A Formation Energy Predictor for Crystalline Materials Using Ensemble Data Mining

Ankit Agrawal[1], Bryce Meredig[2*], Chris Wolverton[2], Alok Choudhary[1]
[1]Department of Electrical Engineering and Computer Science, Northwestern University
[2]Department of Materials Science and Engineering, Northwestern University
* Current affiliation: Citrine Informatics
Corresponding author email: ankitag@eecs.northwestern.edu

Abstract—**Formation energy is one of the most important properties of a compound that is directly related to its stability. More negative the formation energy, the more stable the compound is likely to be. Here we describe the development and deployment of predictive models for formation energy, given the chemical composition of the material. The data-driven models described here are built using nearly 100,000 Density Functional Theory (DFT) calculations, which is a quantum mechanical simulation technique based on the electron density within the crystal structure of the material. These models are deployed in an online webtool that takes a list of material compositions as input, generates over hundred composition-based attributes for each material and feeds them into the predictive models to obtain the predictions of formation energy. The online formation energy predictor is available at http://info.eecs.northwestern.edu/FEpredictor**

Keywords-**Materials informatics, supervised learning, ensemble learning, density functional theory, formation energy**

I. Motivation

The field of materials science and engineering involves conducting experiments and simulations to understand the *science* of materials in order to discover and *engineer* new materials with superior properties. Like most scientific domains, materials science too has strong experimental, theoretical, and computational (simulation) branches of study. Over the last few years, the data being generated by such experiments and simulations is exploding, making it amenable to knowledge extraction via data-driven techniques. The realization of the fourth paradigm of science [1] (data-driven science, unifying the first three paradigms of experiment, theory, and simulation) in materials science has led to the emergence of the new field called materials informatics [2], [3], [4], and a surge in research efforts for data-driven materials property prediction and optimization [5], [6], [7], [8], [9], [10].

In June 2011, the US government launched the Materials Genome Initiative (MGI) [11] to realize the vision of development of advanced materials necessary for economic security and human well-being. In particular, the Materials Genome Initiative "will enable discovery, development, manufacturing, and deployment of advanced materials at least twice as fast as possible today, at a fraction of the cost". Currently, the time lag between the discovery of advanced materials and their deployment stands at more than 20 years, which this initiative aims to reduce to half. The Materials Genome Initiative Strategic Plan released in 2014 [12] also identifies data analytics as one of the key objectives as part of integrating experiments, computation, and theory, in order to realize the vision of MGI.

It is in the spirit and pursuit of the vision and approach of MGI that we discuss and present in this demonstration paper, an online data informatics tool to predict formation energy of a material, which is a crucial material property directly related to its stability. Some of these predictive models were recently used to scan (almost) the entire ternary composition space, and resulted in a first-of-its-kind computational discovery of about 4,500 new stable compounds [13].

