The Phase Diagram of All Inorganic Materials

Vinay I. Hegde,† Muratahan Aykol, Scott Kirklin, and Chris Wolverton
Department of Materials Science and Engineering, Northwestern University, Evanston, IL 60208
Toyota Research Institute, Los Altos, CA 94022

Understanding how the arrangement of atoms and their interactions determine material behavior has been the dominant paradigm in materials science. A complementary approach is studying the organizational structure of networks of materials, defined on the basis of interactions between materials themselves. In this work, we present the “phase diagram of all known inorganic materials”, an extremely-dense complex network of nearly $2.1 \times 10^4$ stable inorganic materials (nodes) connected with $41 \times 10^6$ tie-lines (edges) defining their two-phase equilibria, as computed via high-throughput density functional theory. We show that the degree distribution of this network follows a lognormal form, with each material connected to on average 18% of the other materials in the network via tie-lines. Analyzing the structure and topology of this network has potential to uncover new materials knowledge inaccessible from the traditional bottom-up (atoms to materials) approaches. As an example, we derive a data-driven metric for the reactivity of a material as characterized by its connectedness in the network, and quantitatively identify the noblest materials in nature.

INTRODUCTION

The study of complex systems by modeling them as networks, i.e. a set of connections (edges) linking discrete components (nodes or vertices) has many diverse examples from man-made systems such as the internet, transportation routes, collaborations among actors and scientists, citations in academic papers, friendship, to natural systems such as metabolism and protein-protein interactions in a cell, and neural network in an organism [1–4]. Despite significant variation in the individual components and interconnections, many of these networks show striking similarities in their topology and organizational principles, and the analyses of these topologies often provide new insights for each respective domain.

One of the major pursuits in materials science is unlocking the structure-property relationships in materials [5]. The dominant paradigm in such a pursuit has been the “bottom-up” study of the nature of the individual components making up a material at various scales, and interactions between such components, from electrons to atoms to interfaces, surfaces, grains and so on. An emergent alternative approach is to gain new materials knowledge by studying the similarities or interactions between materials themselves; in other words, a “top-down” view of materials as a network constructed by specific sets of rules [6, 7]. Such a data-driven paradigm has been made possible only recently with developments in high-throughput density functional theory (HT-DFT) and resulting large computational databases of materials properties [8–12].

The Open Quantum Materials Database (OQMD) [11, 13] is one such large database of computed properties of nearly all ordered inorganic materials observed to date (as collected in the Inorganic Crystal Structure Database repository) [14], and a large number of hypothetical compounds based on the decoration of commonly occurring structural prototypes—a total of more than 560,000 compounds. Here, using this database, we uncover a nearly complete phase diagram encompassing all known inorganic materials and more, by computing the 89-dimensional composition-energy convex-hull. With thermodynamically stable compounds represented as nodes and their inter-phase equilibria represented as edges (tie-lines), we show that this “phase diagram of all inorganic materials” is a remarkably dense and well-connected network with 41 million tie-lines between ~21,000 stable materials. By analyzing the properties of this phase diagram as a network, we gain further insight about what its topological features, such as diameter, mean-shortest paths, clustering and degree distribution imply for the collective behavior of the bulk thermodynamic interactions of inorganic materials. We demonstrate that the topological features, such as the form of the degree distribution, can be used to derive data-driven metrics, for example, to quantify the nobility of materials.

RESULTS AND DISCUSSION

The composition-formation energy convex-hull of all inorganic materials is high-dimensional, where each dimension is a chemical element. In the OQMD, there are 89 unique elements that appear in the materials it contains, hence the composition-energy convex-hull is 89 dimensional. Using a divide-and-conquer approach we sample the complete convex-hull (see the Methods section for details) to identify all stable materials and all two-phase equilibria between them. We find that this thermodynamic network of stable materials with ~21,300
nodes is remarkably dense with a total of nearly 41 million edges, and extremely well-connected with \( \sim 3.850 \) edges per node on average ("mean degree" \( \langle k \rangle \)). For comparison, \( \langle k \rangle \) for other widely-studied networks range from 1.4 (network of email messages) to 113.4 (film actors collaboration network) [1]. The connectance of the materials network, or the fraction of maximum possible number of edges that are actually present is 0.18. This is an important statistic for the design of "systems of materials", such as electrodes and electrolytes making up batteries [15] or coating materials separating two reactive materials [16] where longevity of the system relies on stable coexistence of components. This metric implies that, on average, a given stable compound can have a thermodynamically stable coexistence with 18% of all other stable compounds. Around 25% of these edges are between materials that share at least one element in common (hereafter, homochemical edges), whereas the rest of the edges are between materials that belong to completely different chemical subspaces that do not intersect (hereafter, heterochemical edges).

