Software

All of the software necessary to replicate the results of this work is included as part of the Materials-Agnostic Platform for Informatics and Exploration (Magpie), which is freely available under an open-source license.[1] Example script files and the datasets used in this study are also included in a ZIP file supplied along with this document.

Attribute Formulae

As described in the main body of the text, our method employs attributes that fall into several distinct categories. This section describes each of these categories and the attributes in greater detail.

Effective Coordination Number Attributes

We define the effective coordination number of an atom as a function of the sizes of faces on its Voronoi cell:

$$CN_{eff} = \frac{(\sum_n A_n)^2}{\sum A_n^2} \quad (1)$$

where A_n is the area of face n. For a shape with equally-sized faces, the effective coordination number is exactly equal to the number of faces. Additionally, the introduction of a small face will only have a small influence on both the total surface area and the sum of squared surface areas and, therefore, a small change in the effective coordination number.

We compute the maximum, minimum, mean, and mean absolute deviation in coordination number as attributes. The mean absolute deviation of a quantity is computed as:

$$\hat{f} = \frac{1}{N}\sum_i |f_i - \bar{f}| \quad (2)$$

where $\hat{f}$ is the mean absolute deviation, f_i is the value of sample i, N is the number of samples, and $\bar{f}$ is the mean. The mean absolute deviation was selected to measure variance in the properties of atoms in a structure (e.g., coordination number) because it is insensitive to unit cell selection:

1) All symmetrically-distinct images of an atom in a lattice will have the same property
2) All unit cell choices have the same proportion of each type of symmetrically-distinct atoms
3) Consequently, the mean property for any choice of unit cell will have the same mean
4) Therefore, the deviation between the mean property in a unit cell and the property for each atom is unchanged by unit cell choice
5) If the deviation in the property properties do not change with unit cell choice, the mean deviation will also be insensitive to unit cell choice

Structural Heterogeneity Attributes

These attributes are designed to reflect variation in the shape of local bonding environments. The first set of attributes are based on maximum, minimum, and mean absolute deviation (see Eq. 2) in average bond length of each atom. Here, we define bond length as the Voronoi-face-area-weighted average of the distance between an atom and each neighbor:

$$\overline{l_i} = \frac{\sum A_n * \|\overrightarrow{r_n} - \overrightarrow{r_i}\|_2}{\sum A_n} \quad (3)$$

where $\overline{l_i}$ is the mean bond length of an atom *i*, $\vec{r_i}$ is the position of atom *i*, and A_n and $\vec{r_n}$ is the area and of the n^{th} neighbor of atom *i*. To make these attributes insensitive to scaling the volume of a unit cell, they are all normalized by the average $\overline{l_i}$ of all atoms.

Additionally, we create attributes based on the mean, maximum, minimum, and mean absolute deviation in the bond length variance of each atom. The bond length variance captures the distribution in bond lengths between each neighbor of an atom, and is computed by:

$$\widehat{l_i} = \frac{\sum |A_n * \|\vec{r_n} - \vec{r_i}\|_2 - \overline{l_i}|}{\overline{l_i} * \sum A_n} \tag{4}$$

where $\widehat{l_i}$ is the bond length variance and other terms are the same as in Eq. 3. As the variance is normalized by the mean bond length, this term is also insensitive to scaling the volume of the unit cell.

The mean absolute deviation of the volume of the Voronoi cell about each atom is also used as an attribute. This attribute is normalized by the mean volume of all cells in order make it insensitive to changes in the cell volume.

Chemical Ordering Attributes

These attributes are based on Warren-Cowley ordering parameters, which measure how the distribution of atoms on a lattice differs from purely-random.[2] We first compute all *N*-length, non-backtracking paths originating from each atom in the crystal. For each step in these paths, we assign the step a fractional weight corresponding to the size of its face compared to all faces corresponding to other possible (i.e., non-backtracking steps):

$$w_n = \frac{A_n}{\sum_a A_a - \sum_b A_b} \tag{5}$$

where A_n is the area of the face normal to the direction of the step, and the two sums in the denominator are over the faces corresponding to all allowed and back-tracking steps, respectively. The total weight of a path is determined by multiplying the weight of these steps, which results in the sum over the weights of all possible paths being equal to 1. Consequently, the weight for each path can be envisioned as the probability a walker will take a certain path if its probability of making each step is proportional to the area of the face being traversed.

After determining the paths and their effective weights, sum the total weight of all paths ending on each type of atom. If the arrangement of atoms on the lattice is purely random, the likelihood of a type of atoms being at the end of any path is equal to the fraction of atoms of that type in the material. Consequently, the Warren-Cowley ordering parameter can be expressed as

$$\alpha(t, s) = 1 - \frac{\sum_p w_p \delta(t - t_p)}{x_t} \tag{6}$$

where $\alpha(t, s)$ is the weighted ordering parameter for type t in the s^{th} shell, x_t is the atomic fraction of type *t* in the crystal, w_p is weight of path p, t_p is the type of atom at the end of path p, and δ is the delta function.

