A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis

Matthew J. McDermott
Lawrence Berkeley National Laboratory, and University of California, Berkeley
Shyam S. Dwaraknath
University of California, Berkeley
Kristin A. Persson (kapersson@lbl.gov)
Lawrence Berkeley National Laboratory, and University of California, Berkeley

Article

Keywords: chemical reaction, network model, thermochemistry, thermodynamic phase space, pathfinding algorithm, yttrium manganese oxide YMnO3, solid-state chemistry

Posted Date: June 30th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-38000/v1

Version of Record: A version of this preprint was published at Nature Communications on May 25th, 2021. See the published version at https://doi.org/10.1038/s41467-021-23339-x.

A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis

Matthew J. McDermott[1,2], Shyam S. Dwaraknath[1], and Kristin A. Persson[1,2,*]

[1]Energy Technologies Area, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
[2]Department of Materials Science and Engineering, University of California, Berkeley, CA 94720, USA
[*]Corresponding author: Kristin Persson (email: kapersson@lbl.gov)

June 24, 2020

Abstract

Accelerated synthesis of inorganic materials remains a significant challenge in the search for novel, functional materials. Many of the chemical principles which enable "synthesis by design" in synthetic organic chemistry do not exist in solid-state chemistry, despite extensive computed/experimental thermochemistry data. We present a chemical reaction network model constructed from thermochemistry databases that captures features of the thermodynamic phase space which synthesis reactions traverse. Directed edges in the network are assigned weights via a transformation that maps reaction parameters to costs. We devise a computationally tractable approach for suggesting likely reaction pathways via application of pathfinding algorithms and linear combination of lowest-cost paths in the network. We demonstrate initial success of the reaction network in predicting a complex metathesis reaction pathway toward yttrium manganese oxide ($YMnO_3$). The reaction network presents new opportunities for enabling reaction pathway prediction, rapid iteration between experimental/theoretical results, and ultimately, control of synthesis of solid-state materials.

Introduction

Dating back to 18th century mineralogy,[1] solid-state inorganic chemistry is a cornerstone in the design of novel, functional materials and continues to be driven by pressing technological demands. Consequently, the development of new techniques that accelerate materials synthesis/processing is vital for achieving multifunctional materials with complex properties that satisfy today's technological needs. Solid materials with target functionality are often thermodynamically metastable, which can limit their accessibility via conventional solid-state synthesis routes such as the classic "shake and bake" ceramic methods that typically require high temperatures to overcome diffusion barriers and often proceed to global thermodynamic equilibrium.[2] Indeed, solid-state chemistry itself has been dubbed a "black box" which is most effectively probed via systematic and extensive iteration, requiring significant experimental expertise akin to apprenticed artistry.[3] The optimization of synthesis procedures for new materials is hence both highly time- and resource-consuming, demanding human-guided iteration over many combinations of precursors, processing steps, and environmental conditions.

A more efficient approach to synthesizing novel inorganic materials is "synthesis by design", in which a set of guiding principles and relationships is used to quickly devise a synthesis method towards a target material, much like the paradigm central to synthetic organic chemistry.[4,5] Recent work, fueled by advances in solid-state in situ characterization techniques,[6,7] has taken steps in this direction by exploring reaction pathways in select case systems and identifying mechanistic relationships that explain how synthesis conditions (e.g. precursor selection and reaction environment) alter the reaction pathway, leading to selective formation of different target products. For example, Neilson and coworkers demonstrated the use of unconventional solid-state metathesis reactions to kinetically control the reaction pathway towards metastable polymorphs of $CuSe_2$[8] and $YMnO_3$.[9,10] Jiang et al. explored the use of iron silicide reactants to bypass kinetic limitations and achieve low-temperature synthesis of Fe_2SiS_4.[11] Miura et al. demonstrated the synthesis of $MgCr_2S_4$ thiospinel via a metathesis route using novel precursors, which was shown to be thermodynamically favorable through computational phase diagram construction.[12] Bianchini et al. showed that the first phase formed in the synthesis of P2 type $Na_{0.67}MO_2$ (M=Co,Mn) can be predicted by minimizing compositionally unconstrained reaction energies, and that formation of the initial phase can drastically alter the kinetics of the subsequent reaction and final phase selectivity.[13] Each of these studies elucidates an important concept: chemical reaction pathways follow a complex thermodynamic free energy surface which can be carefully manipulated and navigated via thoughtful selection of precursors, processing, and environmental conditions.

Explicit modelling of the free energy surface at an atomistic level (i.e. the potential energy surface of atomic interactions) has been successful in predicting chemical reaction pathways/dynamics in molecules.[14] However, in solid-state chemistry reactions, monitoring the time dependence of each atom's spatial coordinates and interactions over the much larger scale ($\sim 10^{23}$ atoms per mole) becomes intractable. Despite these limitations, modelling of bounded solid-state reaction mechanisms at the atomistic level has been achieved in particular with molecular dynamics (MD)[15] and kinetic Monte Carlo (KMC)-based[16] approaches. Reactive force fields, such as ReaxFF[17] further permit the breaking of chemical bonds and can be used to study specific chemical reaction mechanisms and kinetic parameters.[18] KMC-based methods also explore parts of the potential energy surface, given reaction rate constants that can be approximated with quantum mechanical calculations. However, such methods are ultimately confined to an a priori selection of the relevant domains of the high

dimensional solid-state potential energy surface.

To aid in the development of materials synthesis by design, we propose to leverage recent advances in data-driven methods which have resulted in computational/experimental thermochemical databases[19–22] covering hundreds of thousands of materials and millions of associated reaction energies.[23] In the remainder of this work, we describe a framework for predicting and suggesting solid-state inorganic reaction pathways, which combined with experimental efforts, will ultimately realize inorganic synthesis by design. We propose a chemical reaction network which blends typical thermodynamic phase diagrams with the connectivity and kinetic heuristics derived from transition state theory. The network model serves as a convenient data structure for exploring the underlying free energy surface of thermodynamic phase space in solid-state chemistry via the power and efficiency of existing computational infrastructure for large graph networks. We outline the methodology used to create the chemical reaction network from thermochemical databases and demonstrate its predictive power as a generator of probable reaction pathways, using the recent metathesis synthesis of the multiferroic $YMnO_3$ as a demonstrative case study.

Results

The following subsections describe the i) construction of the chemical reaction network and its relationship to previous models of thermodynamic phase space, ii) prediction of reaction pathways using pathfinding methods, and iii) demonstration in predicting reaction pathways in the experimental metathesis synthesis of $YMnO_3$.

A weighted directed graph of chemical reactions

Following the thermodynamic equilibrium approach developed by Gibbs,[24] we consider solid-state chemical reactions as traversing a thermodynamic phase space governed by a generalized thermodynamic potential or free energy, Φ, where the global minimum represents thermodynamic equilibrium for the system. Figure 1 depicts several models of chemical reactions in thermodynamic phase space, ordered by increasing level of abstraction. The free energy convex hull construction of Figure 1(a) is a purely thermodynamic model of a chemical reaction between a pair of two reactant phases, R_1 and R_2. The convex hull yields the chemical reactions which result in the largest decrease in free energy for a given mole ratio of the two reactants. Figure 1(b) abstracts the thermodynamic model further by incorporating the concept of activation energy, E_a, as defined in transition state theory.[25] This enables inclusion of simple kinetic behavior of reactions, where the height of the activation energy barrier correlates with the rate of reaction. Abstracting further, we can consider these reaction coordinate diagrams as weighted directed graphs, as in the upper portion of Figure 1(b). In these graphs, the cost (or weight) of a chemical reaction edge represents an a priori unknown function of synthesis parameters such as the thermodynamic driving force, activation energy, etc. Figure 1(c) shows the interlinking of many such graph representations within a set of phases, where the nodes represent a combination of phases (e.g. $R_1 + R_2$) and the edges represent chemical reactions with a designated cost. This weighted directed graph, or chemical reaction network, is a densely connected model of thermodynamic phase space where thermodynamic/kinetic features can be combined and transformed into a unique cost representation for each reaction.

