Journal Name

ARTICLE TYPE

Predictive stochastic analysis of massive filter-based electrochemical reaction networks

Daniel Barter,[a,†] Evan Walter Clark Spotte-Smith,[b,c,†] Nikita S. Redkar,[c,d] Shyam Dwaraknath,[e] Kristin A. Persson[b,f] and Samuel M. Blau[a*]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Chemical reaction networks (CRNs) are powerful tools for obtaining mechanistic insight into complex reactive processes. However, they are limited in their applicability where reaction mechanisms are not well understood and products are unknown. Here we report new methods of CRN generation and analysis that overcome these limitations. By constructing CRNs using filters rather than templates, we can capture species and reactions that are unintuitive but fundamentally reasonable. The resulting massive CRNs can then be interrogated via stochastic methods, revealing thermodynamically bounded reaction pathways to species of interest and automatically identifying network products. We apply this methodology to study solid-electrolyte interphase (SEI) formation in Li-ion batteries, generating a CRN with $\sim 86,000,000$ reactions. Our methods automatically recover SEI products from the literature and predict previously unknown species. We validate their formation mechanisms using first-principles calculations, discovering novel kinetically accessible molecules. This methodology enables the *de novo* exploration of vast chemical spaces, with the potential for diverse applications across thermochemistry, electrochemistry, and photochemistry.

Main

Understanding and controlling chemical reactivity has been central to technological advances in manufacturing, transportation, pharmaceuticals, electronics, and beyond. Traditionally, studies of reactivity have been conducted by hand using either trial-and-error experiments or low-throughput molecular simulations. However, pressing challenges in areas including electrochemistry,[1,2] atmospheric chemistry,[3] and catalysis[4,5] present a degree of complexity that cannot be tractably described by intuitively-driven investigation. To that end, recent years have seen the development of a range of automated computational methods to probe complex systems at the atomistic level. Among these methods are techniques to explore the potential energy surface (PES) of a system of interest using quantum chemical methods[6] and methods to analyze chemical reaction networks (CRNs).[7] PES exploration techniques - including *ab initio* molecular dynamics,[8] artificial force-induced reactions,[9] and stochastic surface walking,[10] among others[6] - typically suffer from prohibitively high cost, limiting their use to simple systems involving only small molecules or very short (∼1ps) time scales. On the other hand, because CRNs abstract away much of the complexity of the PES, representing chemical transformations as discrete steps, they do not incur extreme costs. For more tractable exploration of reactive spaces, CRNs are thus highly attractive tools.

Most previous CRNs have been constructed based on human chemical intuition. By applying reaction templates to include only commonly observed mechanisms[11–15] or pruning by the "chemical distance" between species (the number of bonds that must change for a reaction to occur, or the number of reactions required to transform reactants to products) to focus only on starting species and known products of interest,[16] it is possible to create relatively simple networks capable of elucidating reaction pathways in thermochemical systems.

However, chemical intuition is limited and unreliable when describing new reactive spaces,[17] impeding the design of next-generation technologies. In such complex and poorly understood chemical contexts, such as where charged and open-shell species dominate and must be accurately described, the use of CRNs is

[a] *Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, United States.*
[b] *Department of Materials Science and Engineering, University of California, Berkeley, CA 94720, United States.*
[c] *Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, United States.*
[d] *Department of Chemical Engineering, University of California, Berkeley, Berkeley, CA 94720, United States.*
[e] *Luxembourg Institute of Science and Technology, Luxembourg.*
[f] *Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, United States.*
[†] *These authors contributed equally.*
∗ smblau@lbl.gov

challenging. In electrochemistry, for instance, studies of reaction mechanisms[5,18] and even characterization of reaction products[1] can be challenging. As a result, CRN methods that rely on templates or the chemical distance to known products cannot be effectively used to study electrochemical reactivity.

Several methods have been developed that bypass the need for prior knowledge of reactions and products by applying PES exploration to build CRNs.[19,20] In general, these methods face the crippling cost limitations of PES exploration. Recently, however, Zhao and Savoie devised Yet Another Reaction Program (YARP) to accelerate PES exploration-based generation of CRNs containing gas-phase organic molecules.[20] By combining the semi-empirical quantum chemistry code GFN2-xTB[21] with density functional theory (DFT), YARP is able to achieve an admirable 100-fold reduction in cost compared to using DFT alone. In spite of YARP's considerable promise for conventional organic chemistry, its applicability in domains such as electrochemistry is presently limited. Since YARP relies on GFN2-xTB to pre-optimize reaction pathways, it can only be expected to perform well in domains where GFN2-xTB produces reasonable geometries. However, we have previously shown[22] that GFN2-xTB is not an effective tool for accelerating geometry optimization of many classes of molecules that are important in electrochemistry. Specifically, we found that for complex molecules, including radicals, charged species, and metal-coordinated species in solution, GFN2-xTB produces geometries that are inconsistent with DFT nearly half of the time. Given the unreliability of GFN2-xTB and the intractable cost of pure DFT PES exploration, new methods must be developed to allow for the *de novo* study of reactive electrochemistry using CRNs.

We recently developed the first software infrastructure for the construction and analysis of CRNs capable of providing insight into a complex electrochemical cascade, namely the formation of the solid electrolyte interphase (SEI) in lithium-ion batteries. After constructing graph-based CRNs containing thousands of species and millions of reactions, we used shortest-path algorithms to identify optimal pathways to two key SEI products, lithium ethylene dicarbonate (LEDC)[23] and lithium ethylene monocarbonate (LEMC).[2] With this approach, we recovered previously proposed reaction mechanisms and predicted novel thermodynamically and kinetically reasonable pathways.

In our previous work, the networks that we studied were limited in size by the computational cost of both network construction and network analysis (particularly due to the poor scaling of shortest-path algorithms). These costs constrained the number of species as well as the number and types of reactions contained in the networks. Our prior graph-based analysis approach was also not fully predictive; in order to use shortest-path algorithms, products of interest had to be known *a priori*. Further, we could only search for the optimal pathways to a single product of interest at a time. Here we confront the more challenging and general problem of exploring a reactive space without significant *a priori* knowledge of end products. Specifically, we seek to search for many feasible pathways under various starting conditions to a range of products, byproducts, and intermediates, including products that might be unknown at the time of network construction.

To overcome the limitations of our previous work, here we present a new approach to automatically construct and explore CRNs in complex domains such as electrochemistry that is capable of extracting unique, predictive insights from much larger networks than before. We first describe our method of **Hi**gh-**P**erformance **R**eaction **Gen**eration (HiPRGen). HiPRGen efficiently enumerates all stoichiometric reactions and uses extensible filters, rather than prescriptive templates, to eliminate reactions based on physical or practical criteria while retaining a massive set of potentially unintuitive but chemically reasonable reactions. To overcome the scaling limitations of graph-based pathfinding, we use a Monte Carlo-based approach to explore CRNs, where the reactive space is sampled using reaction thermodynamics. This approach automatically yields paths to any molecule formed in the trajectories and heuristically identifies the products of the network as a function of initial conditions. Thus, the combination of HiPRGen with stochastic network analysis allows for the exploration of electrochemical reactive spaces without prior knowledge of reaction mechanisms or end products for the first time. We demonstrate and validate this approach with an application to SEI formation. We first identify 36 products of a HiPRGen-constructed network of roughly 86,000,000 reactions - including both novel species and many products reported in the SEI literature - by analyzing thermodynamically bounded Monte Carlo trajectories. To demonstrate the plausibility of these network products, we then kinetically refine the shortest paths to novel species of interest with first-principles calculations. We discover kinetically accessible formation pathways to multiple novel network product species which may participate in SEI formation, underscoring the critical importance of the inclusion of unintuitive species and reactions in CRN construction. The methods described here can be widely applied to initiate predictive studies of reactivity in domains where existing knowledge is limited.

Results

Template-Free Reaction Network Generation

A CRN is defined by a set of species **S** and a set of reactions **R** linking those species. Inspired by previous work by Kim[16] and Xie[2] where the chemical distance between species was used to selectively include reactions to a CRN without the use of templates, we have devised HiPRGen, which uses filters to remove some unphysical or otherwise undesirable species and reactions while preserving possibly unintuitive mechanisms.

We begin with some large dataset of species, the properties of which are known from e.g. quantum chemical calculations. We then apply a series of filters, where each filter can remove species that are chemically unreasonable or otherwise undesirable under the conditions studied (Figure 1-1). A list of species filters that we have designed and employed is described in Methods and is discussed in more detail in the Supplementary Information.

