# Formulae for Attributes

As described in the manuscript, the first step of our method to create a predictive model of material properties is to compute attributes based on the composition of materials. These attributes are designed to enable a machine learning algorithm to construct general rules that can possibly "learn" chemistry and reflect some kind of chemical intuition. These attributes fall into four categories:

## Stoichiometric Attributes (6 in total)

These attributes capture the fraction of the elements present and are not affected by what those elements are. All are based on $L^p$ norms (i.e. $\|x\|_p = (\sum_{i=0}^{n}|x_i|^p)^{1/p}$) of a vector representing the atomic fraction of the material corresponding to each element. In this work, we use the p=0 norm (which is equivalent to the number of components) and the p=2, 3, 5, 7, and 10 norms. Such a broad range was selected to create attributes that respond to changes in fractions with varied strengths. As an example, the p=7 norm for $Fe_2O_3$ is:

$$\|x\|_7 = \left(\left(\frac{2}{5}\right)^7 + \left(\frac{3}{5}\right)^7\right)^{1/7} \cong 0.605$$

## Elemental-Property-Based Attributes (115 in total)

Most of the attributes created using our method are based on statistics of the elemental properties listed in Table S1. For each property, the minimum, maximum, and range of the values of the properties of each element present in the material is computed along with the fraction-weighted mean, average deviation, and mode (i.e. the property of the most prevalent element). The mean and average deviation are calculated using the following formulae:

$$\bar{f} = \sum x_i f_i \tag{S1}$$

$$\hat{f} = \sum x_i |f_i - \bar{f}| \tag{S2}$$

where $f_i$ is the property of element *i*, $x_i$ is the atomic fraction, $\bar{f}$ is the mean, and $\hat{f}$ is the average deviation. As an example, the mean of and average deviation in the atomic number of $Fe_2O_3$ are:

$$\bar{f} = \frac{2}{5}(26) + \frac{3}{5}(8) = 15.2$$

$$\hat{f} = \frac{2}{5}|26 - 15.2| + \frac{3}{5}|8 - 15.2| = 8.16$$

## Valance Orbital Occupation Attributes (4 in total)

These attributes are the fraction-weighted average of the number of valance electrons in each orbital divided by the fraction-weighted average of the total number of valance electrons. This attribute is exactly equivalent to the one employed by Meredig, Agrawal *et al.*[1] As an example, the fraction of p electrons for $Fe_2O_3$ is computed by

$$F_p = \frac{\frac{2}{5}(0) + \frac{3}{5}(4)}{\frac{2}{5}(8) + \frac{3}{5}(6)} = \frac{6}{17} \cong 0.352$$

### Ionic Compound Attributes (3 in total)

These attributes are designed to determine whether a material is ionically bonded. The first measure is a Boolean denoting whether it possible to form a neutral, ionic compound assuming each element takes exactly one of its common charge states. The other two are based on the "ionic character" of a binary compound, which is computed from the electronegativity difference between its two constituent elements using the relation

$$I(X_A, X_B) = 1 - \exp(-0.25(X_A - X_B)^2) \quad \text{(S3)}$$

where $I$ is the fraction of ionic character, $X_A$ is the electronegativity of element A, and $X_B$ is the electronegativity of element B.[2] The first attribute we used is the maximum ionic character between any two elements in the material. The second is the mean ionic character, which is computed using

$$\bar{I} = \sum x_i x_j * I(X_i, X_j) \quad \text{(S4)}$$

## Comparison of Attributes Sets

To assess the benefit of our expanded attribute set, we compared the predictive accuracy of models created using only the fractions of each element as attributes, using the attribute set proposed by Meredig, Agrawal *et al.*,[1] and using the attributes set proposed in this work. In order to only test the effect of changing the attribute set, we employed the same machine learning algorithm in each case: an ensemble of reduced-error-pruning decision trees created using the rotation forest technique.[3] We used 10-fold cross-validation to estimate the predictive mean absolute error for a model trained on the DFT-computed formation energy, band gap energy, and volume of 228676 compounds taken from the OQMD.[4] As shown in Figure S1, models created using our attributes are, on average, 81% more accurate than models created using only the atomic fractions of elements and 25% better than those using the attribute set of Ref. 1, which clearly shows the benefit of expanding the attribute set.

## Software Used to Perform Machine Learning

All machine learning models described in the associated manuscript, with the exception of those produced using symbolic regression, were created, trained, and evaluated using the Materials Agnostic Platform for Informatics and Exploration (Magpie). This software library package serves to integrate all aspects of constructing predictive models for material properties into one program and enables users to perform all of the requisite tasks through a simple, interactive text interface. As long as users can install the Java Runtime Environment (which is available for many systems), they will be to run this code and, given the provided datasets and scripts, replicate the results of our study with ease.

