nature computational science

ARTICLES

https://doi.org/10.1038/s43588-021-00101-3

Check for updates

# Simultaneously improving reaction coverage and computational cost in automated reaction prediction tasks

Qiyuan Zhao and Brett M. Savoie ✉

**Automated reaction prediction has the potential to elucidate complex reaction networks for applications ranging from combustion to materials degradation, but computational cost and inconsistent reaction coverage are still obstacles to exploring deep reaction networks. Here we show that cost can be reduced and reaction coverage can be increased simultaneously by relatively straightforward modifications of the reaction enumeration, geometry initialization and transition state convergence algorithms that are common to many prediction methodologies. These components are implemented in the context of yet another reaction program (YARP), our reaction prediction package with which we report reaction discovery benchmarks for organic single-step reactions, thermal degradation of a $\gamma$-ketohydroperoxide, and competing ring-closures in a large organic molecule. Compared with recent benchmarks, YARP (re)discovers both established and unreported reaction pathways and products while simultaneously reducing the cost of reaction characterization by nearly 100-fold and increasing convergence of transition states. This combination of ultra-low cost and high reaction coverage creates opportunities to explore the reactivity of larger systems and more complex reaction networks for applications such as chemical degradation, where computational cost is a bottleneck.**

The reaction network prediction problem consists of predicting the transition states, kinetically relevant intermediates and products for a set of reactants. Decades of research has been devoted to this topic for specific applications, ranging from the evaluation of combustion pathways[1,2], cellular metabolism[3–5] and atmospheric chemistry[6,7], to the related inverse problem of retrosynthetic organic reaction planning (that is, generating a reaction network in reverse)[8–10]. Although the details are specific to each application, the problem common to all is resolving which reactions happen and when as a function of relevant environmental variables (for example, temperature, pressure, concentrations, reagents, phase and so on). For applications for which sufficient domain knowledge of plausible reactions exists, workable solutions have been developed to algorithmically generate reaction networks that are then refined with feedback from experiments or further computational characterizations[11–13]. More recently, machine learning has also enabled major advances in data-rich reaction problems, with demonstrations of models capable of predicting retrosynthetic pathways for complex organic molecules that are competitive with the best expert systems and chemical intuition[14,15]. However, for problems where an established set of reactions and reaction data do not exist, a fundamentally different approach is required to predict reaction networks.

In recent years, several groups have recognized the opportunity to automatically elucidate reaction networks by leveraging now mature quantum-chemistry-based transition state characterizations[16–20] and modern computational throughput. As has been recently reviewed in several places[12,21–23], these approaches can roughly be categorized on the basis of whether they utilize single-ended transition state searches (for example, anharmonic downward distortion following[24], artificial force induced reaction[25,26], stochastic surface walking method[27]), double-ended transition state searches (for example, ZStruct[28] and the methods in refs. [29,30]), or reaction templates (for example, NetGen[31], RMG[32], KinBot[33]) to drive reaction discovery, with each coupled to a suitably accurate model chemistry to locate transition states and establish the thermodynamics of products. Although still in a relatively early phase, there are already many demonstrations of such algorithms automatically recapitulating established reaction mechanisms in benchmark systems, as well as authentic predictions of reaction pathways that were previously undocumented[34–39]. Despite this success, computational cost still represents a serious impediment to characterizing complex reaction networks involving large numbers of atoms or reactions occurring in condensed phases. In particular, the underlying reaction exploration algorithm must be sufficiently general such that all kinetically relevant pathways are identified and characterized by quantum chemistry. As such, reaction exploration cannot be entirely naive, as the range of possible bond rearrangements grows roughly factorially with the number of atoms in the reactants and would thus overwhelm even the most economical model chemistry[13]. On the other hand, it has been observed that the number of kinetically relevant reactions grows linearly with respect to number of intermediates for established reaction networks across many domains[40,41]. This implies that the kinetically relevant reactions in a typical network are computationally tractable to characterize, although at present there is no a priori method for robustly distinguishing physical from unphysical reactions besides subjecting them to costly quantum chemistry calculations or relying on established reactivities (that is, the very thing that is missing in many motivating applications for computational prediction).

