to attain moderate and chemically useful accuracies (MAE<100 meV/atom). The final model D, shown in **Fig. 1** and labeled Model 1, has a MAE~80 meV/atom for the test set and was selected using the maximum R^2 of the validation set. The linear function for the final model is provided in the Supplemental Material and online at https://github.com/ademl/predict_Etot_dHf. Terminating the model build at the maximum R^2 from the validation set inhibits overfitting which occurs when the model describes random error in the training set instead of the underlying relationships. Additionally, since we aim to produce simple models with a minimal number of terms, the final model was selected prior to the maximum R^2 of the validation set when the addition of 10 terms gave <0.1% increase in the R^2 from the validation set, as shown by the point D in **Fig. 1a,c**.

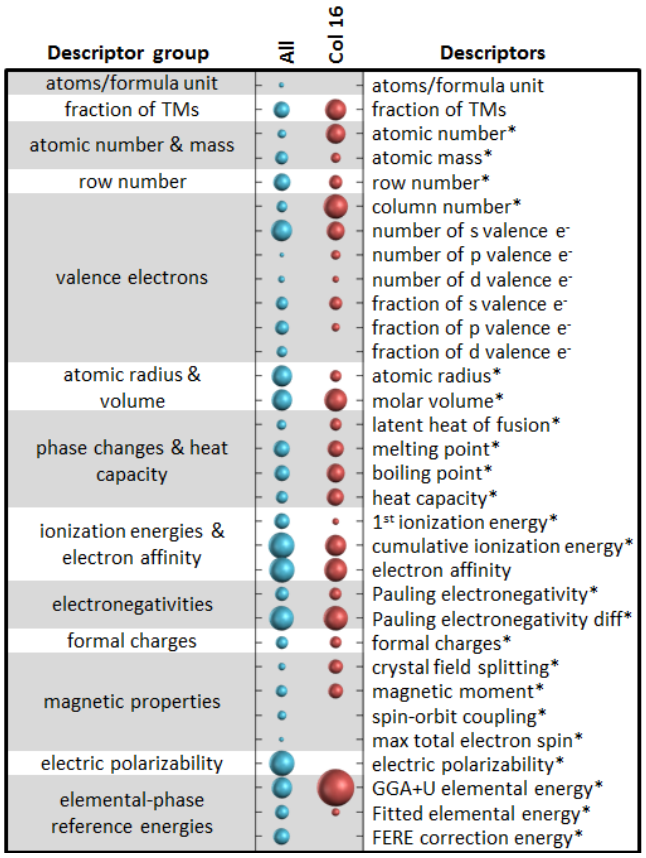

FIG. 3. The relative contributions of candidate descriptors to the final Model 1 for all metal-nonmetal compounds (blue) and Model 2 for chalcogenides, compounds with anions from column 16, (red) provide insight into the properties that most directly relate to the E_{tot} of metal-nonmetal compounds. The areas of the bubbles represent the relative significance of each set of descriptors with the most significant cumulative contributions coming from descriptors of the atomic ionization energies and electron affinities of the metal and nonmetal species respectively, the Pauling electronegativity differences, the atomic electric polarizabilities, and the total energies of the constituent elements in their reference phases. Descriptors for properties marked with an asterisk (*) included the maximum, minimum, range, standard deviation, and stoichiometric weighted mean of the given elemental property.

The full dataset used for training, validation, and testing of Model 1 includes extensive compositional diversity. Alternatively, the dataset can be strategically partitioned into subsets of compounds with greater chemical similarity. For example, partitioning compounds by the anion column number results in Model 2 (**Fig. 1**), which corresponds to the subset of 1,696 chalcogenides, compounds with anions from column 16. Compared to Model 1, Model 2 includes 28 (34%) fewer terms and a similar MAE~80 meV/atom for the test set. Again, the linear function for the final model is provided in the Supplemental Material. This partition, therefore, separates the dataset such that the provided descriptors more accurately describe variations in E_{tot}. The large number of metal oxides in the ICSD yields a sufficiently large dataset to apply this analysis to chalcogenides. On the other hand, pnictides and halides, compounds with anions from columns 15 and 17 respectively, are significantly less prevalent, and E_{tot} models for these subsets exhibit fewer terms and reduced accuracy compared to chalcogenides due to overfitting constraints. Nevertheless, with sufficiently large datasets, the same approach could also be applied to pnictides and halides.