Figure 2 illustrates the generalized graph structure of a reaction network for any chemical system. Here, "chemical system" refers to the set of all N phases p_i $(i = 1, 2, \ldots N)$, which can be produced from a designated set of chemical elements. Each reactant/product node on the graph is created by considering combinations of distinct phases up to a maximum size, n. This corresponds to the set of all nodes, P, given by:

$$\begin{aligned} P = \; & \{p_i | i \leq N\} \\ & \cup \{p_i + p_j | i, j \leq N; i \neq j\} \; \cup \; \ldots \\ & \cup \{p_i + p_j + \cdots + p_n | i, j, \ldots n \leq N; \\ & i \neq j \neq \ldots \neq n\} \end{aligned} \tag{1}$$

Each of these phase combinations is added twice: once as reactants node and again as a products node. While higher values of n enable more complex reactions, in general it suffices to choose $n = 2$, since truly simultaneous reactions among three or more reactants are less likely due to kinetic and steric constraints in a solid composite.

To create the dense set of directed edges at the center of the network, we algorithmically iterate through every possible chemical reaction between all pairs of reactants and product nodes. Using a reaction balancing algorithm, we then solve for the stoichiometric coefficients and add a weighted, directed edge from the reactant node to corresponding product node for every chemical reaction which is successfully balanced. Note that many of the generated trial reactions cannot be stoichiometrically balanced and hence

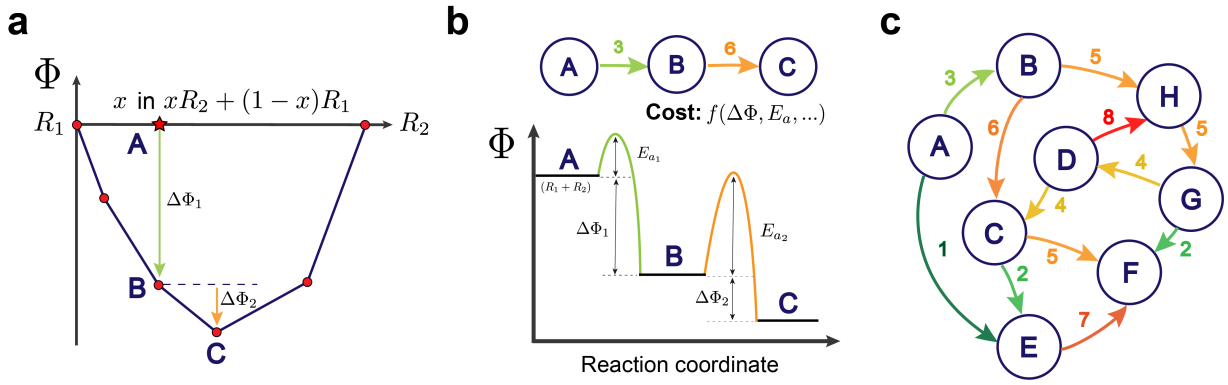

Figure 1: **Models of chemical reactions in thermodynamic phase space, ordered by increasing degree of abstraction**. The label, A, indicates a combination of arbitrary amounts of phases R_1 and R_2. Similarly, the other labels (B-F) are combinations of arbitrary amounts of other phases which result from the chemical reactions. (a) The convex hull construction for consecutive reactions between two reactants, R_1 and R_2. The points drawn indicate chemical reactions at different stoichiometric mixtures of the reactants, and the lines trace the convex hull indicating the thermodynamic equilibrium (minimum free energy) for all ratios of mixing, x. (b) A traditional reaction coordinate diagram used to represent both the free energies of reaction, $\Delta\Phi$, and activation energies, E_a. This is generalized by a weighted, directed graph connecting the three states (top). The cost/weight of the directed edges, shown as the colored number adjacent to each edge, is some function of the free energy change, activation energy, and other reaction features. (c) A chemical reaction network linking together many such possible reaction pathways that may occur within a set of phases.

are excluded from the graph, e.g. there are no x, y, z that satisfy $x\mathrm{Y_2O_3} + y\mathrm{MnO_2} \longrightarrow z\mathrm{YMnO_3}$, but the reaction is balanceable if $\mathrm{O_2}$ is included as an additional product. We also exclude trivial "identity-like" reactions between identical reactants and products, e.g. $\mathrm{Y_2O_3} \longrightarrow \mathrm{Y_2O_3}$. The weight of the reaction edge is determined by a "cost function" that maps features of the chemical reaction (e.g. $\Delta\Phi_{rxn}$) to a single cost value. To facilitate product phases being capable of reacting again (e.g. autocatalytic reactions), zero-weight edges are added which connect each product node to all reactant nodes that contain, as a subset, at least one of the product phases and/or starting reactant phases (regardless of consideration of stoichiometric coefficients). This creates a large degree of cycles in the network, enabling the network to capture multiple step reaction pathways.

Finally, two more nodes are added: one for the synthesis precursors and one for the selected target. These two external nodes act as single-source and destination nodes, defining a net (overall) synthesis reaction. The precursors node connects into the network via zero-weight edges directed towards all reactants nodes that contain, as a subset, at least one of the precursor phases. To allow for an open system, or excess with respect to precursors or specific reactants, zero-weight edges are optionally added from each from each product node to all reactant nodes which include the excess precursor phases. This extra layer of connectivity enables the precursors to react at different steps along the reaction pathway. Lastly, the network links into the target node via a set of zero-weight edges directed from all product nodes which contain the target phase.

The cost function used to determine the weighting of edges is critical to its performance in generating probable reaction pathways. In this work, we employ Dijkstra's algorithm,[26] which uses a priority queue structure to determine the shortest path from a single-source node to destination node. The shortest path is defined as the path which has the smallest sum of all its edge weights. The simplest, and possibly most intuitive, cost function is a direct mapping onto the thermodynamic landscape, such as the measured or calculated Gibbs free energy of reaction, ΔG_{rxn}. However, using reaction energies alone poses several problems: 1) negative reaction energies result in infinite cycles during pathfinding which preclude the use of Dijkstra's algorithm and many other pathfinding methods, 2) kinetic effects and other known heuristics about the reaction are excluded, and 3) reaction costs are affected by stoichiometric scaling. Instead, here we choose a single, positive cost function that maps the Gibbs free energy of reaction, per reactant atom,