The filtered set of species $\mathbf{S}_{filtered}$ is then used to populate buckets each defined by a unique composition (Figure 1-2). Buckets are populated by members - either one or two species which together have the composition of the bucket. This means that any two members of a given bucket define the reactants and products

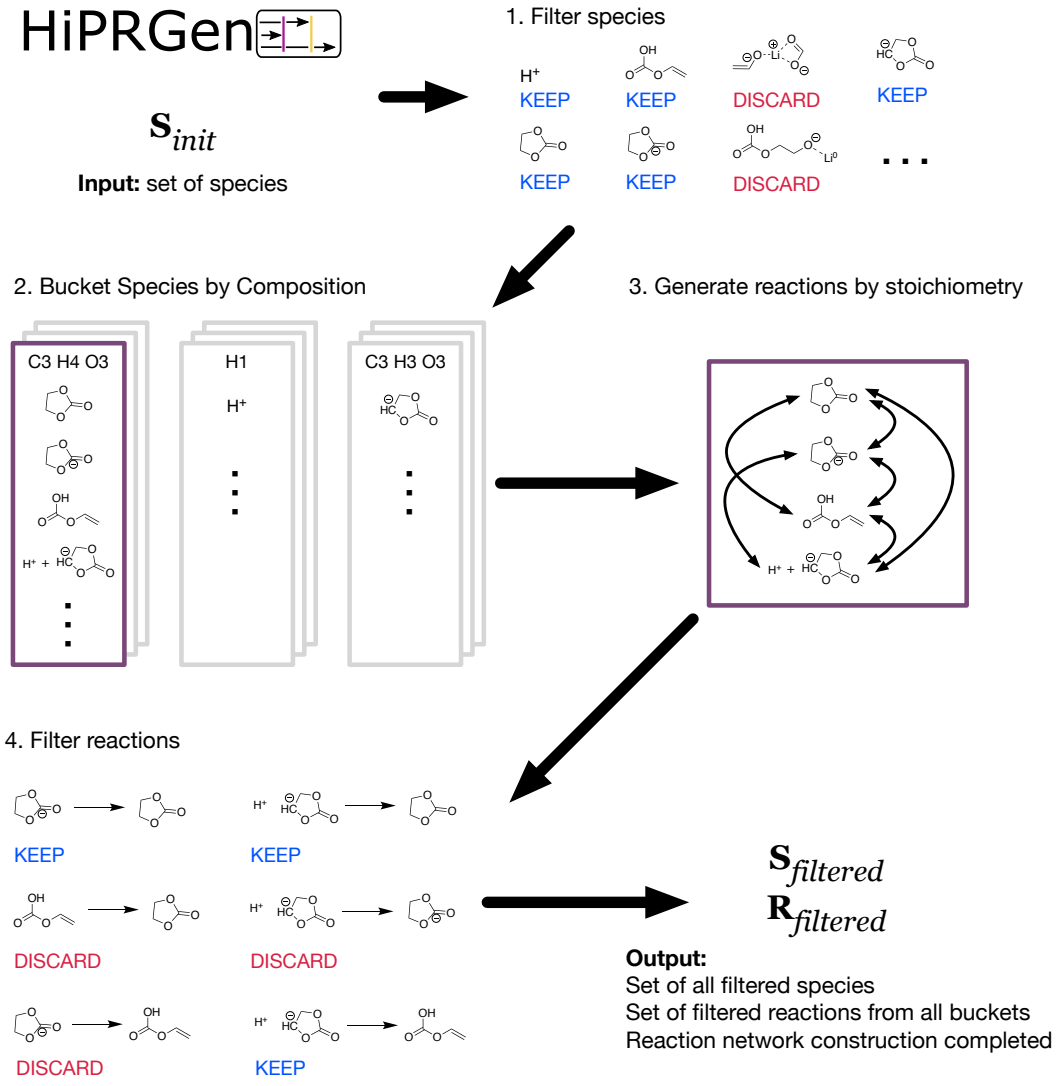

Fig. 1 A schematic overview of the High-Performance Reaction Generation (HiPRGen) method. A set of species $\mathbf{S}_{init}$ is provided as input. **1.** The species contained in $\mathbf{S}_{init}$ are filtered out based on some user-defined criteria. Here, species including non-ionic lithium (Li^0) and species where multiple fragments are connected only by coordination bonds, are removed. **2.** Species are grouped and bucketed based on composition. Each bucket is populated by entries that contain either a single molecule or a pair of molecules that together have the composition of the bucket. **3.** Within each bucket, all stoichiometrically valid reactions are generated. This corresponds to generating all combinations of two members of the bucket. **4.** The generated, stoichiometrically valid reactions are then filtered out, subject to user-defined reaction filters. Here, dissociative redox reactions (where changes in bonding occur simultaneously with reduction or oxidation) and reactions involving more than two bonds changing are removed. After aggregating the reactions generated from each bucket, the end result of the HiPRGen procedure is a set of filtered species $\mathbf{S}_{filtered}$ and a set of filtered reactions $\mathbf{R}_{filtered}$ constituting a reaction network.

of a stoichiometrically balanced chemical reaction containing one or two reactants and one or two products. For each bucket, all combinations of two unique members yield unique reactions (Figure 1-3). Note that, because we allow for electrochemical reactions, charge is not necessarily balanced in these reactions. For a system of several thousand species, there can easily be hundreds of billions or even trillions of stoichiometrically valid reactions. Reaction filters are therefore employed to remove reactions that, despite being stoichiometrically valid, are chemically implausible or otherwise undesirable (Figure 1-4). Due to the massive scale of the reaction filtering problem, we note that computational efficiency and parallelism are imperative. Specifically, reaction filters must use a minimal amount of data to parallelize without out of memory errors and must have very low computational cost in order to allow filtering to complete in a reasonable amount of time (hours to days). All employed reaction filters are described in Methods and are discussed in more detail in the Supplementary Information. Finally, the reactions from each bucket that pass all filters are aggregated. The result of HiPRGen is a set of filtered species $\mathbf{S}_{filtered}$ and filtered reactions $\mathbf{R}_{filtered}$, which constitute a CRN.

We note that HiPRGen can exhaustively enumerate all possible reactions between up to approximately 10,000 species, over an order of magnitude more than any previous approach.[2] Further, to the best of our knowledge, HiPRGen is the first method that combines an exhaustive enumeration of stoichiometrically valid reactions with a range of chemically-motivated filters. In addition to improving the thoroughness and efficiency of reaction generation compared to previous methods (see Supplementary Information), HiPRGen has the benefit that the filters can be modified and extended by the users, making it facile to apply HiPRGen to new chemical domains.

From the CRN generated by HiPRGen, it becomes possible to search for diverse products and reaction pathways to those products. However, even after filtering the set of stoichiometrically valid reactions, the number of remaining reactions can be so vast that a highly scalable method of network analysis is required.

Thermodynamically Bounded Network Analysis Via Stochastic Methods

While it might be desirable to use shortest-path algorithms to identify reaction pathways in graph-based reaction networks, as we did previously,[2,23] such algorithms become computationally intractable as network size increases. We therefore turn to the kinetic Monte Carlo (kMC) algorithm of Gillespie,[24] which, with appropriate modifications,[25] can scale sublinearly with number of reactions. In a kMC simulation, a system evolves from some user-defined initial state in a manner that is non-deterministic but consistent with the rate coefficients provided to the model.

When templates are viable and accurately describe the reactivity in a system, they can be used to approximate reaction kinetics with minimal cost.[11,13] On the other hand, in a template-free network of potentially millions of reactions, it is completely impossible to include accurate rate coefficients for all reactions. For the purposes of stochastic network exploration and analysis, we therefore assign rate coefficients based on the reaction free energy ΔG rather than the reaction energy barrier $\Delta G^{\ddagger}$, which is unknown at this point. Exergonic reactions ($\Delta G \leq 0$) are assumed to be barrierless and all have the same rate coefficient given by the Eyring prefactor ($\frac{k_B T}{h}$), while endergonic reactions ($\Delta G > 0$) have rate coefficients given by the Eyring equation ($\frac{k_B T}{h} exp(-\Delta G^{\ddagger}/k_B T)$) with the reaction free energy serving as an effective barrier ($\Delta G^{\ddagger} = \Delta G$). This means that many reactions that might not happen in reality due to high energy barriers will nonetheless occur in the kMC simulations, but critically, all reactions that will occur in reality will also occur in the kMC simulations. In this way, we can say that the results of these kMC simulations are bounded by reaction thermodynamics.

To analyze a CRN with thermodynamically bounded kMC, we perform a large number of Monte Carlo simulations in parallel (Figure 2a). The result of each simulation is a series of reactions defining the trajectory of the system. We trace each trajectory, examining if a molecule of interest has been formed, and if so, identifying the shortest sequence of reactions leading to its first formation (Figure 2b). Performing this method of stochastic pathfinding over many trajectories, we identify a range of possible pathways to the molecule of interest. We can then rank the identified paths in order to identify the "best" paths among those observed, as defined by some cost function (see Methods). The thermodynamic pathways obtained from network analysis can then be subjected to further analysis to identify complete mechanisms, including transition-states and energy barriers. We emphasize that while graph-based pathfinding becomes intractable above approximately 10 million reactions, pathfinding with thermodynamically bounded kMC can scale to approximately 300 million reactions and consistently recovers the graph-based shortest paths where the methods can be compared (see Supplementary Information).