In this work we introduce several strategies to improve the accuracy and size of reaction networks that can be computationally characterized while limiting the use of domain heuristics. First we employ the recently developed semi-empirical geometry, frequency, non-covalent, extended tight binding (GFN2-xTB) model chemistry developed by Grimme and colleagues[42,43] to preoptimize reaction pathways before more expensive density functional theory (DFT) characterizations. Second, we implement generic

Davidson School of Chemical Engineering, Purdue University, West Lafayette, IN, USA. ✉e-mail: bsavoie@purdue.edu

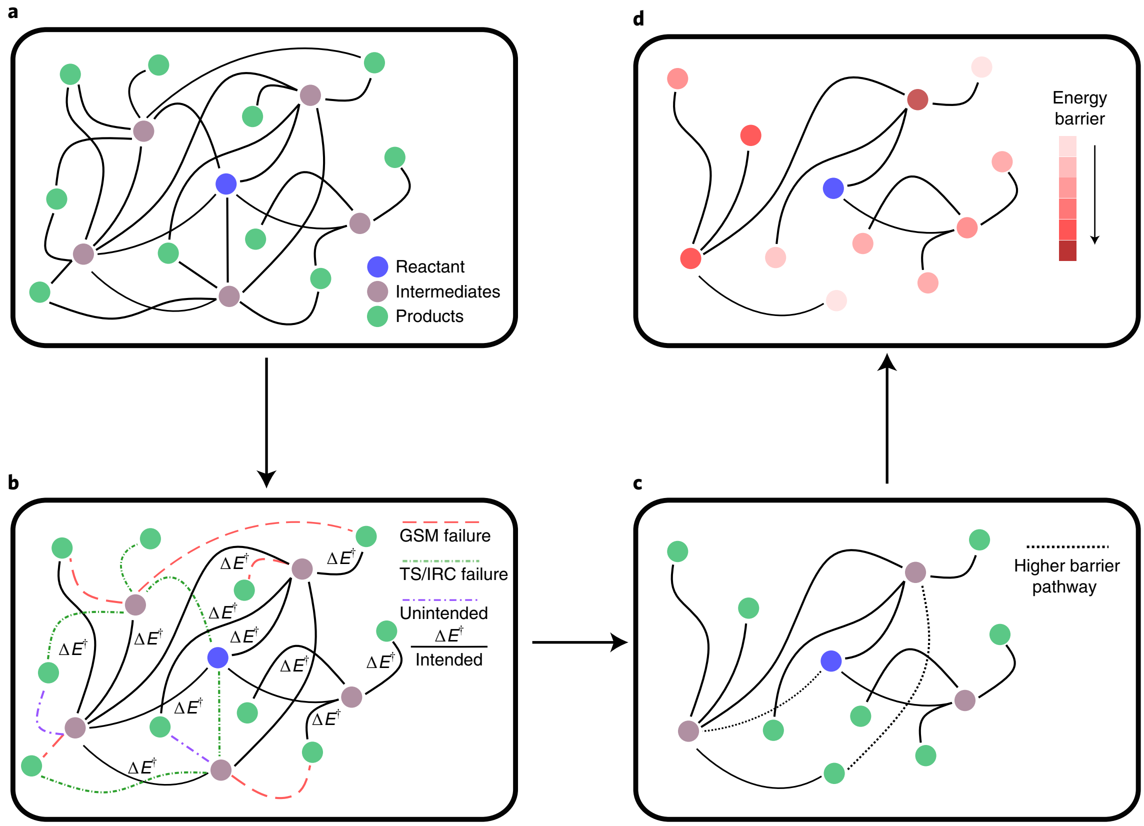


**Fig. 1 | Overview of the YARP methodology.** **a**, Potential products are enumerated using ERSs. **b**, Reaction pathways are constructed by applying GSM at the GFN2-xTB level, Berny optimization and an IRC calculation for validation. **c**, Removal of unphysical reactions based on transition state failures or unintended transition states. **d**, Reaction network analysis is performed to identify the lowest activation energy pathway and determine the thermodynamic and kinetic relevance of each product.

elementary reaction steps (ERSs) that enable reaction exploration while also being well-conditioned for transition state convergence. Third, we establish a robust initial structure optimization procedure based on jointly optimizing the reactant and product geometries before the transition state search. These features are implemented in the context of a recursive reaction enumeration and double-ended transition state search algorithm, which we call Yet Another Reaction Program (YARP). To evaluate the performance of YARP with respect to predicting diverse chemical reactions, we have benchmarked its performance on predicting single-step organic reactions[20], thermal degradation of a $\gamma$-ketohydroperoxide[30,44] and competing Diels–Alder ring-closures in a large organic molecule[39]. In all tasks, we find that YARP recapitulates past reactions, discovers many unknown kinetically relevant pathways and exhibits orders-of-magnitude reduced computational cost.