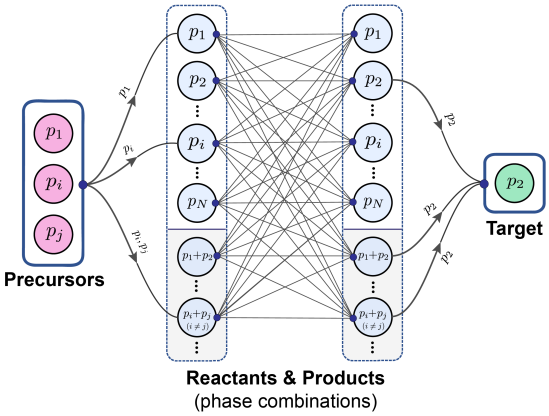

Figure 2: **The generalized graph architecture of a solid-state chemical reaction network.** The reaction network is constructed for a hypothetical chemical system containing N phases p_i ($i = 1, 2, \ldots N$), with nodes made of distinct phase combinations up to maximum size n. The precursors and target nodes link into and out of the densely connected network of reactions, respectively. The edge directions between the reactant and product nodes have been omitted for clarity. These central edges include chemical reactions weighted by their cost value, as well as zero-weight edges which create loops from product nodes back to reactant nodes, such that multiple step reaction pathways can be captured.

to a positive value for each reaction. The choice of a functional mapping also provides the opportunity to create different cost functions for chemical reactions where additional information is known, such as experimental data, kinetic factors and/or other heuristics. One example of a cost function that captures reaction thermodynamics is the *softplus* function, which was originally developed for use as an activation function in neural networks.[27] This function maps the Gibbs free energy of a chemical reaction, ΔG_{rxn}, to a positive cost value, C, via:

$$C = \ln(1 + \frac{273\ \mathrm{K}}{T} e^{\Delta G_{rxn}}) \qquad (2)$$

where T is absolute temperature in Kelvin and ΔG_{rxn} is the Gibbs free energy of reaction in units of eV per reactant atom, divided by unity to be dimensionless. Since molar reaction energies scale with the stoichiometric balancing of the reaction, ΔG_{rxn} must be normalized on a per-atom basis independent of the stoichiometric coefficients. Figure 3 demonstrates the effect of the cost transformation on the distribution of reaction free energies in a chemical system. The softplus cost function transforms highly exergonic ($\Delta G_{rxn} << 0$) reactions into low (near zero) cost events, whereas endergonic ($\Delta G_{rxn} > 0$) reactions exhibit a finite cost that smoothly approaches a linear scaling as $\Delta G_{rxn} \to \infty$. Note that different environmental boundary conditions, such as open elements, can be modelled by replacing ΔG_{rxn} with $\Delta \Phi_{rxn}$, where Φ represents a customized thermodynamic potential. Other monotonically increasing functions were tested (Supplementary Figure 1) however the smooth and differentiable softplus function was preferred. We hypothesize that there indeed exists a *true* cost function for each chemical reaction, which depends on local and global environmental conditions and transforms the thermodynamic/kinetic factors to an appropriate cost that allows the network to accurately capture the real behavior of the system.

Predicting reaction pathways via graph pathfinding

Ideally, predicting a reaction pathway using the network would be equivalent to leveraging existing, powerful algorithms for solving the single-source shortest path problem from graph theory. However, chemical reactions rarely trace a set of linear steps, even in simple syntheses; instead the precursor phases often undergo reactions concurrently in parallel, or react again via autocatalytic reactions. Within the network, parallel reactions can be modeled as simultaneous travel along multiple reaction edges. In the physical world these reactions must obey mass conservation, and phases produced in one reaction may react with phases in another. These so-called "crossover" reactions, along with the possibility of parallel reactions, prohibit the *direct* application of shortest path algorithms.

To solve this issue of parallel paths and crossover reactions, we identify not only the single shortest path from precursors to target phase, but all k-shortest paths in the network, and then attempt to find linear combinations of and interactions between paths which satisfy the stoichiometric constraints of a net/overall reaction which is known a priori. For syntheses that involve multiple targets or byproducts (e.g. CO_2), pathfinding is performed towards each target phase separately, and then all paths are accumulated together; this ensures that generated paths access all targets. With the set of likely paths, we identify intermediate phases and supplement the set with possible crossover reactions which result in further production of the target phase(s). To computationally generate all k-shortest paths, we utilize Yen's algorithm[28] which iteratively produces the next $k-1$

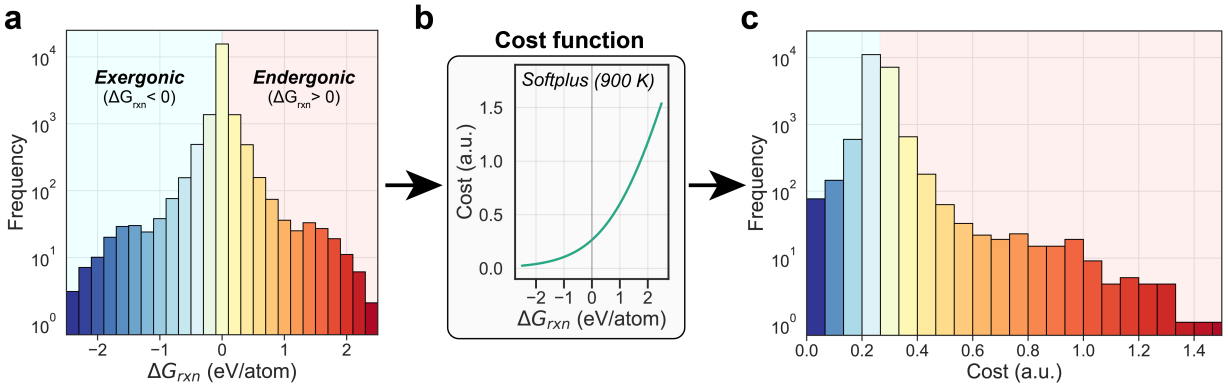

Figure 3: **The effect of the cost function transformation on reaction energies**. (a) Distribution of reaction Gibbs free energies (normalized per reactant atom) in the C-Cl-Li-Mn-O-Y chemical system. (b) Transformation of reaction energies via the softplus cost function described in equation (2). (c) Final distribution of reaction costs (in arbitrary units) after transformation. A reaction free energy of zero corresponds to a cost of ~ 0.265.

shortest paths via deviations from the first shortest path, as calculated with Dijkstra's algorithm. Given that the cost function transformation provides only an approximation to the thermodynamics and kinetics of evolving reactions, the identification of *many* low-cost paths is also practical in the creation of a candidate set of possible paths. While the top shortest paths may not be experimentally feasible, the use of Yen's algorithm to generate many such low-cost paths narrows down the range of probable reactions and facilitates tractable exploration of a combinatorially-dense phase space.

Demonstration of reaction network model for synthesis of $YMnO_3$

To demonstrate the capabilities of the reaction network for predicting possible reaction pathways in a synthesis procedure, we consider the synthesis of yttrium manganese oxide, $YMnO_3$, through the solid-state assisted metathesis reaction reported by Todd & Neilson.[9] The overall reaction,

$$\mathrm{Mn_2O_3} + 2\,\mathrm{YCl_3} + 3\,\mathrm{Li_2CO_3} \longrightarrow 2\,\mathrm{YMnO_3} + 6\,\mathrm{LiCl} + 3\,\mathrm{CO_2} \quad (3)$$

was found to exhibit several steps with distinct intermediate compounds, as determined through in situ temperature-dependent x-ray diffraction performed at a synchrotron beamline.[10] Figure 4(a) shows the chemical reaction network generated for the C-Cl-Li-Mn-O-Y chemical system. The C-Cl-Li-Mn-O-Y phase diagram constructed from Materials Project contains 768 entries; of these, 47 are predicted to be stable at $T = 0$ K. The machine-learned Gibbs descriptor reported by Bartel et al.[29] and NIST-JANAF tables[21] were further used to transform the DFT-derived formation enthalpies into Gibbs free energies of formation at 900 K – the approximate temperature at which $YMnO_3$ was observed to form. The transformation to Gibbs free energy at 900 K reduces the number of stable phases in the phase diagram to 39. We include all of these stable entries as well as metastable entries (not including polymorphs) up to a filter of +20 meV/atom above the hull, resulting in a total of 56 phases considered (Supplementary Table 1). The final reaction network contains 3,194 nodes and 33,428 edges, where 20,116 of these edges represent chemical reactions with a maximum phase combination size of $n = 2$. Costs for all reaction edges were mapped using the softplus function described in equation (2) with $T = 900$ K and plotted in Figure 3b.