III. ANALYSIS OF PROPERTIES GOVERNING CHEMICAL TRENDS

To gain insight into the underlying physics that influences the E_{tot} values and chemical trends of compounds, we examine the cumulative contributions of candidate descriptors to the final Models 1 and 2 (**Fig. 3**) using the absolute value of the descriptor t-statistic, a ratio of the parameter estimate to its standard error. The most significant contributions to the models include GGA+U total energies of the constituent elements in their reference phases. Other significant contributions include atomic ionization energies and electron affinities of the metal and nonmetal species respectively, Pauling electronegativity differences, and atomic electric polarizabilities. The latter set of descriptors reflects properties that somewhat directly, and also intuitively, relate to chemical bonding and the E_{tot} of metal-nonmetal compounds. For example, atomic ionization energies and electron affinities for the metal and nonmetal species, respectively, reflect the required energies to form (partially) ionic species in a metal-nonmetal compound. Similarly, electronegativity differences between the metal and nonmetal species relate to degree of charge transfer and covalency of bonding in the compound while atomic electric polarizabilities reflect the ease of electron density distortions due to the presence of neighboring ions.

We find that, of the considered descriptors, the single term with the strongest correlation to E_{tot} is the stoichiometric weighted mean of the GGA+U total energies of the constituent elements in their reference phases (μ^{GGA+U}), as shown in **Fig. 4**. As with compounds, the total energies of the elements are referenced to the sum of the total energies of the isolated constituent pseudo-atoms in vacuum. Similar trends are observed between the GGA+U cohesive energies of the compounds and the cohesive energies of their constituent elements (**Fig. S1**). The correlation between E_{tot} and μ^{GGA+U} arises mainly due to the large energy difference between isolated pseudo-atoms and bonded atoms in molecules and solids. In other words, the energies of solids and molecules are relatively similar to one another when compared to the very high energies of

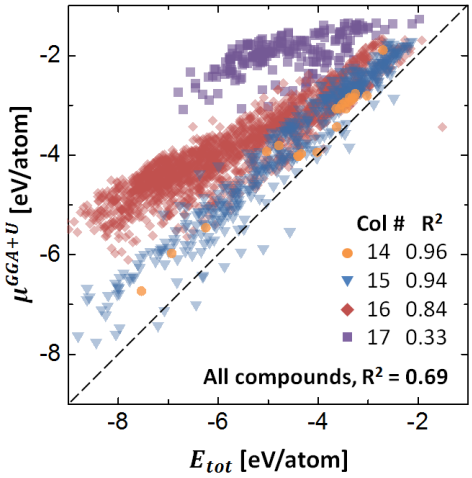

FIG. 4. Of the considered descriptors, the stoichiometric weighted mean GGA+U total energies of the constituent elements in their reference phases (μ^{GGA+U}) exhibit the strongest correlation with the total energies of the compounds (E_{tot}). Correlations are strongest among compounds with anions from the same column of the periodic table.

isolated atoms. As a result, μ^{GGA+U} values contribute most significantly to the E_{tot} models.

It is also apparent from **Fig. 4** that the correlation between calculated E_{tot} and μ^{GGA+U} is strongest within groups of compounds with anions from the same column of the periodic table. This trend results from trends in the energies of the reference phase gas molecules. The extremely low energy of pnictogens (e.g. $\mu_{N_2}^{GGA+U}$=-8.3 eV/atom) results in more negative μ^{GGA+U} values for pnictides compared to chalcogenides ($\mu_{O_2}^{GGA+U}$=-5.0 eV/atom) and halides ($\mu_{F_2}^{GGA+U}$=-1.9 eV/atom). As a result, compounds with anions further to the right of the periodic table exhibit greater stabilization relative to their constituent elements. This correlation among compounds with anions from the same column is consistent with our Model 2 results for partitioning the compounds which uses fewer terms to obtain a similar accuracy for chalcogenides compared to the complete set of all anions. Lastly, **Fig.** 4 shows that nearly all metal-nonmetal compounds exhibit E_{tot} values that are more negative than their composition averaged μ^{GGA+U} values, indicating that the compounds are lower energy and, therefore, more stable than their elemental constituents. This calculated stability is consistent with the existence of these entries in the ICSD, which primarily contains experimental data.