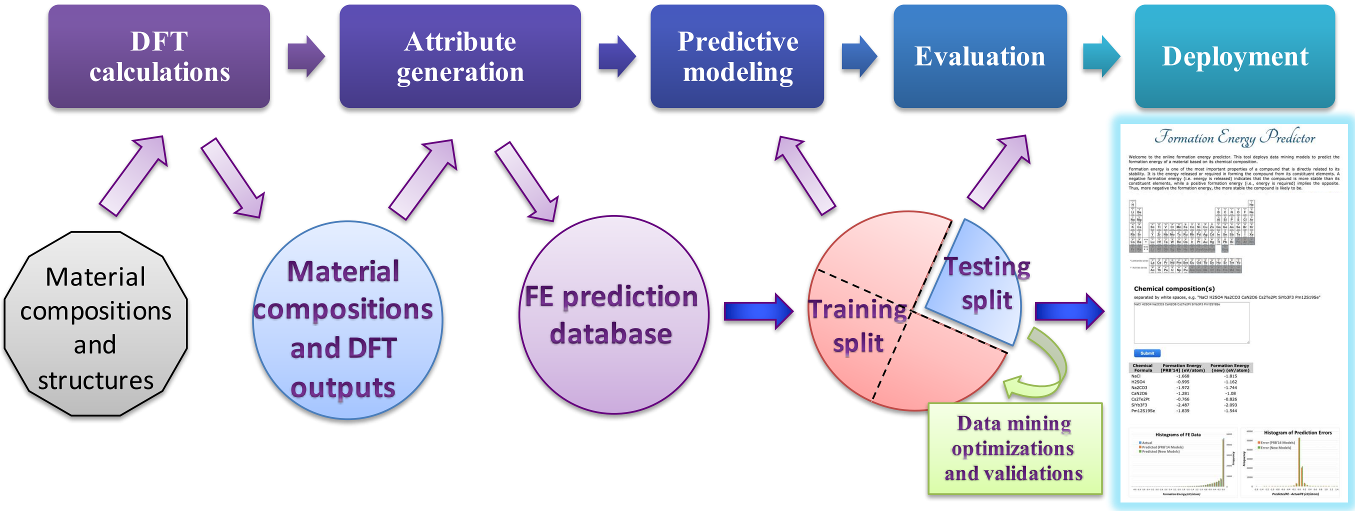
II. Materials science background

Density functional theory (DFT) is a quantum mechanical simulation technique based on the electron density within the crystal structure of the material, and is one of the most commonly used computational tools for studying the electronic scale properties of a material (a many-body system of interacting electrons). DFT calculations are extremely time consuming, and they also require the atomistic structure of the material as an input, which in turn is also very costly computationally. Depending on the size and complexity of the material being studied, a single DFT calculation can take from hours to days to months on modern computing systems.

Formation energy is one of the most important properties of a compound that is directly related to its stability. It is the energy released or required in forming the compound from its constituent elements. A negative formation energy (i.e. energy is released) indicates that the compound is more stable than its constituent elements, while a positive formation energy (i.e., energy is required) implies the opposite. The units of formation energy is electron-volts per atom (eV/atom).

III. Design

The overall data-driven process is depicted as a block diagram in Figure 1. DFT calculations were performed for nearly 100,000 materials with known composition and structure to obtain the values of formation energy. Over hundred composition-based attributes are derived and mapped to the DFT calculated formation energy to construct the formation energy prediction database. Supervised learning techniques are


DFT calculations
Attribute generation
Predictive modeling
Evaluation
Deployment
Material compositions and structures
Material compositions and DFT outputs
FE prediction database
Training split
Testing split
Data mining optimizations and validations
Formation Energy Predictor

Welcome to the online formation energy predictor. This tool deploys data mining models to predict the formation energy of a material based on its chemical composition.

Formation energy is one of the most important properties of a compound that is directly related to its stability. It is the energy released or required in forming the compound from its constituent elements. A negative formation energy (i.e. energy is released) indicates that the compound is more stable than its constituent elements, while a positive formation energy (i.e., energy is required) implies the opposite. Thus, more negative the formation energy, the more stable the compound is likely to be.

Chemical composition(s)

separated by white spaces, e.g. "NaCl H2SO4 Na2CO3 CaN2O6 Cs2Te2Pt SiYb3F3 Pm12S19Se"

NaCl H2SO4 Na2CO3 CaN2O6 Cs2Te2Pt SiYb3F3 Pm12S19Se

Submit

Chemical Formula	Formation Energy [PRB'14] (eV/atom)	Formation Energy (new) (eV/atom)
NaCl	-1.668	-1.815
H2SO4	-0.995	-1.162
Na2CO3	-1.972	-1.744
CaN2O6	-1.281	-1.08
Cs2Te2Pt	-0.766	-0.826
SiYb3F3	-2.487	-2.093
Pm12S19Se	-1.839	-1.544

Histograms of FE Data
Actual
Predicted (PRB'14 Models)
Predicted (New Models)
Frequency
Formation Energy (eV/atom)
Histogram of Prediction Errors
Error (PRB'14 Models)
Error (New Models)
Frequency
PredictedFE - ActualFE (eV/atom)

Fig. 1. The design of the data mining workflow used in this work.

then used to learn predictive models for formation energy, given only the composition of the material, without any explicit structure information. The models are evaluated using standard validation techniques, and the most accurate models are deployed in an online user-friendly web-tool that can predict the formation energy of arbitrary material compositions without the need for any structure information.

