

## Supplementary Materials for

### The phase stability network of all inorganic materials

Vinay I. Hegde, Muratahan Aykol\*, Scott Kirklin, Chris Wolverton\*

\*Corresponding author. Email: c-wolverton@northwestern.edu (C.W.); murat.aykol@tri.global (M.A.)

Published 28 February 2020, *Sci. Adv.* **6**, eaay5606 (2020)

DOI: 10.1126/sciadv.aay5606

#### This PDF file includes:

Section S1. Calculation of the  $T = 0$  K universal phase diagram

Section S2. Degree distribution of the network of all materials

Section S3. New information encoded in the nobility index

Table S1. Sample compute times for calculating the existence of a tie-line between two phases.

Fig. S1. Fitting node connectivity data to candidate distributions.

Fig. S2. Comparison of nobility index versus common elemental properties.

Fig. S3. Comparison of number of compounds formed by an element versus its node degree.

## Supplementary Materials

We describe the approach used to calculate the universal phase diagram, its network representation and determining the node degree distribution in the network in the following sections S1–S2.

### Section S1. Calculation of the $T = 0$ K universal phase diagram

The  $T = 0$  K phase diagram for a given chemical space is determined by the so-called convex hull construction. A phase is thermodynamically stable iff it lies on (i.e. is a vertex of) the convex hull of  $T = 0$  K formation energies of all phases in the chemical space. And phases that are directly connected by a tie-line, i.e., phases that lie on the same facet of the convex hull, are in equilibrium with one another. Determining a binary A-B phase diagram requires constructing a 2-dimensional convex hull of formation energies of all  $A_xB_y$  compounds (composition  $x$  and formation energy being the two dimensions), a ternary A-B-C phase diagram requires constructing a 3-dimensional convex hull of formation energies of all  $A_xB_yC_z$  compounds (compositions  $x$  and  $y$ , and formation energy being the three dimensions), and so on. The determination of an  $d$ -nary phase diagram requires the construction of an  $d$ -dimensional convex hull of formation energies of all the  $N$  phases in the chemical space.

For low dimensions, i.e.  $d = 2$  or  $3$  (binary or ternary systems), finding the convex hull of  $N$  points (total number of phases) has a worst-case time complexity of  $\mathcal{O}(N \log N)$ . For higher dimensions, standard methods of determining convex hulls such as the Quickhull algorithm, have worst-case time complexities of  $\mathcal{O}(N^{\lfloor d/2 \rfloor})$  (39). For random data, even the average-case time complexity at higher dimensions scales as  $\mathcal{O}(\log^{d-1} N)$ , i.e. exponentially with  $d$  (43). Such scaling behaviors mean that for moderately large number of points  $N$  and dimensions  $d$ , finding the convex hull becomes increasingly practically challenging. For instance, to find the convex hull of all known inorganic materials, even restricting ourselves to experimentally reported compounds in the Open Quantum Materials Database (OQMD),  $N \approx 40,000$  and  $d = 89$ , making the calculation of the convex hull practically impossible with a traditional single-shot approach.