However, pathfinding is useful only if one already knows what molecule to search for. Thermodynamically bounded kMC, unlike graph-based pathfinding, can explore a reactive space without a specific target. This is because, while kMC trajectories can be used to search for a specific specie, they are neither produced with any specific specie in mind, nor are they biased towards any such specie. A unique capability that emerges from the exploratory nature of stochastic network analysis is the ability to identify products of a CRN without prior knowledge. To do this, we apply a set of heuristic criteria to the collection of trajectories (Figure 2c). We define a network product as any specie that is on average formed significantly more than it is consumed, that accumulates significantly in an average trajectory, and that can be reached by low-cost reaction pathways (see Methods).

Via this heuristic method, we are able to analyze the structure of the CRN itself. The average trajectory (Figure 2c) satisfies a rate equation of the system,[26,27] as demonstrated by the observed smoothing that indicates convergence from sufficient sampling. This means that our method of defining network products is mathematically rigorous (see Supplementary Information for further detail). While the products of the network are not necessarily the metastable or stable products that would be observed experimentally, the network products provide useful hypotheses

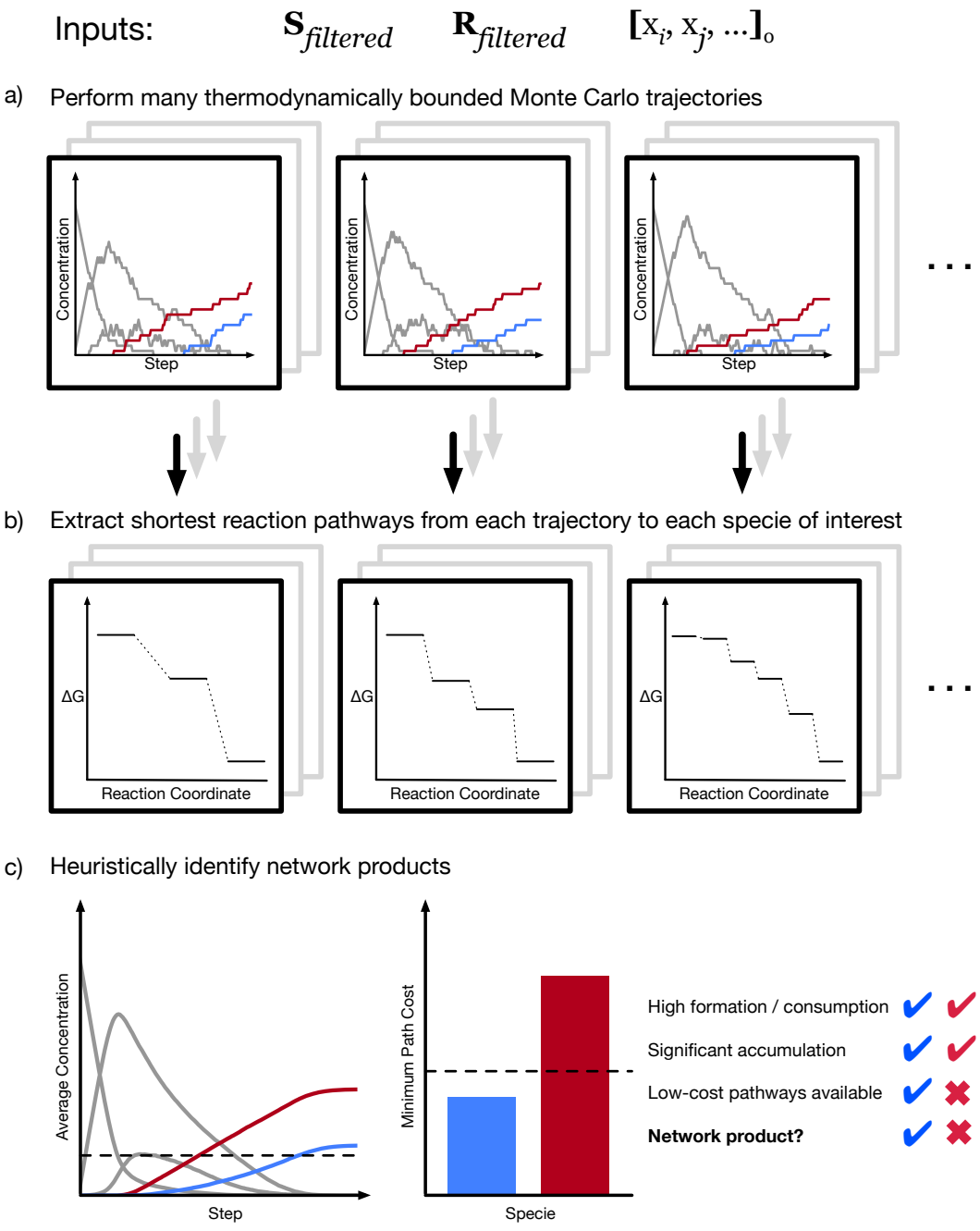

Fig. 2 Methods for analyzing CRNs from stochastic simulations. a) A large number of kinetic Monte Carlo trajectories are calculated, beginning with the same network (defined by $\mathbf{S}_{filtered}$ and $\mathbf{R}_{filtered}$) and the same initial state ($[x_i, x_j, \cdots]_0$, where x_q is the quantity of species q). **b)** In each trajectory, the shortest reaction pathway to some species of of interest can be identified. Note that because these trajectories are stochastic, different trajectories will often yield different shortest pathways to the same product. **c)** To identify products of the network, a set of heuristics are applied. In order to be considered a product of the CRN, a specie must be formed substantially more than it is consumed and must accumulate to a significant degree on average (that is, its average final concentration must be higher than some threshold). In addition, a product specie must be reachable by some low-cost path. In the example provided, both the red and the blue species are formed significantly more than they are produced and accumulate, but only the blue specie can be reached by a low-cost pathway. Therefore, by this heuristic, the blue specie is a network product, while the red specie is not.

regarding what might form in an actual reactive system. We can then interrogate these hypotheses and validate them by either theoretical or experimental means. We note that network products can depend on the initial state of the system.

Automatic Identification of Battery SEI Network Products

Using HiPRGen, we constructed a reaction network that seeks to describe SEI formation in lithium-ion batteries. We begin with an initial set of 8,904 species taken from the Lithium-Ion Battery Electrolyte dataset,[28] which was strategically designed to facilitate studies of electrolyte reactivity and interphase formation (see Methods). Network construction resulted in a CRN containing 5,193 filtered species and 86,001,275 filtered reactions. This is, to our knowledge, by far the largest chemical reaction network ever constructed from first-principles calculations (see Supplementary Information). With this network, we conducted 100,000 kMC trajectories under four conditions, with combinations of two different applied potentials (+0.0V vs. Li/Li^+ and +0.5V vs. Li/Li^+) and two different initial states (one consisting only of Li^+ and ethylene carbonate (EC) and the other consisting of Li^+, EC, and CO_2). Average trajectories for each condition are shown in the Supplementary Information.

Our approach is unique in its capacity to describe the complexity of a reactive space including thousands of charged and open shell species. Figure 3 graphically depicts the systematic reduction in complexity that is inherent to filter-based network construction followed by stochastic analysis. While HiPRGen seeks to preserve the vast range of chemically reasonable reactions that could occur in a system containing charged and radical species, only 312,034 of the over 86 million unique reactions in the network are actually observed in a set of 100,000 kMC trajectories run to completion from one initial condition. Further, of the 3,182 unique species observed to form in the same set of trajectories, just 18 are identified as network products. Only an automated computational infrastructure could possibly navigate such a vast space of interconnecting species and reactions to guide further investigation, and no previously reported method can be applied on this scale.

The utility of our approach is demonstrated through analysis of the 36 network products collected from the set of four conditions previously described (Figure 4). We first note that our automated procedure recovers 15 species that include a majority of the experimentally observed products of SEI formation (Fig. 4 solid dark green). These include gases (H_2, C_2H_4, CO),[29] inorganic species (lithium carbonate (Li_2CO_3) and lithium oxalate ($Li_2C_2O_4$)),[30,31] and alkyl carbonates (including species closely related to LEDC[30–32] and LEMC,[1,31] as well as lithium methyl carbonate, lithium butylene dicarbonate (LBDC, or in this case $LiBDC^-$),[31] and vinyl carbonate).[33] We emphasize that these species are recovered even though network exploration is based only on reaction thermodynamics, with reaction kinetics entirely ignored.

In addition to these well-known species, there are also a number of novel products that have not previously been proposed to participate in SEI formation. Among these are seven additional alkyl carbonates (Fig. 4 dotted light green) which are each very similar to known products in molecular size, composition, bonding, and contained functional groups. Due to the extreme difficulty of experimentally characterizing the SEI and the resulting limited ability to resolve small signal to noise,[34] the likely spectroscopic similarity[35] of these species to the known products means that they may already be unknowingly present in the SEI in small quantities.