Reaction pathway prediction was performed given the initial reactants and final products of the assisted metathesis reaction in equation (3). The 60 shortest paths (20 to each product of $YMnO_3$, LiCl, and CO_2) were identified via Yen's algorithm, resulting in a set of 47 unique reactions. All 60 shortest paths can be seen in Supplementary Tables 2-4. Combined pathways were generated via this set of reactions by solving for linear combinations of reactions, up to a maximum size of four reactions, that obey stoichiometric constraints. Of the total 195,708 pathways considered, only *two* combined reaction pathways fit the stoichiometric constraints of the net reaction.

The shorter of the two combined pathways (in both total and average cost per reaction) involves the for-

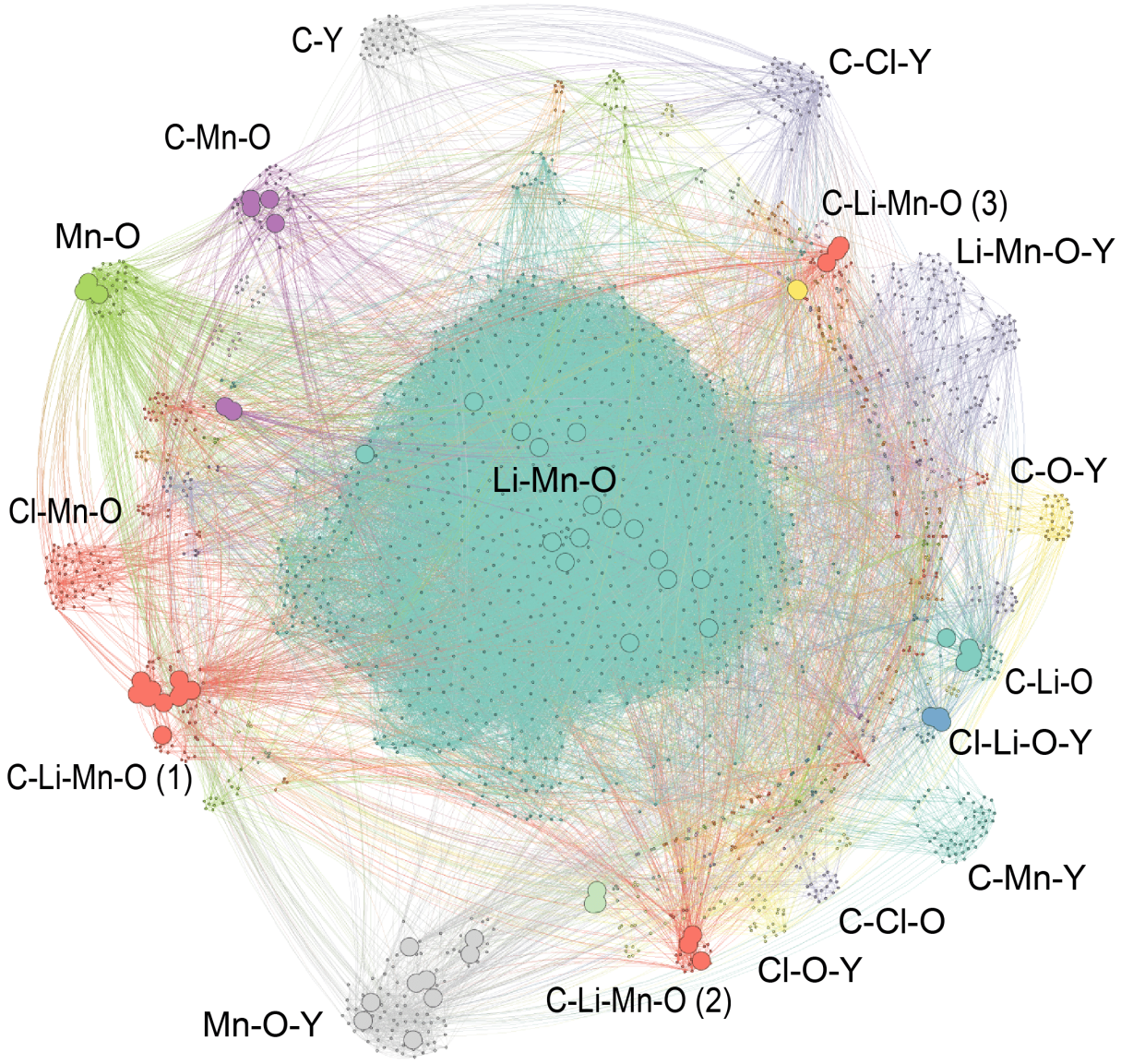

Figure 4: **The reaction network for the C-Cl-Li-Mn-O-Y chemical system**. The network contains 56 phases: 39 stable and 17 within +20 meV/atom above the hull. Chemical subsystems are labeled by color. The larger nodes indicate reactant nodes which are traversed on the 20 shortest pathways from precursors to targets in the $YMnO_3$ assisted metathesis reaction given by equation (3).

mation and reaction of Li_3MnO_3 which, to our knowledge, is an oxygen-deficit distorted rocksalt that is yet to be experimentally verified; in fact, the Materials Project contains tens of thousands of hypothetical compounds, which may or may not be readily synthesizable. According to available computed data, this compound is metastable (+12 meV/atom above the hull) at zero temperature, but is predicted to exhibit thermodynamic stability at 900 K according to the Gibbs free energy model. The proposed pathway is reported below, along with the computed free energy of reaction per reactant atom:

$$1.5Li_2CO_3 + 0.5Mn_2O_3 \longrightarrow Li_3MnO_3 + 1.5CO_2 \quad (\Delta G_{rxn} = 0.199\ \text{eV/atom}) \tag{4}$$

$$\mathrm{YCl_3 + Li_3MnO_3 \longrightarrow YMnO_3 + 3LiCl}$$
$$(\Delta G_{rxn} = -0.225\ \mathrm{eV/atom}) \quad (5)$$

The second ranked pathway by cost, however, closely matches the experimentally reported assisted metathesis pathway. The Li_2CO_3 first decomposes to Li_2O and then reacts to make YOCl:

$$\mathrm{Li_2CO_3 \longrightarrow Li_2O + CO_2}$$
$$(\Delta G_{rxn} = 0.312\ \mathrm{eV/atom}) \quad (6)$$

$$\mathrm{YCl_3 + Li_2O \longrightarrow YOCl + 2\,LiCl}$$
$$(\Delta G_{rxn} = -0.287\ \mathrm{eV/atom}) \quad (7)$$

while simultaneously, $LiMnO_2$ is formed:

$$\mathrm{0.5\,Li_2CO_3 + 0.5\,Mn_2O_3 \longrightarrow LiMnO_2 + 0.5CO_2}$$
$$(\Delta G_{rxn} = 0.099\ \mathrm{eV/atom}) \quad (8)$$

And these two intermediates finally react together in a metathesis reaction to produce $YMnO_3$:

$$\mathrm{YOCl + LiMnO_2 \longrightarrow YMnO_3 + LiCl}$$
$$(\Delta G_{rxn} = -0.086\ \mathrm{eV/atom}) \quad (9)$$

This pathway is nearly identical to the experimentally reported pathway; with the added step that Li_2CO_3 first decomposes to Li_2O and CO_2 before reacting.