We tackle this challenge of calculating the convex hull at high-dimensions by using a divide-and-conquer approach. While the representational complexity of the convex-hull increases at least exponentially with  $d$ , we know from the set of existing materials that not many of them are high-dimensional by themselves. In fact, 99.5% of materials in the OQMD have 4 unique elemental components or fewer. Since the stability of a material is determined only within the chemical subspace of elements that it is made of, we first determine the vertices (i.e. stable materials) of the 89-dimensional convex hull at a reduced computational cost by computing the convex hulls in low-dimensional subspaces for each individual material separately. For instance, to determine if the compound  $\text{CaMnO}_3$  is stable, it is sufficient to construct the convex hull of all phases  $\text{Ca}_x\text{Mn}_y\text{O}_z$  in the Ca-Mn-O subspace. This process of constructing convex hulls separately for each unique chemical subspace yields all the vertices of the full convex hull:  $\sim 2.1 \times 10^4$  stable materials out of the  $> 5.5 \times 10^5$  total materials calculated in the OQMD. Having determined the vertices of the full convex hull, in the second stage, we exhaustively evaluate the existence of a tie-line between any given pair of stable compounds in the OQMD by constructing the convex hull of formation energies in the combined chemical spaces of such candidate pairs, rather than the full 89-dimensional space itself. For example, to determine whether there exists a tie-line between  $\text{Li}_2\text{O}$  and  $\text{NaCl}$ , we construct the Li-Na-Cl-O convex hull, and find that there indeed exists a  $\text{Li}_2\text{O}$ - $\text{NaCl}$  tie-line. In contrast, from a Na-K-F-Cl convex hull we find that  $\text{NaCl}$  and  $\text{KF}$ , in fact, “react” to form a  $\text{NaF}$ - $\text{KCl}$  two-phase equilibrium. Overall, we construct convex hulls for all  ${}^{2.1 \times 10^4}C_2 \approx 2.3 \times 10^8$  stable phase combinations, and find a total  $\sim 41 \times 10^6$  tie-lines.

The computational cost of constructing a convex hull for a unique chemical subspace is expectedly highly dependent on the number of components, and ranges from a few seconds to a few minutes on a standard desktop computer utilizing a single core (some sample times for checking if a tie-line exists between two known materials are provide in Table 1). With a conservative estimate of 15–20 seconds per tie-line, the total time required to exhaustively evaluate all possible tie-lines is more than 1 million CPU hours.

We then represent stable compounds as nodes and tie-lines as edges, thereby generating the “universal phase diagram” as a complete thermodynamic phase stability network of all inorganic materials. We use the *Qhull* library (39) as implemented in the *qmpy* package ([pypi.org/project/qmpy](https://pypi.org/project/qmpy)) for all the convex hull calculations reported in this work.

**Table S1. Sample compute times for calculating the existence of a tie-line between two phases.**

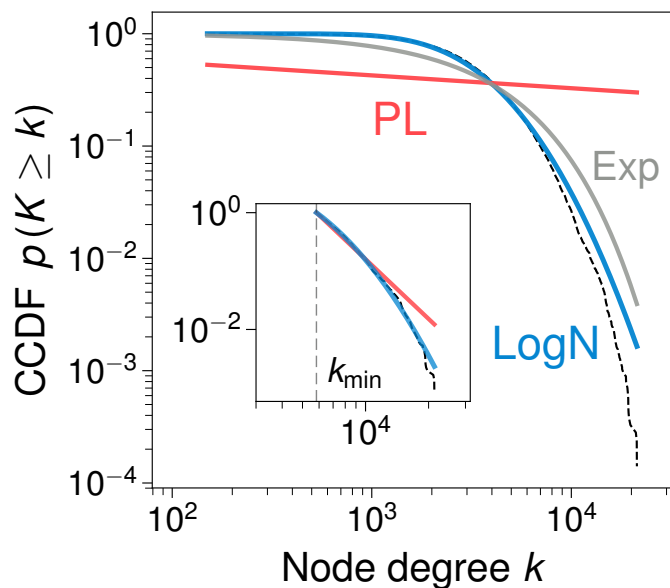
The time required is highly dependent on the number of components, i.e. unique elements in the combined chemical space, and further depends on the number of all known compounds in the chemical space. Each calculation was performed on a standard desktop computer utilizing a single core.

Phase 1	Phase 2	# Components	Time (s)
Na <sub>2</sub> O	KCl	4	~3
Fe <sub>2</sub> S <sub>3</sub>	Li <sub>2</sub> MnO <sub>4</sub>	5	~6
Li <sub>3</sub> PS <sub>4</sub>	SrTiO <sub>3</sub>	6	~8
Ba <sub>2</sub> Li <sub>3</sub> TaN <sub>4</sub>	LiCoO <sub>2</sub>	6	~14
Ba <sub>2</sub> Li <sub>3</sub> TaN <sub>4</sub>	NaCoO <sub>2</sub>	7	~32
Mn <sub>2</sub> Hg <sub>2</sub> SF <sub>6</sub>	Li <sub>4</sub> CrCoO <sub>6</sub>	8	~34
Mn <sub>2</sub> Hg <sub>2</sub> SF <sub>6</sub>	Ba <sub>2</sub> Ca <sub>3</sub> Tl <sub>2</sub> Cu <sub>4</sub> O <sub>12</sub>	9	~65