## Results

**YARP framework.** YARP consists of three stages for automatically characterizing reaction networks: graph-based product enumeration, reaction network construction and reaction network analysis (Fig. 1). In brief, potential products are recursively generated for user-supplied reactants using an automated reaction enumeration scheme and a molecular graph formalism. YARP uses the concept of an ERS—defined as changes to the bond/electron state of the reactants that result in a product with a stable Lewis structure—to enumerate potential products. For full-octet neutral organic systems, the simplest reaction step that yields non-trivial products is breaking two bonds and forming two bonds (b2f2; Extended Data Fig. 1). The extent to which more complex ERSs (for example, b3f3) are necessary is discussed in the context of specific systems below. After ERS-based product enumeration, an algorithm is applied to jointly optimize three-dimensional geometries of the product and reactant states to promote transition state convergence. These geometries are then used to seed a transition state search using the growing string method (GSM) coupled with the GFN2-xTB semi-empirical model chemistry. All reactions generated in this way are subsequently characterized using Berny optimization to locate transitions states at the DFT level and intrinsic reaction coordinate (IRC) calculations to validate that the identified transition states correspond with the putative reaction. At each iteration, this procedure yields a range of products, a subset or all of which can be used as inputs for additional reaction prediction. Detailed descriptions of each stage are provided in the Methods.

Three classes of reaction prediction problems were used to benchmark YARP: the first was predicting a set of single-step organic reactions curated by Zimmerman[20]; the second was to characterize the thermal decomposition network of 3-hydroperoxypropanal, recently benchmarked by Grambow and colleagues[30]; and the third was a case study of applying YARP to compare competing Diels–Alder ring closures in a large ketothioester previously studied by Yang and co-workers[39]. The number of DFT gradient calls is reported to represent the computational cost (Extended Data Fig. 2 and Supplementary Section 1). Statistics associated with the success rate of transition state convergence and whether transition states correspond to intended channels are reported with the following definitions. The success rate corresponds to the fraction of unique reactions that completed all calculations (that is, GSM, Berny and IRC) compared with the total number of unique reactions attempted. The intended rate corresponds to the fraction of unique reactions that completed all calculations and exhibit an intended transition state (that is, based on the IRC results) compared with the total number of unique reactions attempted. Detailed descriptions of the benchmarks and performance statistics are provided in the Methods.

**Single-step reaction prediction benchmark.** The trade-off between reaction exploration and computational cost represents a major limitation on the breadth and depth of reaction networks that can be discovered by automated algorithms. The b2f2 ERS yields a tractable number of reactions that we reasoned would be able to recapitulate typical closed-shell organic reactions without being biased solely towards known reactivities. Nevertheless, limiting the number of ERSs creates the possibility of missing important reaction pathways. To investigate this trade-off, we used YARP to comprehensively characterize all b2f2 reactions for the reactants in the Zimmerman dataset[20] and compared the established reactions with the lowest-barrier channels predicted by YARP (Fig. 2).

We note the small number of DFT-level gradient calls required by YARP to identify the transition states of the attempted reactions (Fig. 2a). The range of gradient calls is from 4 to 50 with an average of ~13 per reaction, where around 75% of reactions needed less than 20 gradient calls. By comparison, Zimmerman reported on average 468 gradient calls for his most robust set of GSM hyperparameters (that is, 11 images with climbing image optimization and an overlap criterion) and ~380 gradient calls on average when using nine images as employed here[20]. Although the reference datasets have minor differences as described in the 'Benchmark Systems' section, YARP clearly exhibits a qualitative reduction in computational cost compared with direct GSM localization at the DFT level.