Discussion

The primary challenge in creating a reaction network model is the high degree of complexity inherent to thermodynamic phase space, which quickly leads to a combinatorial explosion during both the creation of the network and subsequent pathfinding steps. As an example, consider a reaction network with N phases and a maximum phase combination size, n. If during the graph generation every possible chemical reaction between any two nodes is considered, the number of reactions, R, would be:

$$R = \left[\sum_{i=1}^{n} \binom{N}{i}\right]^2 = \left[\binom{N}{1} + \binom{N}{2} + \ \dots\ + \binom{N}{n}\right]^2 \quad (10)$$

For example, for the C-Cl-Li-Mn-O-Y reaction network with $N = 56$ distinct phases, the maximum number of reactions described by equation (10), before stoichiometric balancing, is $R \approx 3.14 \times 10^3$ ($n = 1$), 2.55×10^6 ($n = 2$), and 8.59×10^8 ($n = 3$).

In this work, we present a computationally tractable approach, which effectively introduces a series of filters which reduce the complexity and degrees of freedom of the thermodynamic phase space. These filters include: 1) restricting the number of phases considered via thermodynamic stability arguments, i.e. energy above the hull, 2) limiting the maximum number of phases present on each side of the reaction to a small number, e.g. $n = 2$, 3) using a cost function to prioritize reactions which are likely to occur, and 4) enforcing mass conservation via stoichiometric constraints. The first two filters work together during graph generation to limit the combinatorial size/complexity of the network. This number can be reduced by decreasing either N, n, or both. It is also worth noting that the number of considered reactions can be reduced by considering the connectivity of the compositional phase diagram of the system; for example, chemical reactions may be limited to only those reactions which occur along facets of the phase diagram. Since it is typically optimal to consider as many phases as possible, it is more favorable to reduce n, rather than N. Therefore the choice of $n = 2$ minimizes the combinatorics of the network without inherently sacrificing the complexity of reactions that can occur. Indeed, reaction pathways suggested by the network adhering to the $n = 2$ limit must consist of pseudo-elementary steps which more closely follow the free energy surface. This behavior is nicely illustrated in the second reaction pathway suggested by the C-Cl-Li-Mn-O-Y network (equations (6)-(9)), which closely resembles the experimentally observed pathway. In the experimental synthesis of $YMnO_3$, however, the authors report a reaction step with three products ($\mathrm{YCl_3 + Li_2CO_3 \longrightarrow YOCl + 2\,LiCl + CO_2}$). Due to the selection of $n = 2$, this reaction was not explicitly present in the network. However, the *net* effect of the reaction is indeed incorporated in the second suggested pathway, where this reaction is essentially divided into two smaller steps featuring the decomposition of Li_2CO_3 into Li_2O/CO_2 and the reaction of Li_2O with YCl_3 directly. It is reasonable to postulate that this thermal decomposition step might actually be occurring; the decomposition of lithium carbonate has been well-studied and observed to occur spontaneously at/above temperatures near the maximum temperature of the assisted metathesis synthesis route ($T \sim 900$ K).[30]

The cost function approach provides another simplification of the complex thermodynamic phase space. Here we have shown that a smooth, monotonic transformation of the Gibbs free energy of reaction is sufficient itself in capturing realistic behavior in solid-state materials synthesis, as exemplified

by the network model of $YMnO_3$ synthesis. The cost transformation to positive values allows for the utilization of existing shortest path algorithms such as Dijkstra's algorithm, but also naturally incorporates several realistic/expected features of traversing the thermodynamic phase space. First, the cost function can in principle integrate multiple thermodynamic and kinetic heuristics. While we did not include any kinetic features in this work, we anticipate the future addition of kinetic features, such as the structural (dis)similarity between phases,[31,32] the average number of bonds broken/created, change in the information entropy description of atomic configurations, change in atomic density, etc. The relative weights of each of these features within the cost function, of course, must be carefully examined and validated. For example, the fact that the reaction network encodes modular pseudo-elementary steps is highly conducive to the inclusion of modeled or experimentally obtained kinetic barriers, where the cost associated with a particular step may be high enough to remove the entire pathway from consideration.

Second, shortest path algorithms are naturally cost-additive, which biases the pathway generation towards simpler reaction pathways with fewer steps. This also introduces a trade-off between the number of steps in the pathway and the cost per step, i.e. paths with several low-cost steps may exhibit the same total cost as paths with only one medium/high cost step. The slope of the cost function determines the nature of this trade-off; however, the *softplus* function generally is found to favor shortest paths with fewer steps (Supplementary Tables 2-4). One unique result of the additive nature of the shortest path approach is that the shortest paths towards certain products often involve unanticipated, endergonic ($\Delta G_{rxn} > 0$) reaction steps. This is a major advantage as compared to other thermodynamic models where reaction pathways are often restricted to a cascade of monotonically decreasing free energy steps. Allowing endergonic steps introduces flexibility to the uncertainty (and lack) of thermochemistry data, local vs. global synthesis conditions, etc., allowing the full compositional space to be traversed when considering chemical routes towards targets. This is exemplified in particular by the lithium carbonate decomposition reaction described previously, which is endergonic at low temperatures but is shown to occur spontaneously at elevated temperatures.

Finally, the network can also be used to identify shortest paths to or from *any* nodes in the network. This mechanism may lead to alternative insights about the network, such as lists of the shortest pathways to any target. We anticipate this setup of the network to be useful for speculating possible likely products, when no net reaction information is known a priori. Similarly, the network can be used in "reverse" to identify promising precursors which yield efficient chemical routes towards desired targets.

Conclusions

A chemical reaction network model was designed, implemented with thermodynamic data from the Materials Project, and demonstrated as a predictor of reaction pathways in solid-state chemistry. The framework effectively reduces the large, complex thermochemical landscape to a computationally tractable structure through i) creation of a weighted directed graph representation of the available thermodynamic phase space, ii) mapping of rigorous thermodynamic data and possible heuristics into a versatile cost function and iii) application of existing pathfinding algorithms to identify probable reaction routes. While the framework explores reaction trajectories in the most general way possible, allowing for parallel combined pathways, the combinatorial complexity is reduced by chemically motivated filters such as: 1) restricting the number of phases considered via thermodynamic stability arguments, 2) limiting the maximum number of simultaneously reacting phases, and 3) enforcing mass conservation via stoichiometric constraints. As a demonstration, the framework was shown to successfully identify a complex reaction pathway in a recently elucidated in-situ characterized solid-state synthesis of $YMnO_3$. We envision the methodology to be used to suggest possible synthesis routes and precursors that allow for efficient thermodynamic conditions towards desirable target phases, as well as identification of byproducts and possible thermodynamic sinks along synthesis routes. Future work will benefit tremendously by combining the framework 'live' with automated data collection, *in situ* phase identification, rapid analysis techniques, and automated feedback loops, moving towards active control of inorganic solid-state synthesis.