**Section S2. Degree distribution of the network of all materials**

The probability distribution of node connectivity (number of tie-lines a material has) in the phase stability network of all inorganic materials is heavy-tailed. We examine which of the common heavy-tailed distributions best fit our empirical data. In particular, several well-studied technological, social, and biological networks are thought to have power-law distributions. Is the thermodynamic network of materials similar to other common natural/man-made networks exhibiting power-law behavior or not? To answer this question, we directly compare pairs of heavy-tailed distributions using the method of log likelihood ratios as described in Clauset et al. (42). For the full materials network, we find that a lognormal distribution ( $\mu = 8.06$ ,  $\sigma = 0.65$ ) is the best fit by far (see Fig. 1).

We note that most empirical phenomena obey power-laws only for values greater than some minimum value, i.e. only the tail of the distribution follows a power-law. We investigate if this is indeed the case for the materials network. We find the optimal lower-bound for a power-law behavior,  $k_{\min}$ , for the materials network as the value that minimizes the Kolmogorov-Smirnov distance between the data and the fit (42). We find  $k_{\min}$  for the materials network = ~5800, and the power-law scaling parameter  $\alpha = 4.4$ . We note that a  $k_{\min}$  of 5800 retains only 17% of the overall materials network (i.e. only 17% of all materials have more than 5800 tie-lines each). Furthermore, even over this tail region of the degree distribution, a lognormal distribution is a better fit (see the inset in Fig. 1): the log likelihood ratio  $\mathcal{R}$  for power-law versus lognormal is  $-7.15$  with a  $p$ -value of 0. In other words, even in the  $k_{\min} = 5800$  region (tail) of the materials



**Fig. S1. Fitting node connectivity data to candidate distributions.** The complementary cumulative distribution function of the node degree in the network of all materials is shown as dashed black lines. Power-law (PL), lognormal (LogN), and exponential (Exp) distributions fit to the data are shown as solid red, blue, and grey lines, respectively. The inset shows power-law and lognormal fits to the *tail* of the degree distribution for degree  $k > k_{\min} = 5800$ .

network, the lognormal distribution fits the data far better than a power-law.

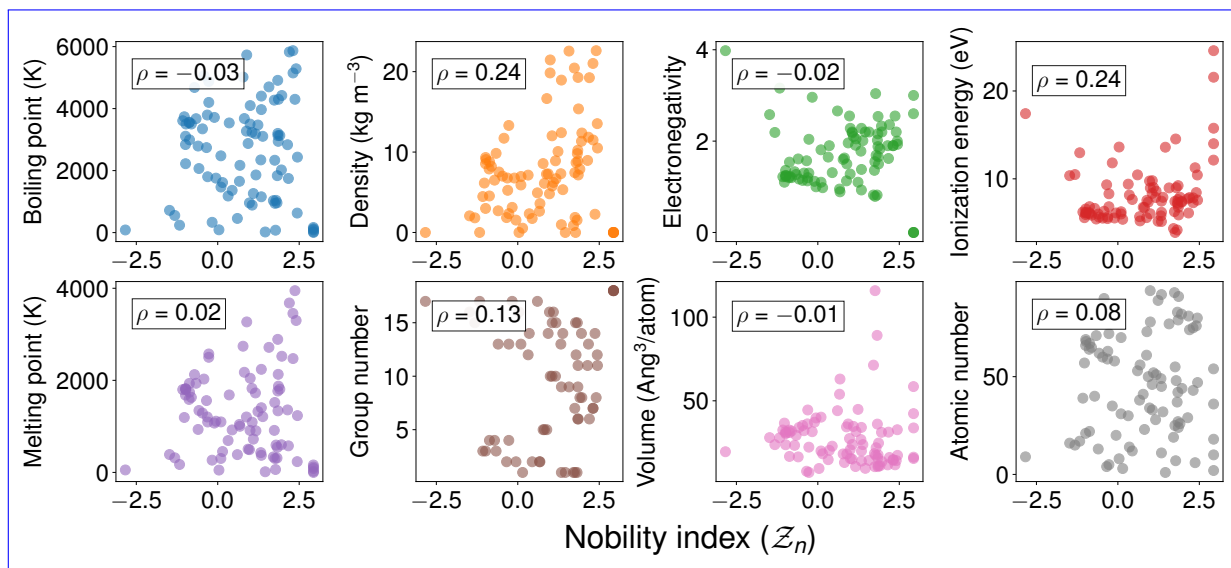
All analyses of fits of degree distributions mentioned above were performed with the *powerlaw* package (41). We note that the graph-theoretic analyses reported in this work (e.g. local clustering and centrality measures) performed with the *graph-tool* package (40), while requiring more than 8 G of memory, take a few hours on a standard desktop utilizing up to 4 cores.