IV. Development and Functionality

A. Data

We used the same dataset as used in [13]. It was generated as a result of a large number of DFT calculations on crystalline compounds from the Inorganic Crystal Structure Database (ICSD). To better teach the model about interactions between all pairs of elements, the binary phase diagrams generated by the DFT calculations were discretized to cover $A_{0.05}B_{0.95}$ to $A_{0.95}B_{0.05}$, for all pairs of elements A and B. In other words, formation energy was obtained at different fractions of all AB systems. The final dataset consisted of 83,728 chemical compositions of crystalline compounds, which corresponded to 9,324 stable ternary compounds and 74,404 discretized points on binary A-B phase diagrams. One attribute for each element was defined that represented the fraction of that element in the compound based on its stoichiometry. The number of such attributes was 112, one for each element. For Fe_2O_3 (ferric oxide or iron(III) oxide), the value of Fe_f would be 0.4 and O_f would be 0.6. Remaining 110 elemental fractions (e.g. Na_f, Cl_f, etc. would be set as 0. Some additional attributes that are based on elemental properties and derivable by composition alone were included in a bid to capture the general chemistry of the compound. These were 17 in number and are listed in Table I, along with their values for Fe_2O_3.

TABLE I
Composition-derived attributes

Attribute	*Value for Fe_2O_3*
Average atomic mass	0.4x55.845 + 0.6x15.999 = 31.94
Average column on periodic table	0.4x8 + 0.6x16 = 12.8
Average row on the periodic table	0.4x4 + 0.6x2 = 2.8
Maximum difference in atomic number	26 - 8 = 18
Average atomic number	0.4x26 + 0.6x8 = 15.2
Maximum difference in atomic radii (pm)	140 - 60 = 80
Average atomic radius	0.4x140 + 0.6x60 = 92.0
Maximum difference in electronegativity	3.44 - 1.83 = 1.61
Average electronegativity	0.4x3.44 + 0.6x1.83 = 2.474
Average number of s valence electrons	0.4x4 + 0.6x2 = 2.8
Average number of p valence electrons	0.4x0 + 0.6x4 = 2.4
Average number of d valence electrons	0.4x6 + 0.6x0 = 2.4
Average number of f valence electrons	0.4x0 + 0.6x0 = 0.0
s fraction of valence electrons	2.8 / (2.8+2.4+2.4+0.0) = 0.368
p fraction of valence electrons	2.4 / (2.8+2.4+2.4+0.0) = 0.316
d fraction of valence electrons	2.4 / (2.8+2.4+2.4+0.0) = 0.316
f fraction of valence electrons	0.0 / (2.8+2.4+2.4+0.0) = 0.0

B. Methods

We used 30 regression schemes including both direct application of regression techniques and constructing their ensembles using ensembling techniques (compatible combinations). Evaluation metrics included the coefficient of correlation (R), explained variance (R^2), Mean Absolute Error (MAE), Root Mean Squared Error ($RMSE$), Relative Absolute Error (RAE), and Root Relative Squared Error ($RRSE$).

C. Results

Table II presents the top five modeling techniques with respect to MAE. All results are based on 10-fold cross-validation. In addition, the training and testing times for each model, and the model size is also listed. WEKA software [14] verison 3.7.13 was used for all analytics with default parameters, unless otherwise stated. The predictive model used in [13] ranks 4^{th} in our current experiments. Statistical

TABLE II

TOP FIVE MODELING TECHNIQUES (10-FOLD CROSS-VALIDATION RESULTS)