Methods

Thermochemical data

While the chemical reaction network can be created from any thermochemical data – computed, experimental, or a combination of both – in this work, we employ the Materials Project (MP), which contains well-benchmarked *ab initio* calculated formation enthalpies for over one hundred thousand dif-

ferent materials as calculated with density functional theory (DFT).[19,33] To capture the temperature dependence of thermodynamic phase space, we employ the machine-learned Gibbs free energy descriptor reported by Bartel et al.,[29] which estimates the finite temperature contribution to the Gibbs free energy of formation of solids, $\Delta G_f(T)$. This contribution incorporates both temperature-dependent enthalpic and entropic effects, although the entropic contribution (TS) typically dominates. The elemental Gibbs free energies used for these formation energy calculations are acquired from FactSage.[20] The Gibbs free energies of formation for non-elemental gases (e.g. CO_2) are acquired from NIST-JANAF experimental thermochemical tables where possible.[21] Thermodynamic stability (energy above the hull) is calculated via phase diagram construction in the *pymatgen* package.[34]

Graph creation, traversal, and visualization

All networks are implemented and analyzed using the *graph-tool* package.[35] Chemical reactions are balanced and combined in a high-throughput manner via the reaction balancing algorithm implemented in *pymatgen*. Graphs are visualized using Graphistry Hub.[36]

Combining chemical reactions via mass conservation

When a net reaction is known a priori, reaction steps identified during the pathfinding can be linearly combined to satisfy the stoichiometric mass constraints of the overall reaction. These constraints correspond to numerically solving the linear system of equations given by:

$$A\boldsymbol{m} = \boldsymbol{c} \tag{11}$$

where $\boldsymbol{m}$ is a vector containing the "multiplicity" of each reaction (i.e. the factor by which the entire reaction is multiplied), A is the matrix containing the stoichiometric coefficients of all phases present in all reactions where reactants/products have negative/positive coefficients respectively, and $\boldsymbol{c}$ is a vector containing the stoichiometric coefficients of the net synthesis reaction. We solve this system of equations for the multiplicity vector, $\boldsymbol{m}$, via application of the Moore-Penrose matrix pseudoinverse as implemented within the *SciPy* package.[37]

References

[1] James L Marshall and Virginia R Marshall. Rediscovery of the Elements: Cronstedt and Nickel. *The Hexagon of Alpha Chi Sigma*, 105(2):24–29, 2014.

[2] Francis J. Disalvo. Solid-state chemistry: A rediscovered chemical frontier. *Science*, 247(4943):649–655, February 1990.

[3] Holger Kohlmann. Looking into the Black Box of Solid-State Synthesis. *European Journal of Inorganic Chemistry*, 2019(39-40):4174–4180, October 2019.

[4] L. Soderholm and J. F. Mitchell. Perspective: Toward "synthesis by design": Exploring atomic correlations during inorganic materials synthesis. *APL Materials*, 4(5):053212, May 2016.

[5] Andreas Stein, Steven W. Keller, and Thomas E. Mallouk. Turning down the heat: Design and mechanism in solid-state synthesis. *Science*, 259(5101):1558–1564, 1993.

[6] Daniel P. Shoemaker, Yung Jin Hu, Duck Young Chung, Gregory J. Halder, Peter J. Chupas, L. Soderholm, J. F. Mitchell, and Mercouri G. Kanatzidis. In situ studies of a platform for metastable inorganic crystal growth and materials discovery. *Proceedings of the National Academy of Sciences of the United States of America*, 111(30):10922–10927, 2014.

[7] Daniel O'Nolan, Guanglong Huang, Gabrielle E. Kamm, Antonin Grenier, Chia-Hao Liu, Paul K. Todd, Allison Wustrow, Gia Thinh Tran, David Montiel, James R Neilson, Simon J. L. Billinge, Peter J. Chupas, Katsuyo S. Thornton, and Karena W. Chapman. A thermal-gradient approach to variable-temperature measurements resolved in space. *Journal of Applied Crystallography*, 53(3), June 2020.

[8] Andrew J Martinolich and James R Neilson. Toward Reaction-by-Design: Achieving Kinetic Control of Solid State Chemistry with Metathesis. *Chemistry of Materials*, 29(2):479–489, 2017.

[9] Paul K. Todd and James R. Neilson. Selective formation of yttrium manganese oxides through kinetically competent assisted metathesis reactions. *Journal of the American Chemical Society*, 141(3):1191–1195, 2019.

[10] Paul K Todd, Antoinette M M Smith, and James R Neilson. Yttrium Manganese Oxide Phase Stability and Selectivity Using Lithium Carbonate Assisted Metathesis Reactions. *Inorganic Chemistry*, 58(22):15166–15174, November 2019.

[11] Zhelong Jiang, Arun Ramanathan, and Daniel P Shoemaker. In situ identification of kinetic factors that expedite inorganic crystal formation and discovery. *J. Mater. Chem. C*, 5:5709, 2017.

[12] Akira Miura, Hiroaki Ito, Christopher J. Bartel, Wenhao Sun, Nataly Carolina Rosero-Navarro, Kiyoharu Tadanaga, Hiroko Nakata, Kazuhiko Maeda, and Gerbrand Ceder. Selective metathesis synthesis of $MgCr_2S_4$ by control of thermodynamic driving forces. *Mater. Horiz.*, 2020.

[13] Matteo Bianchini, Jingyang Wang, Raphaële J Clément, Bin Ouyang, Penghao Xiao, Daniil Kitchaev, Tan Shi, Yaqian Zhang, Yan Wang, Haegyeom Kim, Mingjian Zhang, Jianming Bai, Feng Wang, Wenhao Sun, and Gerbrand Ceder. The interplay between thermodynamics and kinetics in the solid-state synthesis of layered oxides. *Nature Materials*, 2020.

[14] Jeffrey I Steinfeld, Joseph Salvadore. Francisco, and William L Hase. *Chemical kinetics and dynamics*. Prentice Hall, Upper Saddle River, N.J., 1999.

[15] Michael P. Allen. *Introduction to Molecular Dynamics Simulation*. Computational soft matter: from synthetic polymers to proteins (NIC Series), Julich, 2004.

[16] Arthur F Voter. Introduction to the Kinetic Monte Carlo Method. In Kurt E Sickafus, Eugene A Kotomin, and Blas P Uberuaga, editors, *Radiation Effects in Solids*, pages 1–23, Dordrecht, 2007. Springer Netherlands.

[17] Adri C T van Duin, Siddharth Dasgupta, Francois Lorant, and William A Goddard. ReaxFF: A Reactive Force Field for Hydrocarbons. *The Journal of Physical Chemistry A*, 105(41):9396–9409, October 2001.

[18] Daniil V. Ilyin, William A. Goddard, Julius J. Oppenheim, and Tao Cheng. First-principles–based reaction kinetics from reactive molecular dynamics simulations: Application to hydrogen peroxide decomposition. *Proceedings of the National Academy of Sciences*, 116(37):18202–18208, 2019.

[19] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. 1(1), 2013.

[20] C.W. Bale, E. Bélisle, P. Chartrand, S.A. Decterov, G. Eriksson, A.E. Gheribi, K. Hack, I.-H. Jung, Y.-B. Kang, J. Melançon, A.D. Pelton, S. Petersen, C. Robelin, J. Sangster, P. Spencer, and M-A. Van Ende. Factsage thermochemical software and databases, 2010–2016. *Calphad*, 54:35–53, 2016.