Other network products include a number of cyclic species, such as 4,4',5,5'-tetrahydro-2,2'-bi(1,3-dioxolylidene), which we abbreviate as bi-dioxolylidene, and several species containing carboxylate, ester, and oxide functional groups, such as lithium 2-(formyloxy)ethan-1-olate, which we abbreviate as LFEO. Bi-dioxolylidene and LFEO (Fig. 4 dashed purple) were particularly unexpected given how distinct they are from the expected SEI products as well as most of the other novel network products. Evaluating whether or not these products will actually form in the SEI necessitates considering kinetic limitations and reactive competition. However, as we will show, using the thermodynamically shortest paths to guide automated transition state calculations identifies kinetically viable paths to both bi-dioxolylidene and LFEO (see below for analysis of bi-dioxolylidene and see the Supplementary Information for analysis of LFEO).

On the other hand, there are some network products which obviously do not reflect the corresponding chemical system. Specifically, both vinylene carbonate (VC) and propylene carbonate (PC) are known to rapidly decompose when included in battery electrolytes.[36,37] This contradiction indicates that there are missing reactions or, more likely given the thoroughness of the HiPRGen procedure, missing species necessary to facilitate the decomposition of VC and PC. The identification of this gap through the use of stochastic network analysis and network product prediction provides a tractable path forward to expand the CRN via selective addition of missing molecules that enable redox, decomposition, or recombination of network products with other abundant intermediate or product species.

Kinetically Feasible Pathways to Products of Interest

The reaction pathways produced by our stochastic approach are based solely on reaction thermodynamics, with no knowledge of reaction kinetics. In actuality, the dominant reaction pathways are heavily dependent on reaction energy barriers $\Delta G^{\ddagger}$ and rate coefficients. In order for thermodynamically bounded kMC to provide useful chemical insights, it is therefore critical that the thermodynamic reaction pathways can be reasonably translated into complete reaction paths including transition-states (TS) and energy barriers.

Using our stochastic approach, we can identify the N shortest thermodynamic reaction pathways to the network products, ranked by a cost function that we have employed previously (see Methods). We selected two unexpected network products - bi-dioxolylidene and LFEO - and subjected their thermodynamically shortest pathways in order of cost to an automated procedure to identify the TS for each elementary step along each pathway (see Methods), allowing for the construction of complete reaction

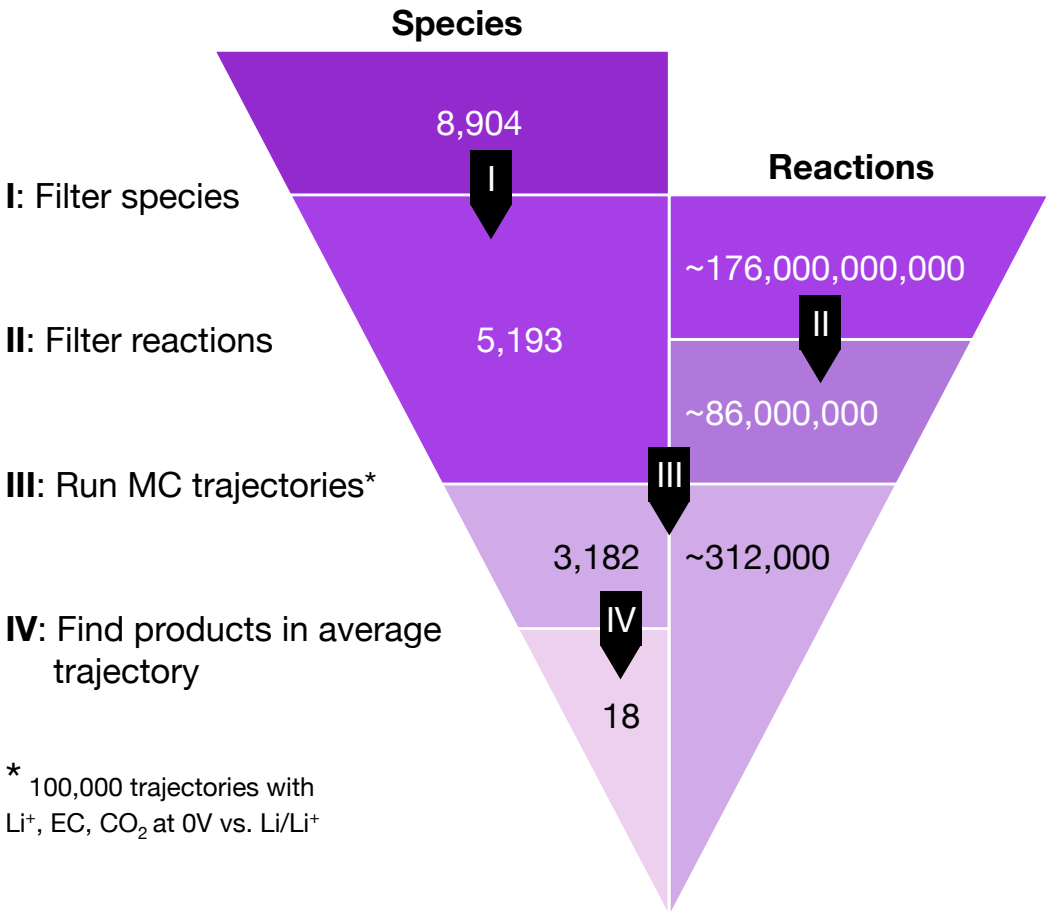

Fig. 3 CRN construction and stochastic analysis for SEI formation yields a systematic reduction of complexity. I. 8,904 initial species are reduced to 5,193 species via HiPRGen species filters. **II.** Of the approximately 176 billion possible stoichiometric reactions between the 5,193 species, just over 86 million pass HiPRGen reaction filters. **III.** 100,000 thermodynamically bounded Monte Carlo trajectories are run to completion with Li^+, EC, and CO_2 as starting species at +0.0V vs. Li/Li^+, and in total only 3182 species and approximately 312,000 unique reactions are observed. **IV.** Heuristic analysis of the average trajectory yields only 18 network product species.

mechanisms. Kinetically viable mechanisms for each were identified within the top 20 pathways, validating the ability of our approach to discover novel accessible species and mechanisms. The pathway to bi-dioxolylidene is shown in Figure 5 and discussed in more detail below, while the pathway to LFEO is provided in the Supplementary Information. Novel chemically interesting pathways to expected product vinyl carbonate discovered in the same manner are also shown in the Supplementary Information.

Bi-dioxolylidene can be formed through a dimerization of the carbene 4,5-dihydro-1,3-dioxol-2-ylidene, which we abbreviate as dioxolylidene. This pathway is based on the overall third-shortest thermodynamic pathway to bi-dioxolylidene from amongst all four scenarios. First, the carbene forms by a facile mechanism involving the reaction of a doubly reduced EC (Li^+EC^{2-}) with CO_2 (Figure 5a). A byproduct of this process is $LiCO_3^-$, which can form Li_2CO_3 simply by coming in contact with an additional Li^+. Once two dioxolylidene have formed via this mechanism, they can combine to form bi-dioxolylidene (Figure 5b). Although bi-dioxolylidene has not previously been reported in the SEI literature (nor, to the best of our knowledge, *anywhere* in the chemical literature), this pathway suggests that dioxolylidene should form in a battery electrolyte if CO_2 is present, and that bi-dioxolylidene formation is plausible if dioxolylidene is formed in abundance. In future work, we intend to combine multiple mechanisms to SEI products into a microkinetic model in order to assess reactive competition and determine what pathways are actually taken and what products actually form.