The characteristic path length or mean node-node distance, \( L \), of a network is defined as the number of edges in the shortest path between two nodes, averaged over all pairs of nodes. The longest node-node distance in the network defines its diameter, \( L_{\text{max}} \). The characteristic path length of the materials network is \( L = 1.8 \), and its diameter \( L_{\text{max}} = 2 \). This remarkably short path length indicates that the materials network has “small-world” characteristics, i.e. despite its large size, the number of edges that need to be traversed from a given node to any other node is relatively small. The extremely small \( L \) for the present materials network is not unexpected. First, noble gases, which only rarely participate in the formation of any materials, are connected with almost all the materials in the network. Since existence of even one such node is sufficient to put an upper bound of 2 on \( L_{\text{max}} \), the small-world character here is arising from the exclusive lack of reactivity of noble gases. Because some material pairs already have tie-lines that connect them immediately, the mean distance \( L \) is slightly smaller than 2. If the noble gases are disregarded, the network diameter \( L_{\text{max}} \) increases to 3 but the mean path length \( L \) remains small (1.9) due to the presence of a few other very-highly connected nodes, e.g. binary halides.

Another metric of interest in a real-world network is transitivity or clustering, quantified by its clustering coefficient, \( C \), which is the mean probability that two nodes connected to the same third node are themselves connected. In other words, if materials A and B each have a tie-line with material C, what is the probability that A and B have a tie-line between themselves? Depending on how the averaging is performed, a global clustering coefficient \( C_g \) or mean local clustering coefficient \( \bar{C}_i \) of a network can be defined [1, 17]. For the materials network, the clustering coefficients are \( C_g = 0.41 \) and \( \bar{C}_i = 0.55 \), comparable to some other real-world networks, and much higher than random networks of the same density [1]. Fig. 1 shows the mean local clustering coefficient for nodes of degree \( k \) as a function of \( k \). The decrease in \( \bar{C}_i \) with increasing degree indicates that stable materials form local, highly-connected communities in the network, and such behavior often suggests a hierarchical network structure [18, 19]. A related metric in a network is the extent of assortative mixing, i.e. tendency of nodes to connect to other “similar” nodes. We calculate the assortativity coefficient, the Pearson correlation coefficient of degree between pairs of connected nodes, for the materials network and find it to be around \( -0.13 \), indicating weakly dissortative mixing behavior. This is also confirmed by the distribution of the mean degree of neighbors of a node of degree \( k \) being a decreasing function of \( k \) (Fig. 1). In other words, materials with a high degree (large number of tie-lines) tend to connect with materials with lower degree (smaller number of tie-lines). The weakly dissortative behavior of the materials network is similar to that observed in most other technological, informational, biological networks, and is likely a virtue of such networks being simple graphs [20].

The degree distribution in the complete thermodynamic network of inorganic materials, specifically the probability \( p(k) \) that a compound has a tie-line with \( k \) other compounds in the network, follows a lognormal form (see Fig. 2). While the scale-free power-law distributions have been shown to exist in many widely-studied networks [3], lognormal distributions as another member of the “heavy-tailed” family, are also relatively common and behave quite similar to power-laws [21]. In fact, sparsity has been shown to be a necessary condition for the emergence of an exact power-law behavior [22], and densification in sparse, scale-free networks leads to distributions that deviate from power-law, and become closer to a lognormal [23]. Considering the extremely dense connectivity of the present network of materials when compared to the generally sparse scale-free networks studied commonly, a lognormal distribution is not unexpected.