It is relevant to note that building a model of E_{tot} with fewer terms (and consequently, reduced accuracy) is best achieved by a sequential build rather than by a secondary down-selection of the most significant terms from a model with more terms. For example, a model including the 20 most significant terms from Model 1 exhibits a significantly larger MAE=219 meV/atom for the test set (when the linear coefficients are re-optimized) than a model that was originally terminated with 20 terms resulting in a MAE=152 meV/atom for the test set. This effect arises from the fact that stepwise addition and removal of terms iteratively optimizes contributions and interdependencies from all provided candidate descriptors. In contrast, a secondary down-selection of terms is restricted to optimizing contributions from only those descriptors that were selected in the original model build.

Ultimately, a relatively large number of terms (30-60) is required to attain chemically useful accuracies (MAE<100 meV/atom) within this linear regression approach and set of descriptors. Additionally, numerous descriptors are interdependent (not necessarily linearly) and many properties are reflected in multiple terms. For example, the mean and minimum GGA+U elemental energies as well as several cross terms containing the mean GGA+U elemental energies are all included in Model 1. This reality inhibits the ability to further elucidate insights from the specific terms and functional forms of the models and also indicates that any truly causal (physically meaningful) connection[38] between the descriptors and $E_{tot}/\Delta H_f$ is complex.

The relatively large number and complexity of terms in our E_{tot} models indicate that no simple combination of the considered descriptors describes the underlying physics influencing E_{tot} sufficiently well within a linear functional. Interestingly, however, our errors in the predicted E_{tot} of binary compounds are larger than those of ternary and higher order chemistries (MAE=110 meV/atom for binary compounds and 80 meV/atom for higher order chemistries from Model 1 training, validation, and test data) indicating that this linear combination of descriptors is most effective at describing the E_{tot}, and consequently ΔH_f, of complex chemistries. Therefore, our model appears to be particularly well suited for predicting the E_{tot} and ΔH_f of unreported compounds which most frequently exhibit ternary and higher order chemistries.

IV. PREDICTING COMPOUND ENTHALPIES OF FORMATION

The final models presented in **Fig. 1** were developed by fitting the E_{tot} of compounds. As discussed, the ΔH_f of the compounds can subsequently be derived from Eq. 1 and the FERE μ_i values. For comparison, the predicted and calculated ΔH_f values from the final Models 1 and 2 are shown in **Fig. 5**. Model 1 includes 82 terms with MAE=68, 83, and 77 meV/atom for the training, validation, and test sets, respectively. Model 2 includes 54 terms with MAE=66, 78, and 79 meV/atom for the training, validation, and test sets, respectively. Because the test sets are fully isolated from the model build, they provide the truest evaluation of the model performance.

The achieved accuracy of MAE~80 meV/atom in predicting E_{tot} and ΔH_f relative to calculated E_{tot} and ΔH_f, respectively, provides reasonable accuracy and provides a route for predicting E_{tot} and ΔH_f values from inputs that are solely composition dependent. Therefore, we further explore the applicability of our $E_{tot}/\Delta H_f$ model to elements not included in training the model and to experimentally unreported compounds, which were also not including in developing the model. We consider two datasets: (1) compounds from the ICSD that contain Mo or Pb, elements that were not included in the original training set, and (2) experimentally unreported metal-chalcogenides of the form A_2BX_4 (X=O, S, Se, Te) predicted stable in Ref. 5.

First, we examine the accuracy of Model 1 in predicting the calculated ΔH_f of 181 compounds from the ICSD that contain Mo or Pb, elements intentionally not included in training the model. As shown in **Fig. 6a**, the predicted ΔH_f of Mo compounds exhibit a systematic error (vertical offset) compared to calculated values. Adding a constant to Eq. 1 (-300 meV/atom, applied only to Mo compounds) provides a simple correction to the model resulting in a MAE=93 meV/atom. The predicted ΔH_f of Pb compounds, on the other hand, do not exhibit a systematic error (**Fig. 6b**, MAE=154 meV/atom). Pb compounds, therefore, do not require a correction to the $E_{tot}/\Delta H_f$ model.

We also consider 95 experimentally unreported A_2BX_4 compounds from Ref [5] that are predicted to be stable but are likely less stable than most compounds included in the training set. This set includes 11 compounds containing Be, which was not included in the original training set. As with Mo and shown in **Fig. 6c**, adding a constant to Eq. 1 (450 meV/atom, applied only to Be compounds) provides a simple correction to Model 1. Therefore, although the model was trained on reported compounds, the ΔH_f of experimentally unreported A_2BX_4 compounds are accurately predicted and, with the correction for Be compounds, exhibit a MAE=72 meV/atom.