Modeling Scheme	*MAE*	*R*	R^2	*RMSE*	*RAE*	*RRSE*	*TrainTime*	*TestTime*	*ModelSize*
	(eV/at)			*(eV/at)*	*(%)*	*(%)*	*(s)*	*(s)*	*(bytes)*
RotationForest_RandomTree	**0.0400**	**0.9887**	**0.9775**	**0.0857**	**10.75**	**15.13**	175.36	9.21	36748860
RandomCommittee_RandomTree	0.0442	0.9869	0.9740	0.0925	11.89	16.35	27.09	0.18	41176273
RandomForest	0.0460	0.9879	0.9759	0.0911	12.36	16.09	159.78	2.31	285658456
RotationForest_REPTree	0.0470	0.9866	0.9734	0.0934	12.64	16.51	291.69	8.41	11057773
RandomCommittee_REPTree	0.0472	0.9855	0.9712	0.0965	12.70	17.05	172.34	0.08	8802355

TABLE III

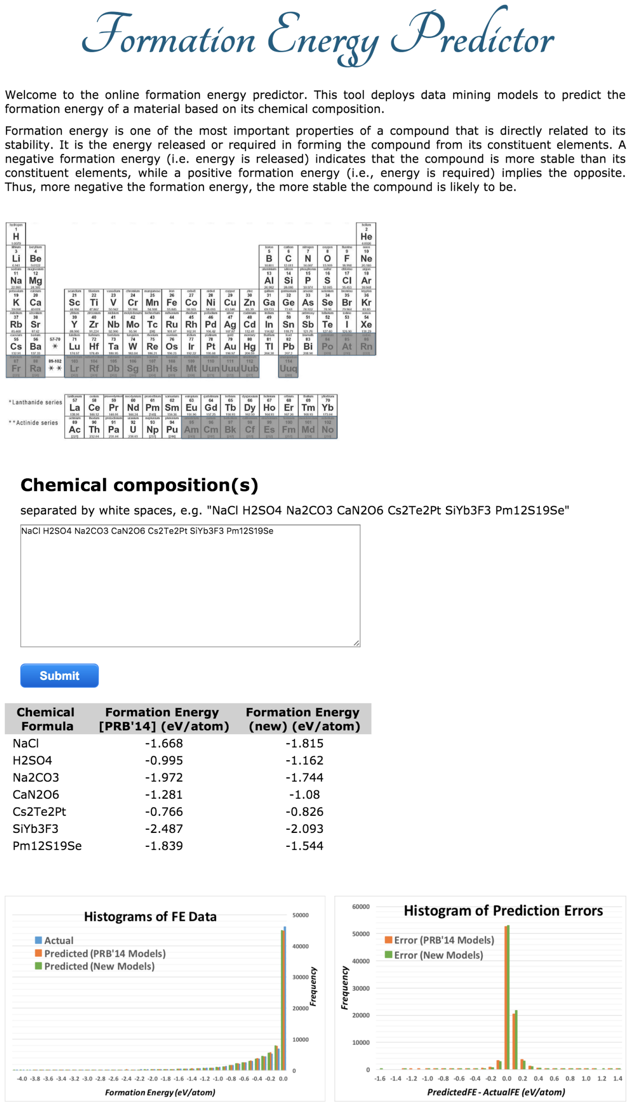
TOP FIVE MODELING TECHNIQUES WITH PART OF TERNARY COMPOSITIONS USED AS TESTING SET

Modeling Scheme	*MAE*	*R*	R^2	*RMSE*	*RAE*	*RRSE*
	(eV/at)			*(eV/at)*	*(%)*	*(%)*
RotationForest_RandomTree	**0.1343**	0.9744	0.9495	0.2035	**15.90**	16.89
RotationForest_REPTree	0.1389	0.9751	0.9508	**0.2025**	16.45	**16.81**
RandomForest	0.1393	**0.9752**	**0.9510**	0.2094	16.49	17.38
RandomCommittee_RandomTree	0.1399	0.9716	0.9440	0.2145	16.56	17.80
RandomCommittee_REPTree	0.1434	0.9690	0.9390	0.2204	16.97	18.29

significance testing revealed that the MAE obtained by the best model ($RotationForest\_RandomTree$) is significantly lower than the MAE from all other models at p=0.05.