Discussion

The method described here, combining HiPRGen with stochastic CRN analysis, is broadly applicable to studies of reactivity involving homogeneous molecular reactions. At present, however, heterogeneous reactions involving surface sites are not supported. The ability to automatically generate reactions involving surfaces, and to perform stochastic simulations of massive reaction networks involving spatial dependence, will be necessary additions to be able to study heterogeneous catalysis and other systems involving surface-mediated reactivity. The other major limitation concerns species generation. The quality of a CRN generated by HiPRGen depends on the species initially provided; if there are key species missing initially, then it is possible that important reaction pathways and products will not be observed. We hope to

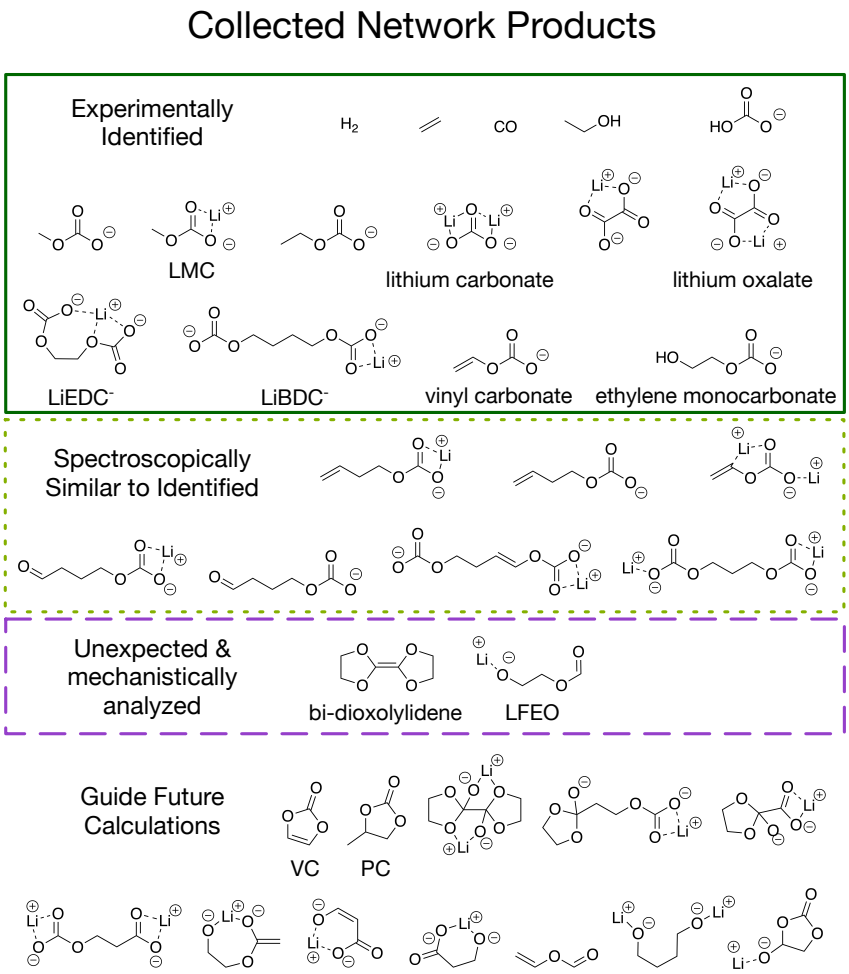

Fig. 4 The 36 total collected network products from four different initial conditions (+0.0V vs. Li/Li^+ with Li^+ and EC as starting species; +0.0V vs. Li/Li^+ with Li^+, EC, and CO_2 as starting species; +0.5V vs. Li/Li^+ with Li^+ and EC as starting species; and +0.5V vs. Li/Li^+ with Li^+, EC, and CO_2 as starting species). 15 of the network products are species which have previously been experimentally identified in the SEI, outlined in green, including dominant gaseous products, dominant molecular inorganic components, and dominant organic components. Seven of the network products, outlined in dotted light green, are species which would have very similar spectroscopic signatures to the dominant organic components, and thus may unknowingly be present in the SEI in small quantities. Two of the network products, bi-dioxolylidene and LFEO, outlined in dashed purple, are unlike any SEI species ever previously proposed and were subjected to further mechanistic analysis; the identification of kinetically accessible pathways validates our approach (see below for analysis of bi-dioxolylidene and see the Supplementary Information for analysis of LFEO). Finally, the remaining 12 network products may indicate that our CRN is missing species, may be kinetically inaccessible, or may be true SEI products, motivating future calculations.

address both of these limitations in future work.

To summarize, we have here described our approach to explore reactivity using CRNs without reliance on prior knowledge. The HiPRGen method allows for the construction of CRNs with vast numbers of chemically reasonable but potentially unintuitive species and reactions. The resulting massive networks can then be analyzed using thermodynamically bounded kMC, allowing for the identification of network products and reaction pathways which can be further refined into full mechanisms. We applied this methodology to study SEI formation in Li-ion batteries, generating a network consisting of over 5,000 species and over 86,000,000 reactions. Automatic network product prediction yielded many known SEI products as well as several products that had not previously been proposed. We identified kinetically viable mechanisms for the formation of multiple novel network products (bi-dioxolylidene and LFEO) based on pathways obtained from CRN analysis and refined using automated DFT, thereby validating that our network product prediction method proposes species that, while unexpected, are kinetically accessible. Moreover, this suggests that thermodynamic shortest pathways, obtained via analysis of appropriately filtered CRNs, can be used to efficiently search for transition-states and kinetically viable mechanisms, facilitating the construction of microkinetic models in the future. With the approach presented here, it will be possible to provide insights into a range of domains where fundamental understandings of reactivity are limited.

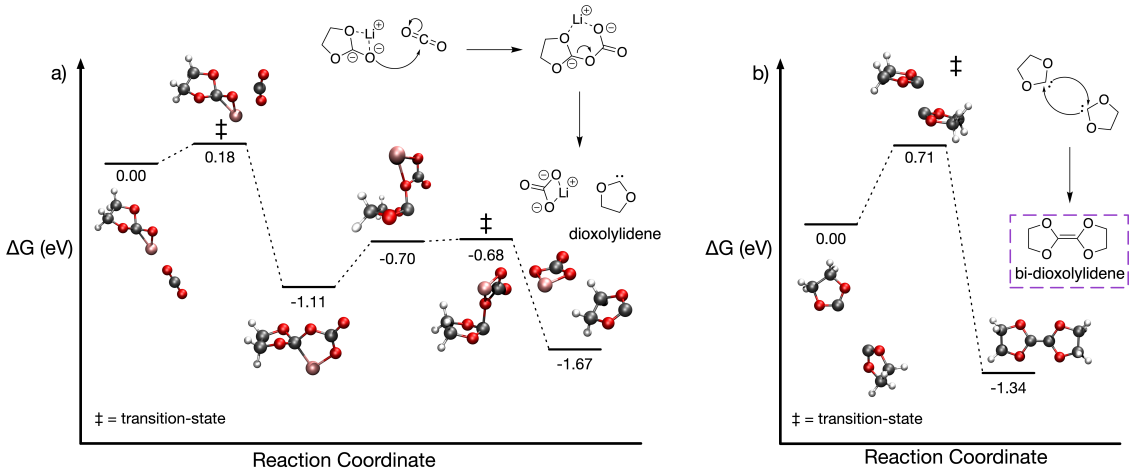

Fig. 5 Kinetically accessible reaction pathway to bi-dioxolylidene identified from the 3rd shortest thermodynamic path. a) The dioxolylidene carbene can be formed by the reaction of a doubly reduced Li^+EC^{2-} with CO_2. CO_2 can add with a small barrier ($\Delta G^\ddagger = 0.18$eV, $\Delta G = -1.11$eV) to Li^+EC^{2-}, resulting in an intermediate specie. After rearrangement of Li^+, this specie can form dioxolylidene by eliminating $LiCO_3^-$ with a small barrier ($\Delta G^\ddagger = 0.02$eV, $\Delta G = -0.99$eV). Once two dioxolylidenes have formed via this facile pathway, then they can dimerize (**b**) to form bi-dioxolylidene ($\Delta G^\ddagger = 0.71$eV, $\Delta G = -1.34$eV).

Methods

Species Selection

Initial species were taken from the Lithium Ion Battery Electrolyte (LIBE) dataset.[28] The LIBE dataset contains the properties of 17,190 species of various charges (-1, 0, 1) and spin multiplicities (1, 2, 3) calculated using density functional theory (DFT) at the ωB97X-V/def2-TZVPPD/SMD level of theory.[38–40] LIBE was designed for study of reactivity in battery electrolytes, making it well suited for the construction of CRNs and especially well suited for interrogating SEI formation.

For this work, network construction began with a subset of LIBE containing all species comprised of only carbon, hydrogen, oxygen, and/or lithium. This subset, which we call LIBE-CHOLi, contains 8,904 species.

Modifying Species Thermodynamics

All calculations on the species present in LIBE were conducted in an implicit solvation environment. While implicit solvation is generally accurate enough for the calculation of properties such as the solvation energy[41] and redox potentials of organic molecules,[42] we have found that even highly accurate implicit solvation methods severely underestimate the stabilization of small ions, especially metal ions, by solvent. This means that species in LIBE containing Li^+ ions with many coordination bonds are in many cases vastly more stable according to DFT than those with fewer coordination bonds, even if the corresponding species without lithium present are significantly less stable (an example is provided in the Supplementary Information). This insufficient stabilization led to inaccurate thermodynamics for reactions where the overall charge of the system was constant but the number of coordinate bonds changed (non-redox reactions).

To correct for insufficient metal ion stabilization, we optimized Li^+EC_n clusters at the ωB97X-D/def2-SVPD/PCM//ωB97X-V/def2-TZVPPD/SMD level of theory, with $n \in 0,1,2,3,4$ to estimate the stabilizing effect of each solvent molecule on Li^+. The lower level of theory (ωB97X-D/def2-SVPD/PCM, $\varepsilon = 18.5$) was used for optimization due to the considerable computational cost of optimizing large clusters. We found (see Supplementary Information) that each EC stabilized Li^+ by ~ 0.7 eV.