These results for Mo, Pb, and Be compounds indicate that our $E_{tot}/\Delta H_f$ model can be extended to elements not included in training the model simply by fitting a constant for any new element. In other words, it is not necessary to retrain the model. In addition, the fitting dataset can be relatively small and, therefore, requires only a small number of additional E_{tot} calculations. The observed

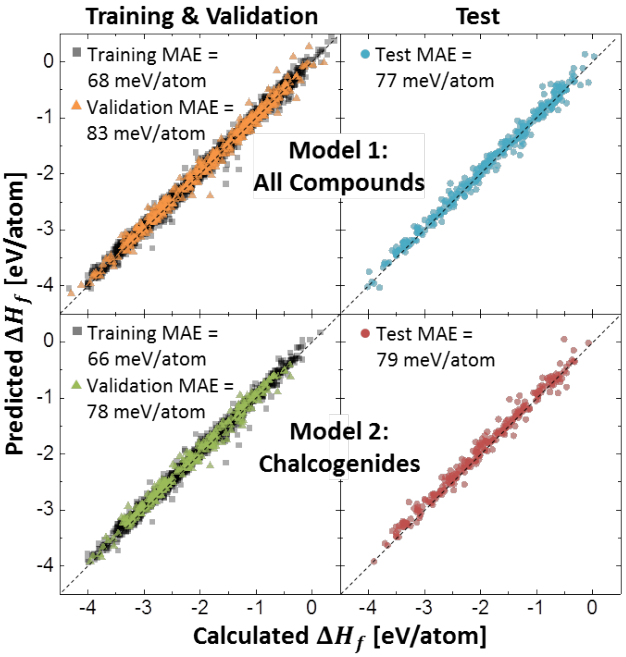

FIG. 5. Predicted enthalpies of formation (ΔH_f) from Model 1 and 2 are in good agreement with calculated ΔH_f for the full set of metal-nonmetal compounds and for chalcogenides, respectively. The training and validation data sets were used to build and select the best models; the test data sets were fully isolated from the model build.

systematic errors in predicting ΔH_f for certain elements arise when the element exhibits descriptor values outside of or near the boundaries of the previous range of values. For example, the atomic number, atomic mass, and heat capacity of Be are smaller than those of other elements that were included in training the model. In combination, these properties result in a systematic error in the predicted E_{tot} and ΔH_f of Be compounds since the model was not trained for this range of values. Similarly, the melting point of Mo is higher than that of most other elements included in training the model resulting in a systematic error in the predicted E_{tot} and ΔH_f of Mo compounds. The properties of Pb, on the other hand, fall within the same ranges of elements included in training the model and no systematic error is observed. Overall, our $E_{tot}/\Delta H_f$ model can be applied to accurately predict the ΔH_f of both elements not originally included in training the model and of experimentally unreported compounds.