To generate attributes that describe the entire cell, the mean of the absolute values of the ordering parameter over all types and each atom is computed for the 1st, 2nd, and 3rd nearest-neighbor shells.

Maximum Packing Efficiency

The largest sphere centered on the position of an atom that can fit inside its Voronoi cell has a radius equal to the distance between the center of an atom and center of the closest face of the cell. To compute the maximum packing efficiency, the sum of the largest possible spheres for each atom is divided by the cell volume.

Local Environment Attributes

These attributes constitute the majority of the structure-based attributes used in our method, and are based on the difference in elemental properties between an atom and each neighbor. The local property difference for each atom is defined as the face-area-weighted mean of the absolute difference in elemental properties between an atom and each of its neighbors

$$\delta_p = \frac{\sum_n A_n * |p_n - p_i|}{\sum_n A_n} \quad (7)$$

where δ_p is the local property difference (for an elemental property, *p*), p_i is the elemental property of the central atom, p_n is the property of neighboring atom n, and A_n is the area of the face corresponding to neighbor *n*.

To create attributes, we compute the mean, mean absolute deviation (Eq. 2), maximum, and minimum in the local property difference for each atom considering 22 different elemental properties (listed in Table **S1**).

Composition-Based Attributes

We also include attributes that are only dependent on the composition of the compound and not its structure. For this purpose, we use attributes described in recent work by Ward *et al.*,[3] which include:

1. **Stoichiometric Attributes** that depend only on the relative fractions of elements in the structure, and not what those elements actually are.
2. **Elemental Property Attributes** based on the mean, maximum, minimum, mode, range, and mean absolute deviation in 22 different elemental properties
3. **Valence Shell Attributes** that are based on the fraction of electrons in the *s*, *p*, *d*, and *f* shells of the constituent elements
4. **Ionicity Attributes**, which include whether it is possible to form a charge-neutral ionic compound at a certain composition and "iconicity" measures based on the electronegativity differences

Deformation Stability Test

To test the stability of our attributes against small perturbations in the crystal structure, we measured the changes in attributes upon introducing a small strain in a unit cell and small displacements on the position of each atom. Specifically, we applied random strains of up to ±5% on each direction and random displacements of up to 0.5 Å onto each atom in a 2x2x2 supercell of a $SrTiO_3$ perovskite. When then created structures that interpolated between -20% and 120% of this randomly-selected displacement vector. As shown in Figure S1, all attributes change continuously as a function of displacement vector.