During reaction network construction, we consider two free energies for each specie: one uncorrected, and one solvent-corrected. The uncorrected free energy is taken directly from LIBE. For the solvent-corrected free energy, we count the total number of coordinate bonds to all Li^+ ions (see the Supplementary Information for a description of the method for deducing metal coordination) and compare this to the maximum expected number of coordinate bonds (assuming that each Li^+ would prefer to be coordinated by four neighbors).[43] If any Li^+ are undercoordinated, then the free energy is lowered by 0.68 eV for each "missing" coordinate bond. When calculating redox free energies, the uncorrected free energy is used; otherwise, the corrected free energy is used (see the Supplementary Information).

Species Filtering

In the HiPRGen package (see Code Availability), we implement a number of filters that remove undesirable species. These filters take as input an object containing information about a molecule, including its species, coordinates, charge, spin multiplicity, partial charges, connectivity, and thermodynamics. Each filter, based on this information, can discard the molecule or pass it onto the next filter. For terminal filters, if the molecule passes, then it is included in the final filtered set $\mathbf{S}_{filtered}$. For this work, the following molecules were filtered out:

- Molecules composed of two or more disconnected fragments
- Metal-centric complexes, where two or more non-metal fragments are connected only by coordinate bonds to Li^+
- Molecules containing neutral or negative metal ions, where

the charges are calculating by applying the Natural Bonding Orbital (NBO) program version 5.0[44,45] to a single-point energy calculation at the ωB97X-V/def2-TZVPPD/SMD level of theory in Q-Chem.

- Molecules with charge less than -1 or greater than 1

In addition to these filters, which define types of molecules to be excluded from the final network, we further reduce the molecules in the network by removing redundant species. In LIBE, all molecules are unique based on the combination of their charge, spin multiplicity, and molecular connectivity. This means that there could be several molecules that differ only by spin multiplicity, or that differ only by the coordination environment of Li^+ ions (what we call "coordimers"). When this occurs (when there are multiple molecules with the same *covalent* connectivity and charge but potentially with different coordination environments or spin multiplicities), we include only that specie with the lowest solvation-corrected free energy in the final filtered set of species $\mathbf{S}_{filtered}$.

These filters are explained further in the Supplementary Information. We emphasize that these filters are particular to the chemistry being studied in this work, but that HiPRGen has been engineered to enable straightforward addition, removal, or modification of filters in order to be easily applied across diverse chemical applications.

Reaction Filtering

Reaction filters take as input a reaction, defined by a collection of reactants and a collection of products, and either discard the reaction or pass it onto the next filter until a terminal filter is reached. The following types of reactions were filtered out in this work:

- Endergonic reactions with $\Delta G > 0$eV. The reaction free energies for non-redox reactions were taken as $\Delta G = \sum_{i=1}^{m} G^C_{product,i} - \sum_{j=1}^{n} G^C_{reactant,j}$, where G^C is the solvation-corrected free energy, m is the length of the product collection, and n is the length of the reactant collection. For redox reactions, $\Delta G = G^0_{product} - G^0_{reactant} + \Delta q(G_e)$, where G^0 is the uncorrected free energy, Δq is the change in charge of the reaction, and G_e is the free energy of the electron (for +0.0 V vs. Li/Li^+, $G_e = -1.4$eV; for +0.5 V vs. Li/Li^+, $G_e = -1.9$eV).
- Reactions involving a charge change $|\Delta q| > 1$
- Redox reactions involving more than one reactant or more than one product
- Unimolecular dissociative redox reactions in which $|\Delta q| > 0$ and covalent bonds form or break
- Reactions involving more than two reactants or products (this is actually enforced by the species bucketing procedure of HiPRGen and is not a separate filter)
- Reactions involving spectators (species that do not change as a result of the reaction) e.g. A + B to A + C.
- Rections involving more than two bond changes
- Reactions in which two bonds form simultaneously or two bonds break simultaneously
- Reactions in which covalent bonds change and metal ions coordinate/decoordinate (note that reactions in which metal ions remain coordinated but change their coordinate bonds are allowed)

Motivations for each of these filters, along with examples, are provided in the Supplementary Information. Like species filters, the reaction filters can be easily modified and extended by end users to suit a broad range of chemical applications. Of course, removing filters will yield a larger final collection of reactions. We note that for the size of the species collection presented in this work, some filters are necessary to obtain a tractable number of reactions in the final collection.

Monte Carlo Methods

We developed a high-performance implementation of Gillespie's direct method[24], with dependency graph and logarithmically scaling sampler optimizations[25], which we call Reaction Network Monte Carlo (RNMC). RNMC is heavily based on the Stochastic Parallel Particle Kinetic Simulator (SPPARKS) package[46,47] but with modifications to allow simulating networks with hundreds of millions of reactions and thousands of species. RNMC shares the reaction network and dependency graph between all running simulators and uses a lockless data structure for the dependency graph that allows it to be computed dynamically by all of the simulators in parallel.

Using RNMC, we performed 100,000 simulations under each of the four chosen conditions (+0.0V without CO_2, +0.0V with CO_2, +0.5V without CO_2, and +0.5V with CO_2). For simulations without CO_2, the initial state consisted of 30 Li^+ and 30 EC; for those with CO_2, the initial state also included 30 CO_2. Because all reactions were exergonic and no energy barriers were considered, all rate coefficients were constant and equal. Each simulation was conducted to "completion" - that is, until there were no further reactions available for further simulation. Due to the relatively small number of initial species, most simulations took between roughly 200 and 500 steps. We note that simulating to completion - especially with so few simulation steps - is only possible because the system contains only exergonic reactions and therefore contains no loops.

Identification of Thermodynamic Reaction Pathways

To identify a single reaction pathway to a specie of interest, we trace through an individual Monte Carlo trajectory. If the specie of interest is formed in that trajectory, then we trace back the series of reactions leading to the first formation of that specie (see Supplementary Information for an illustration of this process). For instance, if we are searching for pathways to specie X, we might find that it is first formed by the reaction $V + W \rightarrow X$. We then look for the first reaction(s) forming V and W, and then for the first reaction(s) forming the reactants of those formation

reactions, until all reactions can occur from only starting species of the simulation. The series of reactions obtained in this way define a reaction pathway to X.

In general, we are not interested in a single reaction pathway but rather the myriad pathways to the specie of interest. Therefore, for each specie of interest, we repeat the pathway identification procedure above for each trajectory, collecting all unique pathways. We then rank these pathways by some cost function. Here, the cost Φ of a given reaction is defined as $\Phi = exp(\Delta G/k_B T) + 1$, where ΔG is the reaction free energy (uncorrected for a redox reaction, and solvation-corrected otherwise).[23] The total cost of a reaction pathway is the sum of the costs of the individual reactions. We note that, because all reactions included in our network are exergonic, the constant term tends to dominate, though this cost function retains a preference for highly exergonic reactions over those that are only slightly exergonic.

Identification of Network Products

After all simulations have completed, the resulting trajectories are analyzed to determine product species. Products are defined by three criteria: the ratio of formation and consumption, relative accumulation, and availability of low-cost pathways.

To determine the ratio of formation and consumption, each trajectory was interrogated to find all reactions involving each specie. If a given specie is a reactant of an identified reaction, then that means it was consumed; if it is a product of the reaction, then that means it was formed. If the ratio of the total number of instances of formation across all trajectories to the total number of instances of consumption across all trajectories is greater than some threshold (here chosen as 1.5, meaning that three of the species are produced for every two consumed), then it could be a network product.

For relative accumulation, we take the average of all trajectories. The expected value of a specie is the average of the final state - how many of the molecule will persist once the average simulation has completed. If this expected value is greater than some threshold (here 0.1, meaning that one of this specie is produced and is present in the final state for every ten simulations), then that specie could be a product.

Finally, for those species with formation/consumption ratios and expected values that pass the chosen thresholds, we perform thermodynamic pathfinding analysis. If the pathway with the lowest cost has a cost less than some threshold (here 10.0), then we consider the specie to be a product of the network.

The species reported in Figure 4 are network products in at least one - but not necessarily all - of the four conditions considered (see Supplementary Information for details). We note that we add one additional constraint to the products reported here: spin multiplicity. While open-shell species can be products of the network (see Supplementary Information for examples), they are highly unlikely to be stable or meta-stable (long-lived radicals are generally rare). In the hopes of extracting useful chemical insights from network products, we therefore only consider network products that are singlets.

Kinetic Refinement of Reaction Mechanisms

The thermodynamic reaction pathways obtained via stochastic analysis were interrogated to determine the actual elementary steps. For the network products considered here (for instance bi-dioxolylidene), several low-cost thermodynamic reaction pathways were selected. For each elementary step along these pathways - excluding coordination reactions and redox reactions - we attempted to locate the TS using the AutoTS workflow,[48] an end-to-end workflow to identify TS and reaction pathways that is built on top of the Jaguar electronic structure code (version 11.2).[49] All AutoTS calculations were conducted at a ωB97X-D/def2-SVPD(-f)/PCM level of theory,[39,50,51] with water as the solvent. In some cases, for reactions involving two bonds changing, AutoTS identified two TS (for instance, one to form a bond and one to break a bond); these were optimized separately.