It is important to note that since we had discretized all binary systems while building the training data, we expect the above accuracy numbers to be over-optimistic. This is because having the discretized binary compositions of a given system split across training and testing sets could artificially boost the accuracy. We thus performed a more difficult test for these models in order to evaluate their true predictive power by withholding 5% of the training data that consisted entirely of ternaries (compounds with three elements, e.g. $CaCO_3$) as the test set, and redid the entire modeling comparison with all the 30 modeling configurations. Results for the above-described setting are summarized in Table III. As expected, the accuracy went down, but the models were still found to be very useful for predictive purposes (R^2 of ∼0.95 and RAE of ∼15%). Rotation forest ensembling with random tree as the base modeling technique still gave the best MAE=0.1343 eV/atom, which was again found to be significantly better than all other techniques at p=0.05.

It is important to remember that DFT itself is a simulation technique and has discrepancies with experimentally observed formation energy values. The current MAE between DFT and experiment is 0.136 eV/atom [15], which is comparable to the MAE of the best model developed here. Further, it has also been reported that there is discrepancy even across experiments, with a surprising large MAE of 0.082 eV/atom [15]. Given these numbers, we believe that the machine learning models developed and deployed in this work could be very useful for quickly estimating the formation energy of materials with reasonable accuracy, which can help scan a large number of compositions in a short time without expensive DFT calculations or experiments.



Formation Energy Predictor

Welcome to the online formation energy predictor. This tool deploys data mining models to predict the formation energy of a material based on its chemical composition.

Formation energy is one of the most important properties of a compound that is directly related to its stability. It is the energy released or required in forming the compound from its constituent elements. A negative formation energy (i.e. energy is released) indicates that the compound is more stable than its constituent elements, while a positive formation energy (i.e., energy is required) implies the opposite. Thus, more negative the formation energy, the more stable the compound is likely to be.

H He
Li Be B C N O F Ne
Na Mg Al Si P S Cl Ar
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te I Xe
Cs Ba * Lu Hf Ta W Re Os Ir Pt Au Hg Tl Pb Bi Po At Rn
Fr Ra ** Lr Rf Db Sg Bh Hs Mt Uun Uuu Uub Uuq
*Lanthanide series La Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb
**Actinide series Ac Th Pa U Np Pu Am Cm Bk Cf Es Fm Md No

Chemical composition(s)

separated by white spaces, e.g. "NaCl H2SO4 Na2CO3 CaN2O6 Cs2Te2Pt SiYb3F3 Pm12S19Se"

NaCl H2SO4 Na2CO3 CaN2O6 Cs2Te2Pt SiYb3F3 Pm12S19Se

Chemical Formula	Formation Energy [PRB'14] (eV/atom)	Formation Energy (new) (eV/atom)
NaCl	-1.668	-1.815
H2SO4	-0.995	-1.162
Na2CO3	-1.972	-1.744
CaN2O6	-1.281	-1.08
Cs2Te2Pt	-0.766	-0.826
SiYb3F3	-2.487	-2.093
Pm12S19Se	-1.839	-1.544

Fig. 2. A screenshot of the deployed formation energy predictor.

D. Formation energy predictor

We have created an online formation energy predictor that can take as input a list of chemical compositions satisfying the charge balance condition respecting the common oxidation states of individual elements, and generate predictions of formation energy for each composition. Two models are deployed in this tool: $RotationForest\_REPTree$ (the one used in [13]) and $RotationForest\_RandomTree$ (that was found to be most accurate in our current experiments). The results are presented in the form of a sortable table of compositions with predicted formation energy values from the two models. In addition, histograms of actual/predicted formation energy values, and that of the prediction errors are also presented. The screenshot of the formation energy predictor is depicted in the Figure 2, and the tool is available online at: http://info.eecs.northwestern.edu/FEpredictor.