[21] Jr Malcolm W. Chase. *NIST-JANAF thermochemical tables*. Fourth edition. Washington, DC : American Chemical Society ; New York : American Institute of Physics for the National Institute of Standards and Technology, 1998., 1998.

[22] Christopher W. Bale and Gunnar Eriksson. Metallurgical thermochemical databases—a review. *Canadian Metallurgical Quarterly*, 29(2):105–132, 1990.

[23] Anubhav Jain, Yongwoo Shin, and Kristin A Persson. Computational predictions of energy materials using density functional theory. *Nature Reviews Materials*, 1(1):15004, 2016.

[24] J. W. Gibbs. On the equilibrium of heterogeneous substances. *American Journal of Science*, 16(96):441–458, December 1878.

[25] Henry Eyring. The activated complex in chemical reactions. *The Journal of Chemical Physics*, 3(2):107–115, 1935.

[26] E W Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

[27] Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, page 451–457, Cambridge, MA, USA, 2000. MIT Press.

[28] Jin Y. Yen. Finding the k shortest loopless paths in a network. *Management Science*, 17(11):712–716, 1971.

[29] Christopher J. Bartel, Samantha L. Millican, Ann M. Deml, John R. Rumptz, William Tumas, Alan W. Weimer, Stephan Lany, Vladan Stevanović, Charles B. Musgrave, and Aaron M. Holder. Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry. *Nature Communications*, 9(1):4168, December 2018.

[30] A N Timoshevskii, M G Ktalkherman, V A Emel'kin, B A Pozdnyakov, and A P Zamyatin. High-temperature decomposition of lithium carbonate at atmospheric pressure. *High Temperature*, 46(3):414–421, 2008.

[31] Nils E. R. Zimmermann and Anubhav Jain. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC Adv.*, 10:6063–6081, 2020.

[32] Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.*, 18:13754–13769, 2016.

[33] Anubhav Jain, Geoffroy Hautier, Shyue Ping Ong, Charles J Moore, Christopher C Fischer, Kristin A Persson, and Gerbrand Ceder. Formation enthalpies by mixing GGA and GGA + U calculations. *Physical Review B*, 84:45115, 2011.

[34] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013.

[35] Tiago P. Peixoto. The graph-tool python library. May 2017.

[36] Graphistry. *Graphistry Hub*, 2020 (accessed June 18, 2020).

[37] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

Author Contributions

M.J.M. and S.S.D. conceived the idea of the presented work. M.J.M. designed and developed the *reaction-network* code with feedback from S.S.D. and K.A.P. M.J.M. wrote the manuscript with guidance of S.S.D. and K.A.P.

Acknowledgements

This work was supported as part of GENESIS: A Next Generation Synthesis Center, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under Award Number DE-SC0019212. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

The authors would like to thank J. Neilson and P. Todd for their helpful discussion regarding the reaction network model, as well as E. Persson for math skills, and L. Meyerovich for assistance with graph visualization.

Code availability

The *reaction-network* package was created in Python to implement the reaction network model described in this work. The code is free and available for use by the general community, via GitHub at `https://github.com/GENESIS-EFRC/reaction-network`.

Figures

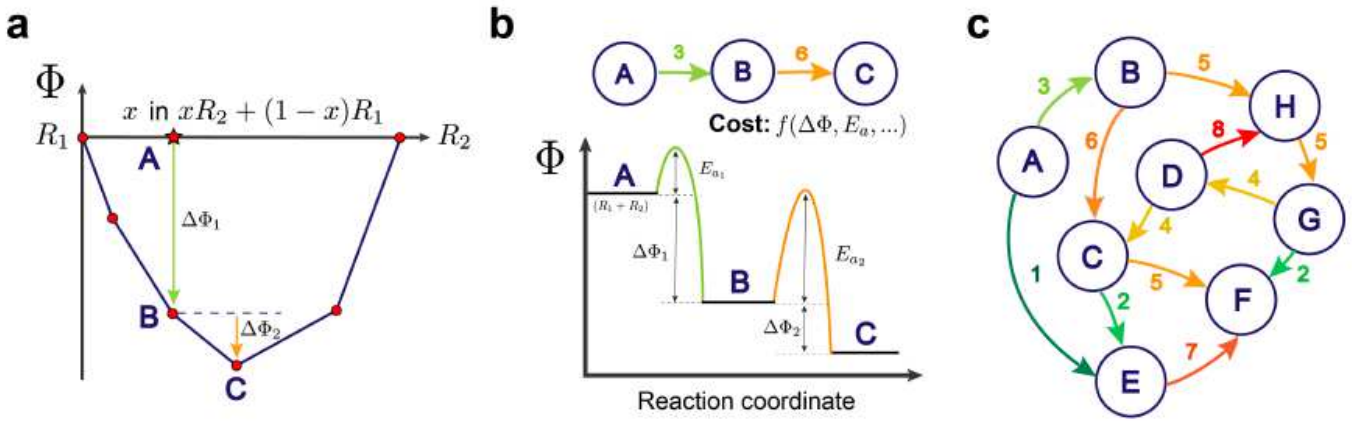

Figure 1

Models of chemical reactions in thermodynamic phase space, ordered by increasing degree of abstraction. The label, A, indicates a combination of arbitrary amounts of phases R1 and R2. Similarly, the other labels (B-F) are combinations of arbitrary amounts of other phases which result from the chemical reactions. (a) The convex hull construction for consecutive reactions between two reactants, R1 and R2. The points drawn indicate chemical reactions at different stoichiometric mixtures of the reactants, and the lines trace the convex hull indicating the thermodynamic equilibrium (minimum free energy) for all ratios of mixing, x. (b) A traditional reaction coordinate diagram used to represent both the free energies of reaction, ΔΦ, and activation energies, Eα. This is generalized by a weighted, directed graph connecting the three states (top). The cost/weight of the directed edges, shown as the colored number adjacent to each edge, is some function of the free energy change, activation energy, and other reaction features. (c) A chemical reaction network linking together many such possible reaction pathways that may occur within a set of phases.

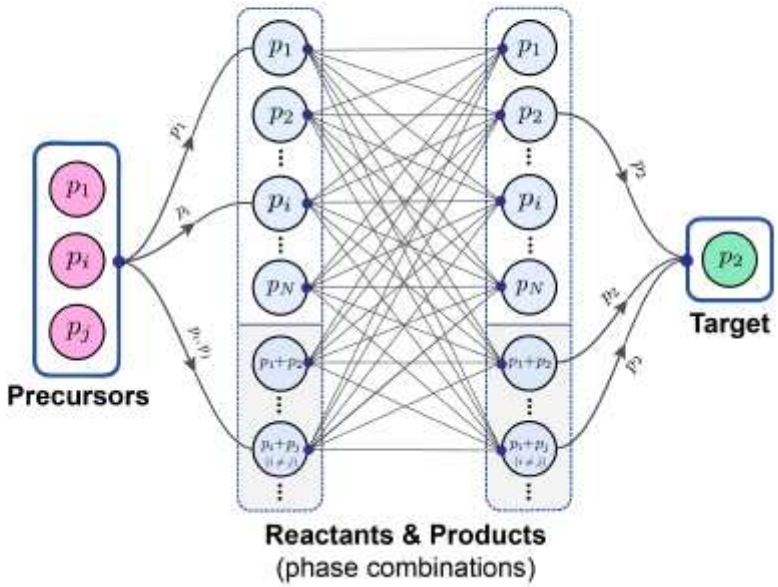

Figure 2

The generalized graph architecture of a solid-state chemical reaction network. The reaction network is constructed for a hypothetical chemical system containing N phases pi (i =1, 2, . . . N), with nodes made of distinct phase combinations up to maximum size n. The precursors and target nodes link into and out of the densely connected network of reactions, respectively. The edge directions between the reactant and product nodes have been omitted for clarity. These central edges include chemical reactions weighted by their cost value, as well as zero-weight edges which create loops from product nodes back to reactant nodes, such that multiple step reaction pathways can be captured.