With the clustering behavior and hierarchical structure revealed in Fig. 1, we now search for chemical fingerprints that would lead to such topological features in the network. Dependence of \( \langle k \rangle \) on the number of elemental components, \( N \) (\( N = 2 \) for binary, \( N = 3 \) for ternary, etc.) in Fig. 3 shows that higher-order materials tend to have fewer tie-lines with other materials, hinting at the existence of a chemical hierarchy in the network. Given that high-\( N \) compounds need to compete with low-\( N \) compounds in their chemical subspace to be stable (e.g. in a ternary system \( X-Y-Z \), ternary compounds compete with all other compounds in the three binaries \( X-Y, Y-Z \) and \( X-Z \)) whereas the stability of lower-order compounds are not affected by higher-order compounds, existence of a hierarchical structure to the network, in particular in its homochemical connections,
is intuitive. However, since the fraction of homochemical tie-lines increases with \( N \) (Fig. 3), the heterochemical tie-lines should contribute to the downwards trend in \( \langle k \rangle \) with \( N \). To understand the reason behind this connectivity behavior, we now turn to analyzing how \( N \)-ary compounds are distributed in the network.

Since the distribution of stable \( N \)-ary compounds decays fast after ternaries (Fig. 5), the 89-dimensional composition simplex is highly sparse, populated only at low dimensions, and underpopulated or unpopulated in high dimensions. In an \( N \)-dimensional compositional simplex (e.g. a ternary with \( N = 3 \)), if the interior of the simplex is empty, the nodes on faces at lower dimensions will form heterochemical tie-lines spanning the empty space. For example, in a ternary system \( X-Y-Z \), if the ternary space and binary spaces are empty except \( Y-Z \), the compounds in the \( Y-Z \) space will form heterochemical tie-lines with \( X \). When generalized to higher-dimensional spaces, lower-\( N \) materials have a higher chance of forming such heterochemical tie-lines than the higher-\( N \) materials because they simply have more facets in a simplex open to them for forming such heterochemical tie-lines, thereby giving rise to the overall trend in Fig. 3.

Next, we attempt to find potential reasons for why the distribution of stable \( N \)-ary materials in Fig. 5 peaks at a certain \( N \). In other words, does this distribution hint at the existence of an infinite, underexplored opportunity space for the discovery of systems beyond ternaries or a fundamental limitation on how high \( N \) can be? The distributions of formation energies of \( N \)-ary materials in the OQMD (Fig. 4) reflect the consequence of the competition between higher and lower dimensional materials. Materials of higher \( N \) appear to need significantly lower
FIG. 4. Distributions of formation energies of stable N-ary materials in the OQMD. The y-axes show the probability density function (pdf) and therefore the areas under the histograms are not proportional to the number of materials in each N-ary system.

formation energies than lower N materials, in particular beyond ternaries, to become stable. The distributions of all stable materials in the OQMD and those with an ICSD source are nearly the same (Fig. 5), with the most notable difference at $N = 3$ corresponding to ternary hypothetical materials that were found to be stable after HT prototype searches in the OQMD. Given that there is no obvious reason for the underlying distribution of formation energies of N-ary compounds to differ significantly for $N > 1$, (particularly at low temperatures, where entropic effects can be neglected), it is not surprising that only a few high-N materials would survive as stable materials, if the corresponding lower N systems already have several stable compounds. Widom [24] further argued that the distributions similar to that in Fig. 5, i.e. a peak near $N = 3$ or 4, arise from the competition between the exponential growth of number of combinations of elements and the diminishing volume-to-surface ratio in the composition simplex, as N increases. At this point, we do not know if there exists a fundamental law limiting access to higher N materials. However, combined with the hierarchy observed in the formation energies, these arguments imply that the scarcity of known high-N stable materials may simply be originating from the topology of the convex-hull itself, rather than a mere consequence of those spaces being underexplored.

Since the “phase diagram of everything” practically encompasses all known inorganic crystalline materials as well as a large number of hypothetical materials, the number of tie-lines (or node degree) emerges as a natural metric of nobility of a crystalline material—it is simply the count of other materials it is determined to have no reactivity against. The nobility here is therefore the measured non-reactiveness of a material against all others in the database. Thus, while reactivity or nobility have no concrete definitions, we are able to approach chemical nobility of inorganic materials in solid-solid and solid-gas reactions from a data-driven stand point, instead of bottom-up theory, intuition or heuristic approaches [25]. Since the number of tie-lines follow a lognormal distribution, a natural standard score for nobility can be defined as

$$z_n = \frac{\ln(k) - \mu}{\sigma}$$  \hspace{1cm} (1)\n
where $k$ denotes the degree, i.e. number of tie-lines of the material, and $\mu$ and $\sigma$ are the mean and standard deviation of the distribution. The present values of $\mu$ and $\sigma$ for OQMD as extracted from the fitted lognormal distribution are 8.06 and 0.65, respectively. The metric is agnostic of textbook classifications such as metal, non-metal, metalloid, ionic, covalent and so on, and would work equally for all such classes. The accuracy in predicted nobilities is bounded of course by that of DFT in predicting relative stabilities of inorganic materials [26].