In the ZIP archive provided with this document, we have provided the Java archive for Magpie (Magpie.jar) and all of its required software libraries, which includes the Weka machine learning library.[5] Documentation for both the Java and text interfaces of this program is available online.[6] The source code for this program is freely available under a permissive license, and technical details of this code will also be described in a future publication.

## Hierarchical Model used to Predict Band Gap Energies

A representation of the hierarchical model used to predict band gap energies in the "Accurate Models for Arbitrary Properties" section of the manuscript is shown in Figure S1. Each submodel employed by

this composite model was an ensemble of Reduced-Error Pruning Trees created using the random subspace approach. This composite model was chosen because it was found to most accurately locate compounds with a band gap between 0.9 and 1.7 eV out of around 50 other models. The input file and code used to create and test this model are available in the ZIP archive provided with this document.

# References


1. Meredig, B., Agrawal, A., *et al.* Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89,** 094104 (2014).

2. Callister, W. D. *Materials Science and Engineering: An Introduction*. (Wiley, 2007).

3. Rodríguez, J. J., Kuncheva, L. I. & Alonso, C. J. Rotation forest: A new classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell.* **28,** 1619–30 (2006).

4. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *Jom* **65,** 1501–1509 (2013).

5. Hall, M. *et al.* The WEKA data mining software. *ACM SIGKDD Explor. Newsl.* **11,** 10 (2009).

6. http://bitbucket.org/wolverton/magpie

7. Wierzbicki, A. P. A mathematical basis for satisficing decision making. *Math. Model.* **3,** 391–405 (1982).

8. http://reference.wolfram.com/language/note/ElementDataSourceInformation.html

9. Villars, P., Cenzual, K., Daams, J., Chen, Y. & Iwata, S. Data-driven atomic environment prediction for binaries using the Mendeleev number. *J. Alloys Compd.* **367,** 167–175 (2004).

## Figures

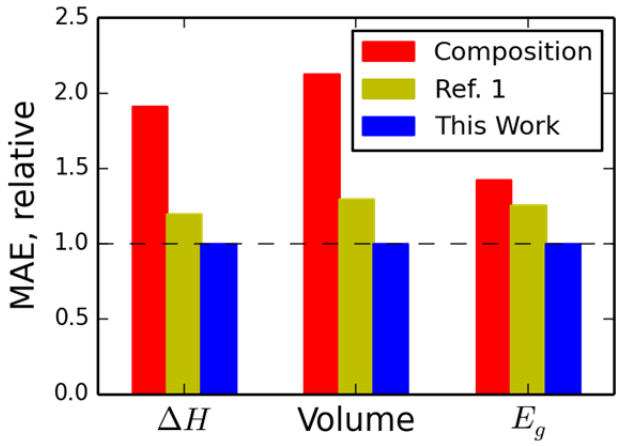


**Figure S1.** Relative magnitude of mean absolute errors in cross-validation of machine learning models that used three different attribute sets: only the composition (i.e., atomic fractions of each element), the attribute set of Meredig, Agrawal *et al.*,[1] and the set proposed in this work. Each model used ensembles of decision trees and were trained on the DFT formation energy, volume, and band gap energy 228676 compounds.