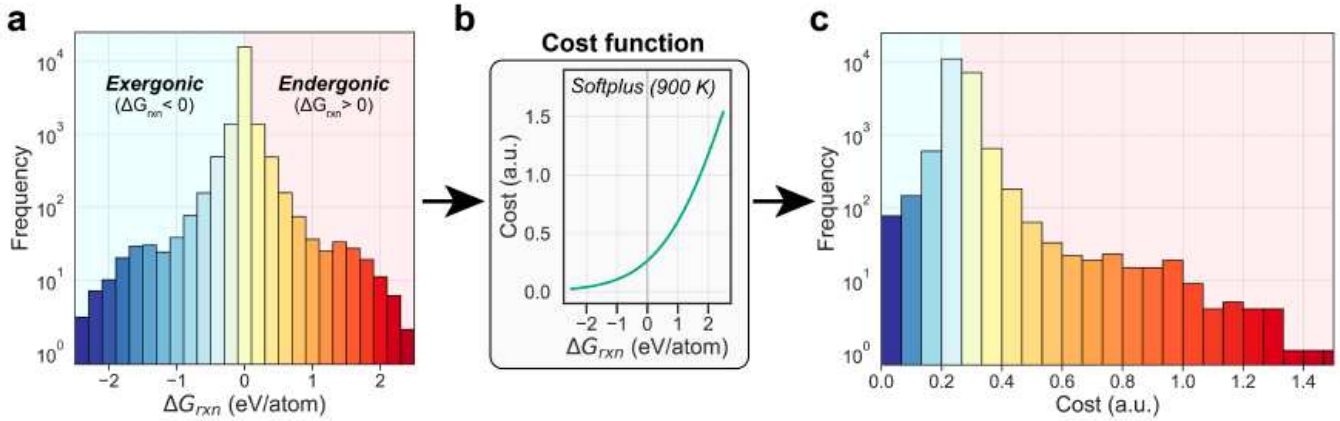

Figure 3

The effect of the cost function transformation on reaction energies. (a) Distribution of reaction Gibbs free energies (normalized per reactant atom) in the C-Cl-Li-Mn-O-Y chemical system. (b) Transformation of reaction energies via the softplus cost function described in equation (2). (c) Final distribution of reaction costs (in arbitrary units) after transformation. A reaction free energy of zero corresponds to a cost of ~0.265.

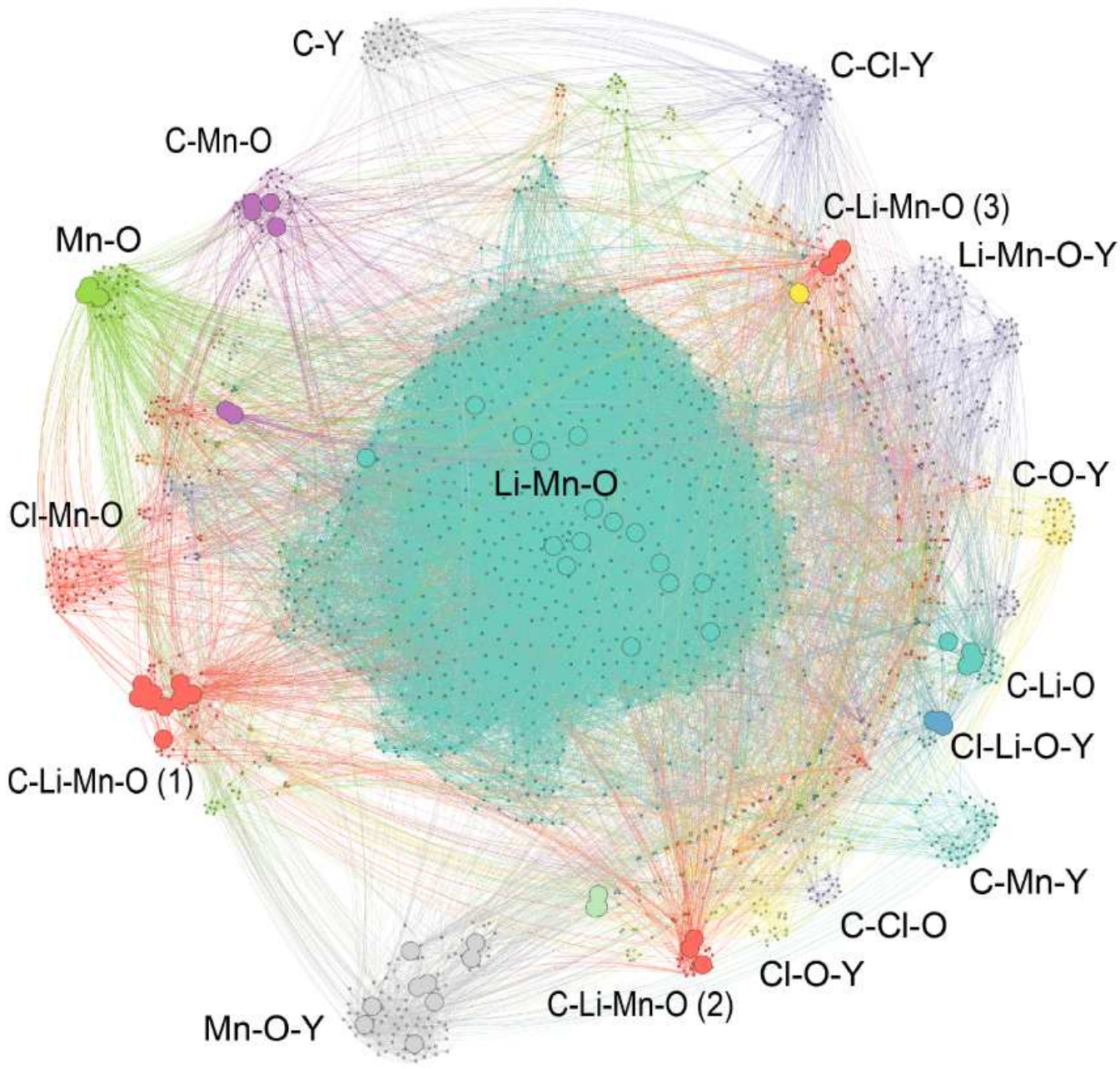

Figure 4

The reaction network for the C-Cl-Li-Mn-O-Y chemical system. The network contains 56 phases: 39 stable and 17 within +20 meV/atom above the hull. Chemical subsystems are labeled by color. The larger nodes

indicate reactant nodes which are traversed on the 20 shortest pathways from precursors to targets in the YMnO3 assisted metathesis reaction given by equation (3).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- rnSInaturecommsubmit2.pdf