The first question we are after is what tie-lines tell us about the reactivity or nobility of elements. Noble gases and F, not too surprisingly, form the bounds of the nobility score in Fig. 6, as the noblest and most reactive, respectively, not only among the elements but in fact among all materials in the phase diagram. Following F, most reactive elements are P, S and Cl. Alkali and alkaline earth elements, which are often considered as highly reactive metals, are found to be relatively noble in solid-solid and solid-gas reactions compared to early d-block or lanthanide elements, which are, along with Al, the most reactive metals. The most reactive d-block metals are Y and Sc, and the most reactive alkali and alkaline earth elements are Li and Be respectively. The nobility metric tends to increase down a group for metals, and increases
solid-state reactions. Further, we find that the noblest element is EuF$_2$ forming tie-lines with almost 99% of all stable materials, immediately following the noble gases. This material was known to be stable against oxidation and hydrolysis [28], but this extent of solid-solid inertness is likely unknown to date. EuF$_2$ is an example among many IA/IIA–VIIA compounds such as RbYbF$_3$, RbBr, CsBr, YbClF, CsYbF$_3$, Rbi, KCl, CsCl, KBr, SrF$_2$, KI, CaF$_2$, NaCl, LiF, and others, that have a highly ionic character. ThO$_2$, YbO, BeO, Lu$_2$O$_3$, and Tm$_2$O$_3$ are found to be the top 5 most noble metal-oxides on the basis of the number tie-lines they form.

While some of the findings above are in line with chemical intuition, relative nobilities in certain cases, e.g. silver vs gold, deviate from it. Part of that deviation in Fig. 6 is originating from the context these elements are historically considered as noble or reactive for, such as whether an element oxidizes or corrodes readily when exposed to air, reacts with water or certain acids, tends dissolve in water or electrolytes (e.g. in the electrochemical series) and sometimes considering how vigorous such reactions would seem. More fundamental approaches to finding descriptors for reactivity go back to electronegativity related concepts, followed by many interrelated theories based on perturbation theory, derivatives of electronic energy (such as hardness and softness) and others [29–33]. The $z_n$ score, as derived from the tie-lines in the present network provides a general, solid-state reaction-driven bulk thermodynamic metric, emerging directly from the data, and hence complementary to chemical intuition based on chemical-agent or reaction-vigor, or metrics derived from the electronic structure.

**CONCLUSION**

To summarize, we compute the complete phase diagram of inorganic materials, encompassing all known ordered compounds and more, based on the formation energies derived from high-throughput density functional theory. Analyzing the topology of this phase diagram in the context of complex networks, we show that it is an extremely dense, small-world network with very small diameter and mean-shortest-path, and a high clustering coefficient. We show that the topology of the network has a hierarchy to it, with the hierarchical structure strongly associated with the number of components in materials. We further demonstrate how topological features of this network can uncover material behavior determined by their collective interactions, such as quantifying relative nobilities of materials. We believe that the present analysis based on considering the “phase diagram of all inorganic materials” as a complex network simply serves as a preliminary demonstration of opportunities in understanding and extracting knowledge out of such materials networks. For example, more detailed analyses of the topology, such as identification of communities/groups in the network and how much those topological features overlap with chemical intuition can provide new insights. As another example, while some of its features resemble other complex networks, the extremely-high connectance and the lognormal degree distribution of the presented thermodynamic network imply that its underlying generative mechanisms may be unique, and developing generative models for such materials networks can have significant impact on the knowledge discovery for materials in the future.

**METHODS**

**Network analysis:**
Properties of the networks were analyzed using networkx and graph-tool [34] packages. Degree distributions were analyzed using the powerlaw pack-
age [35]. Computations of convex-hulls were performed using the Qhull library (http://www.qhull.org), as implemented in qmpy (https://github.com/wolverton-research-group/qmpy).