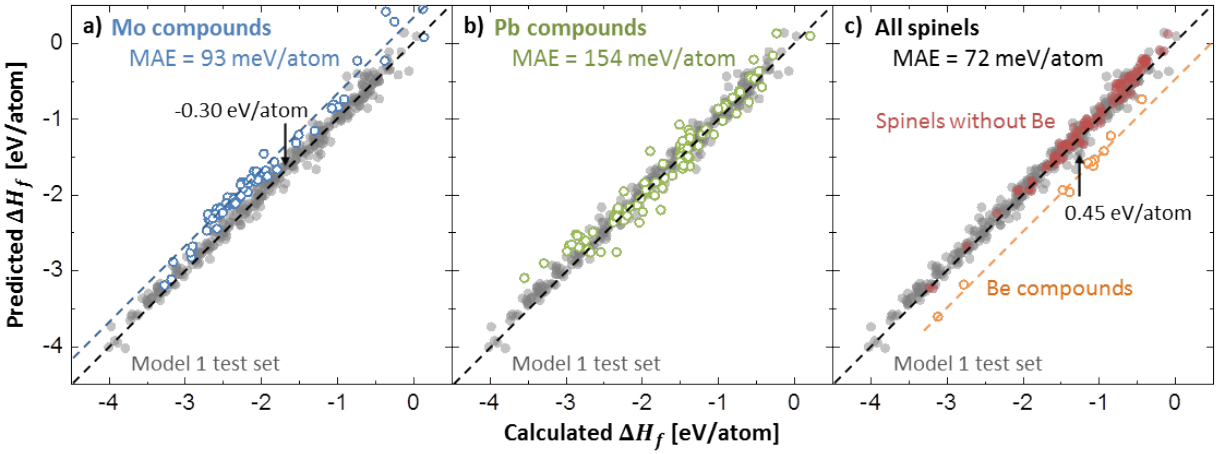

FIG. 6. The enthalpies of formation (ΔH_f) of compounds containing the elements a) Mo, b) Pb, and c) Be, which were not included in training the model, are accurately predicted by adding element-specific constants to the $E_{tot}/\Delta H_f$ model. c) In addition, although the model was trained solely on reported compounds, the ΔH_f of experimentally unreported spinel metal-chalcogenide compounds are accurately predicted. Gray symbols correspond to the Model 1 test set which contains only compounds from the ICSD and only elements that were included in training the model. Open symbols correspond to compounds with elements not included in training the model. Closed red symbols in c) correspond to new spinel compounds composed only of elements that were included in training the model.

V. CONCLUSIONS

Motivated by our aim to provide insights into the underlying physics that dictates the chemical trends in E_{tot}, and therefore ΔH_f, we have developed a linear model for the E_{tot} of metal-nonmetal compounds. Our model reproduces E_{tot} and FERE ΔH_f values with a mean absolute error ~80 meV/atom and with inputs that are solely composition dependent. The most significant contributions to the model include calculated total energies of the constituent elements in their reference phases, atomic ionization energies and electron affinities, Pauling electronegativity differences, and atomic electric polarizabilities. These descriptors reflect properties that most directly relate to chemical bonding and the E_{tot} of metal-nonmetal compounds. Our model can also be applied to accurately predict the ΔH_f of both elements not originally included in training the model and of experimentally unreported compounds. Partitioning of compounds by one metric of chemical similarity, the anion column number in the periodic table, reduces the number of required terms while retaining similar accuracy. Nevertheless, within the chosen approach, a relatively large number of terms are required to attain chemically useful accuracies. Therefore, our findings demonstrate significant motive to identify more suitable descriptors of E_{tot} and to investigate other approaches for developing simple, linear or nonlinear models of E_{tot}.

ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation (NSF) under grants DMR-1309980 and DMR-1309957. The research was performed using computational resources sponsored by the Department of Energy's Office of Energy Efficiency and Renewable Energy and located at the National Renewable Energy Laboratory. The authors thank Logan Ward and Gary Haith for valuable discussions and suggestions.

[1] W. Kohn and L.J. Sham, Phys. Rev. **140**, 1133 (1965).
[2] P. Hohenberg and W. Kohn, Phys. Rev. **136**, 864 (1964).
[3] R.M. Martin, *Electronic Structure: Basic Theory and Practical Methods*, 1st ed. (Cambridge University Press, 2008).
[4] G.H. Jóhannesson, T. Bligaard, a V Ruban, H.L. Skriver, K.W. Jacobsen, and J.K. Nørskov, Phys. Rev. Lett. **88**, 255506 (2002).
[5] X. Zhang, V. Stevanović, M. d'Avezac, S. Lany, and A. Zunger, Phys. Rev. B **86**, 014109 (2012).
[6] S. Curtarolo, G.L.W. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, Nat. Mater. **12**, 191 (2013).
[7] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K.A. Persson, APL Mater. **1**, 011002 (2013).
[8] J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, JOM **65**, 1501 (2013).
[9] S. Kirklin, J.E. Saal, B. Meredig, A. Thompson, J.W. Doak, K. Muratahan, S. Ruehl, and C. Wolverton, Npj Comput. Mater. **1**, (2015).