### Section S3. New information encoded in the nobility index

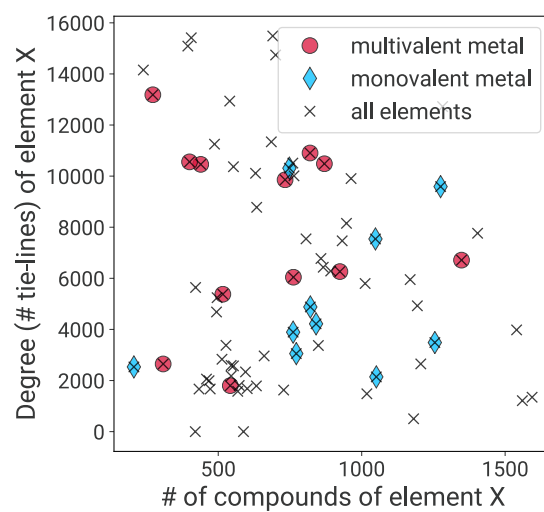
A comparison of the nobility index  $\mathcal{Z}_n$  of elements against elemental properties such as electronegativity, boiling point, melting point, atomic volume, etc., as collected by Ward et al. (33) shows little correlation between  $\mathcal{Z}_n$  and other properties, with Pearson correlation coefficients close to 0 for most properties (see Fig. 2 for a sample comparison set). This indicates that the nobility index defined in this work truly encodes new information about an element/a material not adequately captured by other common properties.

Further, data-driven metrics such as nobility index capture materials knowledge that is not immediately intuitive or is sometimes even counter-intuitive. For instance, intuition derived only from a few elements and some of their compounds may imply that multivalent elements (e.g.

transition metals) are likely to have a higher number of tie-lines than monovalent elements (e.g. alkali metals) simply by the virtue of a higher number of compound-forming possibilities. However, data from all materials known so far shows no correlation between number of compounds formed by an element and its total number of tie-lines (i.e. nobility; see Fig. 3). In fact, monovalent metals seem to form more compounds on average than their multivalent counterparts!



**Fig. S2. Comparison of nobility index versus common elemental properties.** There is little to no correlation between the nobility index of an element and any of its properties such as (counterclock-wise from top-left) boiling point, density, electronegativity, first ionization energy, atomic number, atomic volume, group in the periodic table, and melting point. The Pearson correlation coefficient  $\rho$  for each comparison is on the top-left of the corresponding panel.



**Fig. S3. Comparison of number of compounds formed by an element versus its node degree.** Multivalent metals indicated are all transition metals (Ti, V, Cr, Mn, Fe, Co, Ni, Mo, W, Hf, Pd, Pt), and monovalent metals indicated are mostly alkali/alkaline earth metals (Li, Na, K, Rb, Be, Mg, Ca, Sr, Al, Zn).