V. Significance

From a research perspective, this work investigates the applicability of predictive modeling techniques to predict material properties. In particular, here we focus on predicting the formation energy of a material given its chemical composition, by comparing over 30 supervised modeling configurations on a dataset of DFT calculations, and have identified the most accurate model till date for this problem, which is significantly better than the one used in [13].

From a practical point of view, we have deployed the most accurate predictive models in a user-friendly web-tool for easy access. Most data-driven models are, in general, not simple equations as in the case of something like linear regression, and are more like black box models that are not easily usable with traditional spreadsheet software. To make the developed predictive models for formation energy readily accessible for use by the materials science and engineering community, we have created an online formation energy predictor that can take arbitrary materials compositions and predict their stability. The primary advantage of this tool is the capability of quickly and accurately predicting formation energy using just the chemical composition of the material without needing any structure information, which is hard to obtain and is an essential prerequisite for performing a DFT calculation. The deployed tool is expected to be a useful resource for researchers and practitioners in the materials science and engineering community, to assist in their search for better materials with improved properties.

VI. Acknowledgment

This work is supported in part by the following grants: NIST award 70NANB14H012; AFOSR award FA9550-12-1-0458, DARPA award N66001-15-C-4036, NSF awards IIS-1343639, CCF-1409601; DOE awards DE-SC0007456, DE-SC0014330.

References

[1] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009. [Online]. Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/

[2] A. Agrawal and A. Choudhary, “Perspective: Materials informatics and big data: Realization of the fourth paradigm of science in materials science,” *APL Materials*, vol. 4, no. 053208, pp. 1–10, 2016.

[3] S. R. Kalidindi and M. D. Graef, “Materials data science: Current status and future outlook,” *Annual Review of Materials Research*, vol. 45, no. 1, pp. 171–193, 2015.

[4] K. Rajan, “Materials informatics: The materials ”gene” and big data,” *Annual Review of Materials Research*, vol. 45, no. 1, pp. 153–169, 2015.

[5] K. Gopalakrishnan, A. Agrawal, H. Ceylan, S. Kim, and A. Choudhary, “Knowledge discovery and data mining in pavement inverse analysis,” *Transport*, vol. 28, no. 1, pp. 1–10, 2013.

[6] A. Agrawal, P. D. Deshpande, A. Cecen, G. P. Basavarsu, A. N. Choudhary, and S. R. Kalidindi, “Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters,” *Integrating Materials and Manufacturing Innovation*, vol. 3, no. 8, pp. 1–19, 2014.

[7] R. Liu, A. Kumar, Z. Chen, A. Agrawal, V. Sundararaghavan, and A. Choudhary, “A predictive machine learning approach for microstructure optimization and materials design,” *Nature Scientific Reports*, vol. 5, no. 11551, 2015.

[8] R. Liu, Y. C. Yabansu, A. Agrawal, S. R. Kalidindi, and A. N. Choudhary, “Machine learning approaches for elastic localization linkages in high-contrast composite materials,” *Integrating Materials and Manufacturing Innovation*, vol. 4, no. 13, pp. 1–17, 2015.

[9] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, “A general-purpose machine learning framework for predicting properties of inorganic materials,” *npj Computational Materials*, vol. 2, no. 16028, 2016.

[10] A. Agrawal and A. Choudhary, “A fatigue strength predictor for steels using ensemble data mining,” in *Proceedings of 25th International Conference on Information and Knowledge Management (CIKM) (Demo)*, 2016.

[11] Materials Genome Initiative for Global Competitiveness, June 2011; OSTP 2011.

[12] Materials Genome Initiative Strategic Plan, National Science and Technology Council Committee on Technology, June 2014.

[13] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, “Combinatorial screening for new materials in unconstrained composition space with machine learning,” *Physical Review B*, vol. 89, no. 094104, pp. 1–7, 2014.

[14] M. Hall, E. Frank *et al.*, “The weka data mining software: An update,” *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[15] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, “The open quantum materials database (oqmd): assessing the accuracy of dft formation energies,” *npj Computational Materials*, vol. 1, p. 15010, 2015.