**Computation of the 89-dimensional composition-formation energy convex-hull:** While representation complexity of the convex-hull increases exponentially [36], we know from the existing set of 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 (Fig. 5). 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 spaces of such candidate pairs, rather than the 89-dimensional space itself. We then represent stable compounds as nodes and tie-lines as edges, thereby generating the “universal phase diagram” as a complete thermodynamic network of all inorganic materials.

**ACKNOWLEDGEMENTS**

V.I.H was supported by the Accelerated Materials Design and Discovery program at Toyota Research Institute.

[1] M. E. J. Newman, SIAM Review 45, 167 (2003).
[2] H. Ebel, L.-J. Mielsch, and S. Bornholdt, Physical Review E 66 (2002), 10.1103/PhysRevE.66.035103.
[3] A.-L. Barabási, Science 325, 412 (2009).
[4] R. Guimerá and L. A. Nunes Amaral, Nature 433, 895 (2005).
[5] M. F. Ashby, *Materials selection in mechanical design* (Butterworth-Heinemann, 2011).
[6] M. Aykol, V. I. Hegde, S. Suram, L. Hung, P. Herring, C. Wolverton, and J. S. Hummels, ArXiv e-prints (2018), arXiv:1806.05772 [cond-mat.mtrl-sci].
[7] O. Isayev, D. Fournes, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, and S. Curtarolo, Chemistry of Materials 27, 735 (2015), arXiv:arXiv:1412.4966v3.
[8] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, Nat. Mater. 12, 191 (2013).
[9] 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).
[10] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. Persson, APL Mater. 1, 011002 (2013).
[11] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, JOM 65, 1501 (2013).
[12] J. Hachmann, R. Olivares-Amaya, S. Atahan-Envenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, and A. Aspru-Guzik, Journal of Physical Chemistry Letters 2, 2241 (2011).
[13] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, npj Computational Materials 1, 15010 (2015).
[14] A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch, Acta Crystallographica Section B Structural Science 58, 364 (2002).
[15] S. P. Ong, Y. Mo, W. D. Richards, L. Miara, H. S. Lee, and G. Ceder, Energy Environ. Sci. 6, 148 (2013).
[16] M. Aykol, S. Kim, V. I. Hegde, D. Snyderacker, Z. Lu, S. Hao, S. Kirklin, D. Morgan, and C. Wolverton, Nature Communications 7, 13779 (2016).
[17] D. J. Watts and S. H. Strogatz, Nature 393, 440 (1998).
[18] E. Ravasz and A.-L. Barabási, Phys. Rev. E 67, 026112 (2003).
[19] G. Szabó, M. Alava, and J. Kertész, Phys. Rev. E 67, 056102 (2003).
[20] S. Maslov, K. Sneppen, and A. Zaliznyak, Physica A: Statistical Mechanics and its Applications 333, 529 (2004).
[21] M. Mitzenmacher, Internet Mathematics 1, 226 (2003).
[22] C. I. Del Genio, T. Gross, and K. E. Bassler, Physical Review Letters 107, 178701 (2011).
[23] M. Scholz, Journal of Data Mining and Digital Humanities 2015, 1 (2015), arXiv:1010.0803v5.
[24] M. Widom, Journal of Statistical Physics 167, 726 (2017), arXiv:1610.02054.
[25] B. Hammer and J. K. Norskov, Nature 376, 238 (1995).
[26] M. Aykol, S. S. Dwaraknath, W. Sun, and K. A. Persson, Science Advances 4, eaau0148 (2018).
[27] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, npj Computational Materials 2, 16028 (2016), arXiv:1606.09551.
[28] K. Lee, H. Muir, and E. Catalanlo, Journal of Physics and Chemistry of Solids 26, 523 (1965).
[29] W. Yang and R. G. Parr, Proceedings of the National Academy of Sciences 82, 6723 (1985).
[30] H. Chermette, Journal of Computational Chemistry 20, 129 (1999).
[31] G. Klopman, Journal of the American Chemical Society 90, 223 (1968).
[32] R. G. Pearon, Inorganic Chemistry 27, 734 (1988), arXiv:00277a030 [10.1021].
[33] T. P. Peixoto, figshare (2014), 10.6084/m9.figshare.1164194.
[34] J. Alstott, E. Bullmore, and D. Plenz, PLoS ONE 9, e85777 (2014).
[35] H. Sartipizadeh and T. L. Vincent, Arxiv (2016), arXiv:1603.04422.