[10] G. Hautier, C.C. Fischer, A. Jain, T. Mueller, and G. Ceder, Chem. Mater. **22**, 3762 (2010).
[11] G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder, Inorg. Chem. **50**, 656 (2011).
[12] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, Sci. Rep. **3**, 1 (2013).
[13] B. Meredig, A. Agrawal, S. Kirklin, J.E. Saal, J.W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, Phys. Rev. B **89**, 094104 (2014).
[14] J. Yan, P. Gorai, B. Ortiz, S. Miller, S.A. Barnett, T. Mason, V. Stevanović, and E.S. Toberer, Energy Environ. Sci. **8**, 983 (2015).
[15] V. Stevanović, S. Lany, X. Zhang, and A. Zunger, Phys. Rev. B **85**, 115104 (2012).
[16] R. Gautier, X. Zhang, L. Hu, L. Yu, Y. Lin, T.O.L. Sunde, D. Chon, K.R. Poeppelmeier, and A. Zunger, Nat. Chem. **7**, 308 (2015).
[17] V. Chevrier, S. Ong, R. Armiento, M. Chan, and G. Ceder, Phys. Rev. B **82**, 1 (2010).
[18] J. Bhattacharya and C. Wolverton, J. Electrochem. Soc. **161**, A1440 (2014).
[19] A.M. Deml, A.M. Holder, R.P. O'Hayre, C.B. Musgrave, and V. Stevanovic, J. Phys. Chem. Lett. **18**, 1948 (2015).
[20] A.M. Deml, V. Stevanović, C.L. Muhich, C.B. Musgrave, and R. O'Hayre, Energy Environ. Sci. **7**, 1996 (2014).
[21] O. Kubaschewski, C.B. Alcock, and P.J. Spencer, *Materials Thermochemistry*, 6th ed. (Pergamon Press, New York, 1993).
[22] D.D. Wagman, W.H. Evans, V.B. Parker, R.H. Schumm, I. Halow, S.M. Bailey, K.L. Churney, and R.L. Nuttall, J. Phys. Chem. Ref. Data **11**, Supplement No. 2 (1982).
[23] L. Wang, T. Maxisch, and G. Ceder, Phys. Rev. B **73**, 195107 (2006).
[24] A. Jain, G. Hautier, S. Ong, C. Moore, C. Fischer, K. Persson, and G. Ceder, Phys. Rev. B **84**, 045115 (2011).
[25] S. Grindy, B. Meredig, S. Kirklin, J.E. Saal, and C. Wolverton, Phys. Rev. B **87**, 075150 (2013).
[26] M. Pandey and K.W. Jacobsen, Phys. Rev. B **91**, 235201 (2015).
[27] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R.H. Taylor, L.J. Nelson, G.L.W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy, Comput. Mater. Sci. **58**, 227 (2012).
[28] NREL High Performance Computing Center Materials Database, http://materials.nrel.gov/ (2015).
[29] A. Miedema, P.F. de Chatel, and F.R. de Boer, Physica **100**, 1 (1980).
[30] A.P. Gonçalves and M. Almeida, Phys. B **228**, 289 (1996).
[31] P.C.P. Bouten and A. Miedema, J. Less Common Met. **71**, 147 (1980).
[32] K.H.J. Buschow, P.C.P. Bouten, and A.R. Miedema, Reports Prog. Phys. **45**, 937 (1982).
[33] N. Saunders and A.P. Miodownik, *Calphad (A Comprehensive Guide)* (Pergamon Press, Oxford, 1998).
[34] H.. Lukas, S.. Fries, and B. Sundman, *Computational Thermodynamics – The Calphad Method* (Cambridge University Press, 2007).
[35] M.W.M. Hisham and S.W. Benson, J. Phys. Chem. **89**, 3417 (1985).
[36] M.W.M. Hisham, S.W. Benson, and O. Compounds, J. Phys. Chem. **90**, 885 (1986).
[37] M.W.M. Hisham and S.W. Benson, J. Phys. Chem. **92**, 6107 (1988).
[38] L.M. Ghiringhelli, J. Vybiral, S. V Levchenko, C. Draxl, and M. Scheffler, Phys. Rev. Lett. **114**, 105503 (2015).
[39] JMP® Pro, (2014).
[40] A. Belsky, M. Hellenbrandt, V.L. Karen, and P. Luksch, Acta Crystallogr. Sect. B Struct. Sci. **B58**, 364 (2002).
[41] S.L. Dudarev, G.A. Botton, S.Y. Savrasov, C.J. Humphreys, and A.P. Sutton, Phys. Rev. B **57**, 1505 (1998).
[42] J.P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[43] P. Blöchl, Phys. Rev. B **50**, 17953 (1994).
[44] G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 (1996).
[45] H. Monkhorst and J. Pack, Phys. Rev. B **13**, 5188 (1976).
[46] J.C. Slater, J. Chem. Phys. **41**, 3199 (1964).
[47] W.M. Haynes, editor , *CRC Handbook of Chemistry and Physics*, 95th ed. (CRC Press, 2014).
[48] W.W. Porterfield, *Inorganic Chemistry*, 2nd ed. (Academic Press, 1993).