In cases where AutoTS was unable to find a TS for a given reaction, we searched using the single-ended growing string method (SE-GSM), as implemented in the pyGSM code.[52] SE-GSM calculations were conducted with a Q-Chem backend (version 5.3.2) at the ωB97X-D/def2-SVPD/PCM level of theory.[53] To be as consistent as possible, TS found using SE-GSM in Q-Chem were re-optimized in Jaguar at the ωB97X-D/def2-SVPD(-f)/PCM level of theory.

For each TS, we confirmed that the optimized structure possessed one imaginary frequency and confirmed that it connected the expected endpoints. For cases where the endpoints consist of two molecules that are not covalently bound (typically bound only by coordination to Li^+), we allow small imaginary frequencies (less than 75i cm^{-1}). These small imaginary modes can prove extremely difficult to remove using conventional geometry optimization methods, especially when they involve the motion of Li^+ ions, and typically do not significantly affect the free energy. We note that in some cases, the barriers that we report are based on the difference between the TS and the reactants or products at infinite separation, rather than the entrance or exit complex. The electronic energies of all optimized structures (TS and endpoints) were corrected using a single-point calculation at a higher level of theory (ωB97X-V/def2-TZVPPD/SMD) in Q-Chem. The SMD parameters used were the same used for the construction of the LIBE dataset.[28] We note that we used Q-Chem for these calculations, rather than Jaguar, because the SMD implicit solvent model is not implemented in Jaguar at the time of this writing.

All AutoTS and pyGSM calculations were automated using workflows that we have implemented in the MPcat code (see Code Availability). These workflows are designed for high-throughput transition-state searches and reaction pathway analysis. Note that we use a fork of the original pyGSM code for SE-GSM. All data used to construct mechanisms (molecular structures, thermodynamics, vibrational frequencies, and frequency modes) are provided in the Javascript Object Notation (JSON) format in the file "reaction_pathways.json".

Code Availability

All codes discussed here (HiPRGen, RNMC, MPcat, and pyGSM) are released open source on Github. A Python implementation of the HiPRGen method can be found at `https:`

//github.com/BlauGroup/HiPRGen. RNMC, a performant kinetic Monte Carlo code in C++ and based on SPPARKS, can be found at https://github.com/BlauGroup/RNMC. AutoTS and SE-GSM calculations were performed using the automated workflows defined in MPcat, which can be found at https://github.com/espottesmith/MPcat. SE-GSM calculations specifically used a fork of the original pyGSM code, which can be found at https://github.com/espottesmith/pyGSM.

Acknowledgements

This work was supported by both the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231 and the Joint Center for Energy Storage Research, an Energy Innovation Hub funded by the US Department of Energy, Office of Science, Basic Energy Sciences. Computational resources were provided by the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility under Contract No. DE-AC02-05CH11231, and by the Department of Energy's Office of Energy Efficiency and Renewable Energy located at the National Renewable Energy Laboratory. This research also used the Lawrencium computational cluster resource provided by the IT Division at the Lawrence Berkeley National Laboratory (Supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231). Schrödinger, Inc. provided access to Jaguar and AutoTS software, as well as technical advice regarding their use.

Author Contributions

D.B. and S.M.B. conceived of the study and approach; D.B. implemented HiPRGen, RNMC and stochastic analysis methods with feedback from S.M.B.; D.B., E.W.C.S.-S., S.D., and S.M.B. designed species and reaction filters; D.B., S.D., and S.M.B. designed the methods for network product identification; E.W.C.S.-S. and N.S.R. performed quantum chemical calculations; E.W.C.S.-S. and S.M.B. analyzed data; K.A.P. and S.M.B. secured funding; E.W.C.S.-S. wrote the original manuscript; all authors edited the manuscript.

Notes and references

1 L. Wang, A. Menakath, F. Han, Y. Wang, P. Y. Zavalij, K. J. Gaskell, O. Borodin, D. Iuga, S. P. Brown, C. Wang, K. Xu and B. W. Eichhorn, *Nature Chemistry*, 2019, **11**, 789–796.

2 X. Xie, E. W. Clark Spotte-Smith, M. Wen, H. D. Patel, S. M. Blau and K. A. Persson, *J. Am. Chem. Soc.*, 2021, **143**, 13245–13258.

3 G. Isaacman-VanWertz, P. Massoli, R. O'Brien, C. Lim, J. P. Franklin, J. A. Moss, J. F. Hunter, J. B. Nowak, M. R. Canagaratna, P. K. Misztal, C. Arata, J. R. Roscioli, S. T. Herndon, T. B. Onasch, A. T. Lambe, J. T. Jayne, L. Su, D. A. Knopf, A. H. Goldstein, D. R. Worsnop and J. H. Kroll, *Nat. Chem*, 2018, **10**, 462–468.

4 S. Back, K. Tran and Z. W. Ulissi, *ACS Catalysis*, 2019, **9**, 7651–7659.

5 Y. Y. Birdja, E. Pérez-Gallent, M. C. Figueiredo, A. J. Göttle, F. Calle-Vallejo and M. T. M. Koper, *Nat Energy*, 2019, **4**, 732–745.

6 A. L. Dewyer, A. J. Argüelles and P. M. Zimmerman, *WIREs Computational Molecular Science*, 2018, **8**, e1354.

7 J. P. Unsleber and M. Reiher, *Annual Review of Physical Chemistry*, 2020, **71**, 121–142.

8 J. S. Tse, *Annual Review of Physical Chemistry*, 2002, **53**, 249–290.

9 S. Maeda, Y. Harabuchi, M. Takagi, T. Taketsugu and K. Morokuma, *The Chemical Record*, 2016, **16**, 2232–2248.

10 C. Shang and Z.-P. Liu, *J. Chem. Theory Comput.*, 2013, **9**, 1838–1845.

11 C. W. Gao, J. W. Allen, W. H. Green and R. H. West, *Computer Physics Communications*, 2016, **203**, 212–225.

12 C. F. Goldsmith and R. H. West, *J. Phys. Chem. C*, 2017, **121**, 9970–9981.

13 M. Liu, A. Grinberg Dana, M. S. Johnson, M. J. Goldman, A. Jocher, A. M. Payne, C. A. Grambow, K. Han, N. W. Yee, E. J. Mazeau, K. Blondal, R. H. West, C. F. Goldsmith and W. H. Green, *J. Chem. Inf. Model.*, 2021, **61**, 2686–2696.

14 D. Rappoport and A. Aspuru-Guzik, *J. Chem. Theory Comput.*, 2019, **15**, 4099–4112.

15 A. Wołos, R. Roszak, A. Żądło Dobrowolska, W. Beker, B. Mikulak-Klucznik, G. Spólnik, M. Dygas, S. Szymkuć and B. A. Grzybowski, *Science*, 2020, **369**, 1–12.

16 Y. Kim, J. W. Kim, Z. Kim and W. Y. Kim, *Chem. Sci.*, 2018, **9**, 825–835.

17 X. Jia, A. Lynch, Y. Huang, M. Danielson, I. Lang'at, A. Milder, A. E. Ruby, H. Wang, S. A. Friedler, A. J. Norquist and J. Schrier, *Nature*, 2019, **573**, 251–255.

18 F. Calle-Vallejo and M. T. M. Koper, *Electrochimica Acta*, 2012, **84**, 3–11.

19 S. Maeda and K. Morokuma, *J. Chem. Theory Comput.*, 2012, **8**, 380–385.

20 Q. Zhao and B. M. Savoie, *Nat Comput Sci*, 2021, **1**, 479–490.

21 C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.

22 S. M. Blau, E. W. C. Spotte-Smith, B. Wood, S. Dwaraknath and K. A. Persson, *ChemRxiv*, 2020.

23 S. M. Blau, H. D. Patel, E. W. Clark Spotte-Smith, X. Xie, S. Dwaraknath and K. A. Persson, *Chemical Science*, 2021, **12**, 4931–4939.

24 D. T. Gillespie, *Journal of Computational Physics*, 1976, **22**, 403–434.

25 M. A. Gibson and J. Bruck, *J. Phys. Chem. A*, 2000, **104**, 1876–1889.

26 J. C. Baez, *Advances in Mathematical Physics*, 2018, **2018**, e7676309.

27 J. C. Baez and J. Biamonte, *arXiv:1209.3632 [math-ph, physics:quant-ph]*, 2019.

28 E. W. C. Spotte-Smith, S. M. Blau, X. Xie, H. D. Patel, M. Wen, B. Wood, S. Dwaraknath and K. A. Persson, *Sci Data*, 2021, **8**, 203.