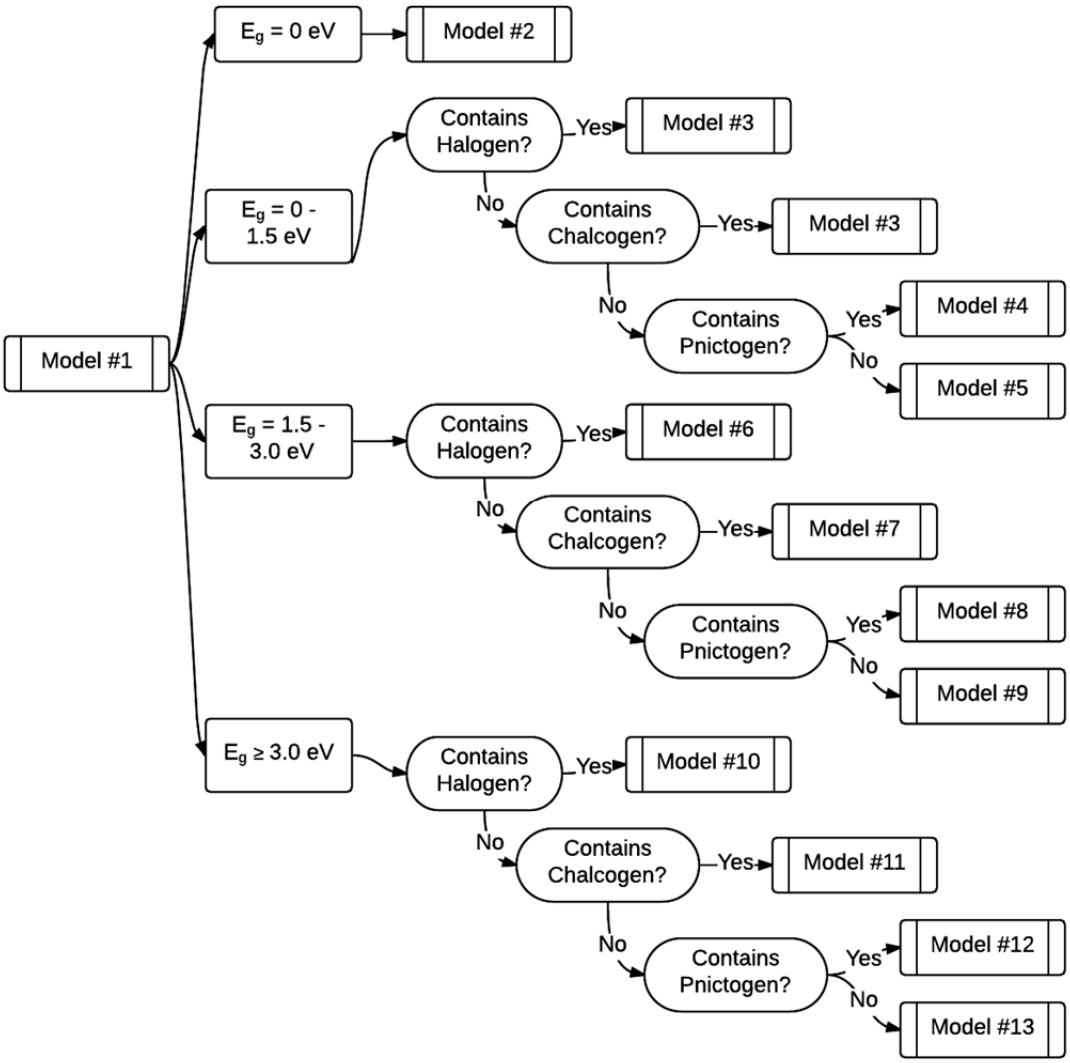


**Figure S2**. Hierarchical model used to predict band gap energies of crystalline compounds. Each rectangle with rounded corners represents a machine learning model. The model on the far left (Model #1) is trained to predict the range on which the band gap was most likely to lie. The models on the right are trained to predict actual value of the band gap energy. Depending on results of Model #1 and the composition of an entry, a different machine model would be used. For example, Model #3 will be used for all halogen-containing compounds predicted to have a band gap energy between 0 and 1.5 eV by Model #1.

# Tables

**Table S1.** Elemental properties used to compute elemental-property-based attributes. Elemental property is taken from that dataset available with the Wolfram programming language,[8] unless otherwise specified

| | | | | |
|---|---|---|---|---|
| Atomic Number | Mendeleev Number[9] | Atomic Weight | Melting Temperature | Column |
| Row | Covalent Radius | Electronegativity* | # s Valence Electrons | # p Valence Electrons |
| # d Valence Electrons | # f Valence Electrons | Total # Valance Electrons | # Unfilled s States† | # Unfilled p States† |
| # Unfilled d States† | # Unfilled f States† | Total # Unfilled States† | Specific Volume of 0 K Ground State‡ | Band Gap Energy of 0 K Ground State‡ |
| Magnetic Moment (per atom) of 0 K ground state‡ | Space Group Number of 0 K Ground State‡ | | | |

*Electronegativities for Eu, Yb, Tb, Pm taken to be the average of that of the element with one greater and one less atomic number (e.g. the average of Sm and Gd is used for Eu)

†Computed as the number of electrons in a partially-occupied orbital subtracted from the total number of electrons allowed in that orbital. Unoccupied orbitals always count as 0. Example: an element with a electronic configuration of $[Ar]3d^3 4s^2$ has 0 unfilled s orbitals, 7 filled d orbitals, and 0 unfilled p and f orbitals by the measure defined here.

‡Data taken from OQMD.org