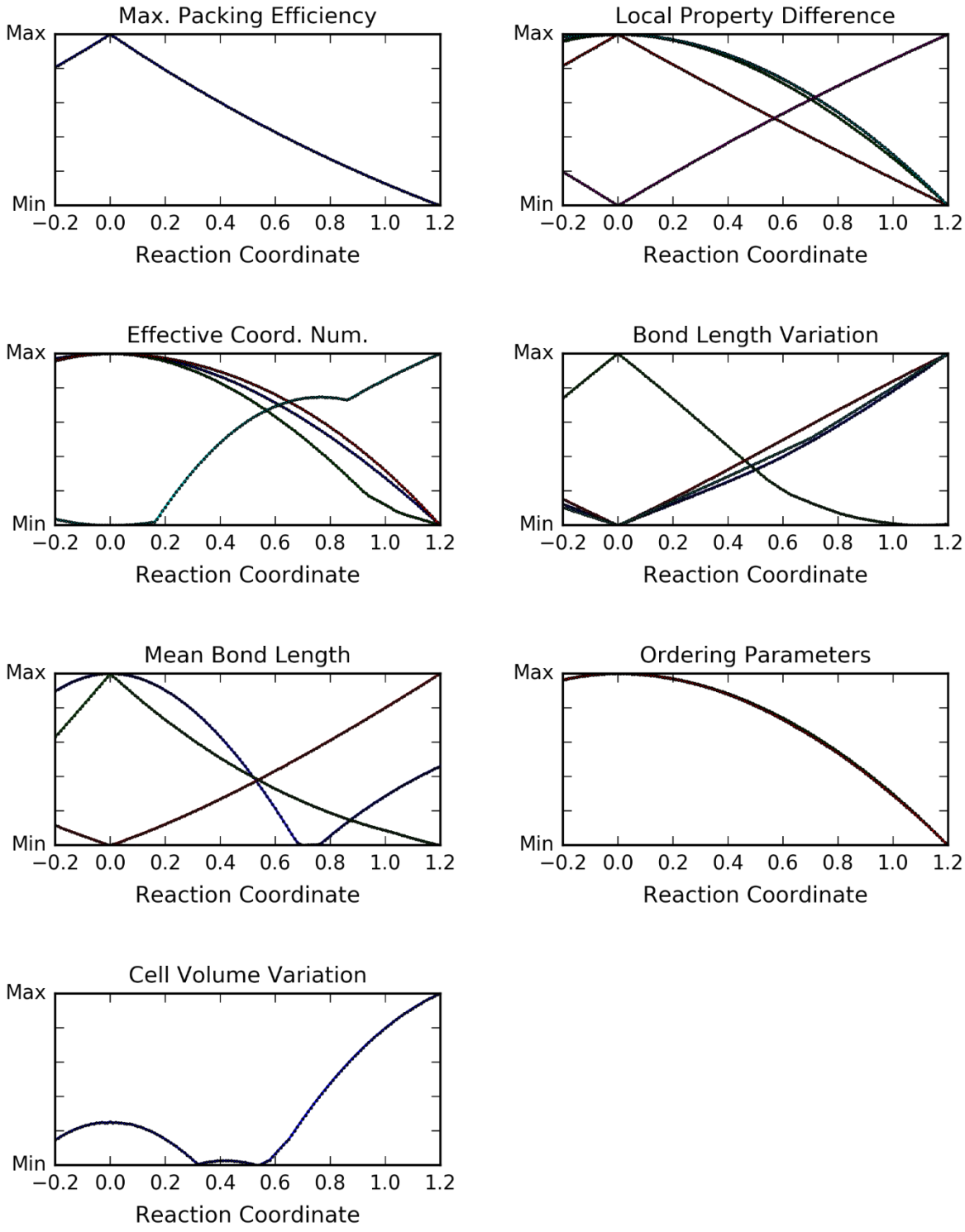

Figure S1. Variation in representative examples of each kind of Voronoi-tessellation-based attribute as a function of position along a randomly assigned displacement. A coordinate of 0 corresponds to an undistorted, cubic perovskite with composition $SrTiO_3$.

Coulomb Matrix Implementation

We used the Sine matrix formation of the Coulomb matrix technique demonstrated by Faber *et al.*[4] As the source code used by Faber *et al*. is not publicly-available (at least at the time of writing this manuscript), we implemented a version of this algorithm using Magpie. Our implementation uses Cholesky Decomposition, using the Apache Commons Math Library, to perform the Kernel-Ridge Regression. All other details are identical to the approach described in the paper by Faber *et al*.

As the dataset used in this study was not available when this work was performed, we validated our implementation of the Coulomb Sine matrix approach by attempting to emulate the test performed by Faber *et al.*, which is similar to the one described in the "Comparison to Existing Techniques with Cross-Validation" section of the paper associated with this document. To match the Materials Project dataset used by Faber *et al.*, we randomly selected entries from the OQMD[5] that correspond to compounds from the ICSD to create a training and validation set. Faber *et al*. report a mean absolute error of 0.37 eV/atom with a training set size of 3000. We found a mean absolute error of 0.375 ± 0.04 eV/atom (average of 20 tests) for our implementation, which we assumed was small enough to be attributed to differences in the training data (i.e., using OQMD rather than Materials Project data).

To account for the similarity between our tests and those described by Faber *et al*., we retuned the metaparameters for the Sine matrix using those reported in their original work of $\sigma = 4 \cdot 10^4$ and $\lambda = 10^{-4}$ as a starting guess. Using a logarithm grid with a spacing factor of 2 for λ and 10 for σ, we found that a combination of $\sigma = 6.3 \cdot 10^4$ and $\lambda = 10^{-2}$ produced the lowest MAE (0.37 eV/atom) for a test with a training set size of 3000 and a test set of 1000 entries out of 51 tested combinations.

PRDF Implementation

We implemented the PRDF method of Schütt *et al.* using Magpie. As with the Coulomb matrix approach, our implementation uses Cholesky Decomposition to perform the ridge regression.

For metaparameters for this model, which include λ and σ parameters from the ridge regression and the cutoff distance and bin spacing of the radial distribution function, we used a set of metaparameters found to result in the lowest error in a model with a training set size of 3000 entries using the methods described in "Validation using Experimentally-Studied Compounds." The λ and σ parameters were varied on a logarithmic grid with spacings of $10^{0.5}$ and $10^{0.25}$, respectively. The cutoff distance and number of bins were varied from 1 to 16 Å in steps of 2 Å and from 2 to 20 in steps of 2, respectively. We found the set of parameters of $\lambda = 3 \cdot 10^{-5}$, $\sigma = 100$, a distance cutoff of 16 Å, and 8 RDF bins to be the optimal choice out of 610 combinations considered.

References

[1]https://bitbucket.org/wolverton/magpie
[2] J. Cowley, Phys. Rev. **77**, 669 (1950).
[3] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, Npj Comput. Mater. **2**, 16028 (2016).
[4] F. Faber, A. Lindmaa, O.A. von Lilienfeld, and R. Armiento, Int. J. Quantum Chem. **115**, 1094 (2015).
[5] J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, JOM **65**, 1501 (2013).
[6]http://reference.wolfram.com/language/note/ElementDataSourceInformation.html
[7] P. Villars, K. Cenzual, J. Daams, Y. Chen, and S. Iwata, J. Alloys Compd. **367**, 167 (2004).

Tables

Table S1. (Reproduced from Ref. 3) Elemental properties used to compute elemental-property-based attributes. Elemental property is taken from that dataset available with the Wolfram programming language,[6] unless otherwise specified

Atomic Number	Mendeleev Number[7]	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