29 B. Rowden and N. Garcia-Araez, *Energy Reports*, 2020, **6**, 10–18.

30 P. Verma, P. Maire and P. Novák, *Electrochimica Acta*, 2010, **55**, 6332–6341.

31 B. L. D. Rinkel, D. S. Hall, I. Temprano and C. P. Grey, *J. Am. Chem. Soc.*, 2020, 22.

32 S. J. An, J. Li, C. Daniel, D. Mohanty, S. Nagpure and D. L. Wood, *Carbon*, 2016, **105**, 52–76.

33 R. Mogi, M. Inaba, Y. Iriyama, T. Abe and Z. Ogumi, *Journal of Power Sources*, 2003, **119-121**, 597–603.

34 J. Nanda, G. Yang, T. Hou, D. N. Voylov, X. Li, R. E. Ruther, M. Naguib, K. Persson, G. M. Veith and A. P. Sokolov, *Joule*, 2019, **3**, 2001–2019.

35 S. Tsubouchi, Y. Domi, T. Doi, M. Ochida, H. Nakagawa, T. Yamanaka, T. Abe and Z. Ogumi, *Journal of The Electrochemical Society*, 2012, **159**, A1786–A1790.

36 M. Nie, D. P. Abraham, D. M. Seo, Y. Chen, A. Bose and B. L. Lucht, *J. Phys. Chem. C*, 2013, **117**, 25381–25389.

37 M. Nie, J. Demeaux, B. T. Young, D. R. Heskett, Y. Chen, A. Bose, J. C. Woicik and B. L. Lucht, *J. Electrochem. Soc.*, 2015, **162**, A7008.

38 A. V. Marenich, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. B*, 2009, **113**, 6378–6396.

39 D. Rappoport and F. Furche, *J. Chem. Phys.*, 2010, **133**, 134105.

40 N. Mardirossian and M. Head-Gordon, *Physical Chemistry Chemical Physics*, 2014, **16**, 9904–9924.

41 A. V. Marenich, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. B*, 2009, **113**, 4538–4543.

42 J. J. Guerard and J. S. Arey, *J. Chem. Theory Comput.*, 2013, **9**, 5046–5058.

43 M. I. Chaudhari, J. R. Nair, L. R. Pratt, F. A. Soto, P. B. Balbuena and S. B. Rempe, *J. Chem. Theory Comput.*, 2016, **12**, 5709–5718.

44 E. Glendening, J. Badenhoop, A. Reed, J. Carpenter, J. Bohmann, C. Morales and F. Weinhold, *Theoretical Chemistry Institute, University of Wisconsin: Madison, WI, USA*, 2001.

45 F. Weinhold, C. Landis and E. Glendening, *International Reviews in Physical Chemistry*, 2016, **35**, 399–440.

46 S. Plimpton, A. Thompson and A. Slepoy, *Stochastic Parallel PARticle Kinetic Simulator*, Sandia National Lab. (SNL-NM), Albuquerque, NM (United States) Technical Report SPPARKS, 2008.

47 S. Plimpton, C. Battaile, M. Ch, L. Holm, A. Thompson, V. Tikare, G. Wagner, X. Zhou, C. G. Cardona and A. Slepoy, *Crossing the Mesoscale No-Man's Land via Parallel Kinetic Monte Carlo*, 2009.

48 L. D. Jacobson, A. D. Bochevarov, M. A. Watson, T. F. Hughes, D. Rinaldo, S. Ehrlich, T. B. Steinbrecher, S. Vaitheeswaran, D. M. Philipp, M. D. Halls and R. A. Friesner, *J. Chem. Theory Comput.*, 2017, **13**, 5780–5797.

49 A. D. Bochevarov, E. Harder, T. F. Hughes, J. R. Greenwood, D. A. Braden, D. M. Philipp, D. Rinaldo, M. D. Halls, J. Zhang and R. A. Friesner, *International Journal of Quantum Chemistry*, 2013, **113**, 2110–2142.

50 J.-D. Chai and M. Head-Gordon, *Physical Chemistry Chemical Physics*, 2008, **10**, 6615–6620.

51 B. Mennucci, *WIREs Computational Molecular Science*, 2012, **2**, 386–404.

52 C. Aldaz, J. A. Kammeraad and P. M. Zimmerman, *Physical Chemistry Chemical Physics*, 2018, **20**, 27394–27405.

53 E. Epifanovsky, A. T. B. Gilbert, X. Feng, J. Lee, Y. Mao, N. Mardirossian, P. Pokhilko, A. F. White, M. P. Coons, A. L. Dempwolff, Z. Gan, D. Hait, P. R. Horn, L. D. Jacobson, I. Kaliman, J. Kussmann, A. W. Lange, K. U. Lao, D. S. Levine, J. Liu, S. C. McKenzie, A. F. Morrison, K. D. Nanda, F. Plasser, D. R. Rehn, M. L. Vidal, Z.-Q. You, Y. Zhu, B. Alam, B. J. Albrecht, A. Aldossary, E. Alguire, J. H. Andersen, V. Athavale, D. Barton, K. Begam, A. Behn, N. Bellonzi, Y. A. Bernard, E. J. Berquist, H. G. A. Burton, A. Carreras, K. Carter-Fenk, R. Chakraborty, A. D. Chien, K. D. Closser, V. Cofer-Shabica, S. Dasgupta, M. de Wergifosse, J. Deng, M. Diedenhofen, H. Do, S. Ehlert, P.-T. Fang, S. Fatehi, Q. Feng, T. Friedhoff, J. Gayvert, Q. Ge, G. Gidofalvi, M. Goldey, J. Gomes, C. E. González-Espinoza, S. Gulania, A. O. Gunina, M. W. D. Hanson-Heine, P. H. P. Harbach, A. Hauser, M. F. Herbst, M. Hernández Vera, M. Hodecker, Z. C. Holden, S. Houck, X. Huang, K. Hui, B. C. Huynh, M. Ivanov, A. Jasz, H. Ji, H. Jiang, B. Kaduk, S. Kähler, K. Khistyaev, J. Kim, G. Kis, P. Klunzinger, Z. Koczor-Benda, J. H. Koh, D. Kosenkov, L. Koulias, T. Kowalczyk, C. M. Krauter, K. Kue, A. Kunitsa, T. Kus, I. Ladjanszki, A. Landau, K. V. Lawler, D. Lefrancois, S. Lehtola, R. R. Li, Y.-P. Li, J. Liang, M. Liebenthal, H.-H. Lin, Y.-S. Lin, F. Liu, K.-Y. Liu, M. Loipersberger, A. Luenser, A. Manjanath, P. Manohar, E. Mansoor, S. F. Manzer, S.-P. Mao, A. V. Marenich, T. Markovich, S. Mason, S. A. Maurer, P. F. McLaughlin, M. F. S. J. Menger, J.-M. Mewes, S. A. Mewes, P. Morgante, J. W. Mullinax, K. J. Oosterbaan, G. Paran, A. C. Paul, S. K. Paul, F. Pavošević, Z. Pei, S. Prager, E. I. Proynov, A. Rak, E. Ramos-Cordoba, B. Rana, A. E. Rask, A. Rettig, R. M. Richard, F. Rob, E. Rossomme,

T. Scheele, M. Scheurer, M. Schneider, N. Sergueev, S. M. Sharada, W. Skomorowski, D. W. Small, C. J. Stein, Y.-C. Su, E. J. Sundstrom, Z. Tao, J. Thirman, G. J. Tornai, T. Tsuchimochi, N. M. Tubman, S. P. Veccham, O. Vydrov, J. Wenzel, J. Witte, A. Yamada, K. Yao, S. Yeganeh, S. R. Yost, A. Zech, I. Y. Zhang, X. Zhang, Y. Zhang, D. Zuev, A. Aspuru-Guzik, A. T. Bell, N. A. Besley, K. B. Bravaya, B. R. Brooks, D. Casanova, J.-D. Chai, S. Coriani, C. J. Cramer, G. Cserey, A. E. DePrince, R. A. DiStasio, A. Dreuw, B. D. Dunietz, T. R. Furlani, W. A. Goddard, S. Hammes-Schiffer, T. Head-Gordon, W. J. Hehre, C.-P. Hsu, T.-C. Jagau, Y. Jung, A. Klamt, J. Kong, D. S. Lambrecht, W. Liang, N. J. Mayhall, C. W. McCurdy, J. B. Neaton, C. Ochsenfeld, J. A. Parkhill, R. Peverati, V. A. Rassolov, Y. Shao, L. V. Slipchenko, T. Stauch, R. P. Steele, J. E. Subotnik, A. J. W. Thom, A. Tkatchenko, D. G. Truhlar, T. Van Voorhis, T. A. Wesolowski, K. B. Whaley, H. L. Woodcock, P. M. Zimmerman, S. Faraji, P. M. W. Gill, M. Head-Gordon, J. M. Herbert and A. I. Krylov, *J. Chem. Phys.*, 2021, **155**, 084801.