A dramatic reduction in gradient calls is moot if YARP cannot also successfully localize transition states. We thus evaluated the transition state success and intended rates averaged across all unique reactions involving each reactant (Fig. 2b). The success rate varies between 75% and 100% with an overall average of 86.2% for all 20 reactants, whereas the intended rate varies between 33.3% and 100%, with an average of 57.9%. The comparison between the success rate and the intended rate illustrates that the latter depends on reactant complexity, whereas the former is consistently high. In particular, the number of potential products and the dimensionality of the potential energy surface increase with reactant size and complexity, which also increases the occurrence of unphysical reactions in unbiased reaction searches. For some simple systems such as **R2**, **R4** and **R10** (reactant labels are shown in Fig. 2c), where the number of heavy atoms are 3, 2 and 3, respectively, both the success rate and intended rate reach 100%; however, for **R7** (a stable four-membered ring) and **R9** (an aromatic compound with ten heavy atoms) the intended rate is only 33.3%, whereas the success rate is still over 80%. We note that even with a comprehensive transition state search, neither the success rate nor the intended rate would necessarily reach 100% when comprehensive reaction enumeration is used. As the transition state algorithm within YARP is playing the role of discriminating between physical and unphysical reactions, it may be that some of the enumerated reactions are poorly conditioned and no physical transition state exists that directly connects those reactants and products. For instance, considering only the subset of reactions that were characterized by both Zimmerman and YARP, YARP exhibits a success rate of 100%, which is the same as in an earlier work. The intended rate was not reported by Zimmerman, but it is similarly high in YARP at 98% for this subset of reactions.

The goal of performing a benchmark that includes a range of organic reactions and functional groups was to establish the extent to which YARP can (re)discover diverse reactions using only the simple b2f2 ERS. To show the kinetically favorable reactions discovered by YARP, the b2f2 reaction with the lowest activation energy for each reactant was compared with the lowest barrier reaction reported by Zimmerman[20] (Fig. 2c). The green region includes 13 reactions that are the lowest energy barrier reactions in both the Zimmerman dataset and the b2f2 enumeration performed by YARP. For the three reactions in the blue region, YARP identified the same reactions as Zimmerman, but also discovered reactions with lower activation energies that were not identified in the previous study. For **R17** and **R18**, in the red and yellow regions, respectively, the lowest activation energy reactions in the Zimmerman dataset are b3f3 reactions that are not discovered by single-step b2f2 enumeration. In particular, R17 represents a Diels–Alder reaction, which is a concerted b3f3 reaction that would not be discovered even by repeated application of the b2f2 ERS. Nevertheless, when running YARP to include b3f3 reactions, the Diels–Alder reaction is discovered as the lowest energy barrier reaction (Extended Data Fig. 3). By contrast, the lowest energy barrier reaction of **R18** predicted by YARP was not previously investigated, which makes it uncertain which reaction is in fact more likely. Furthermore, Zimmerman reported only one reaction pathway for **R19** (dimethylphosphine + ethene) and **R20** (trimethylphosphine + oxirane), and neither is a b2f2 reaction step and thus cannot be compared here. To summarize these comparisons, YARP was able to automatically discover all of the previously reported b2f2 reactions and identify them as important channels among all of the investigated reactions. Furthermore, YARP discovered several lower activation energy reactions that were not previously reported. On the other hand, all reactions that were missed by YARP can be rationalized by the fact that they either required repeated applications of the b2f2 ERS—which was not pursued in this benchmark—or the inclusion of complementary ERSs. In the latter case, we note that such reactions are relatively contextual (for example, atoms capable of expanded octets or the presence of multiple double bonds) and it may be desirable to leverage further elementary chemical information available from the reactant graph when considering whether to apply more complex ERSs.

*Reaction mechanisms discovered by YARP.* One rationale for using simple ERSs such as b2f2 is to better condition the transition state localization by reducing the occurrence of reactions that possess multiple transition states. We note that across all of the b2f2 reactions that were successfully localized in this study, 98.5% exhibit only a single transition state in the minimum energy pathway determined by GSM. Nevertheless, these b2f2 reactions still exhibit a broad range of mechanistic diversity with respect to when bonds break and form relative to the transition state (Fig. 3). Based on the bond-breaking and -forming histograms in Fig. 3a, we observe the intuitive result that bond-breaking events skew earlier than the transition state, whereas bond-forming events skew later than the transition state. Similarly, the aggregated bond-breaking and -forming distributions are centered about the transition state, indicating that the transition state of most reactions consists of at least one change in bond order (that is, more than 80% of bond changes occur within ±20 steps of the transition state). We can further distinguish between several types of sequential reactions and concerted reactions that are predicted by YARP (**I**–**IV** in Fig. 3b). Among the sequential reactions, we observe at least one bond-breaking event before the transition state in all cases, followed by b1f2 type transition states (**I**), b1f1 transition states (**II**) and transition states only associated with conformational rearrangement (**III**). Typical examples of **I** are bimolecular reactions, where the reaction can proceed in mainly concerted fashion after an initial bond-breaking event. Typical examples of **II** are reactions involving ring closures, where the transition state is associated with a partial bond rearrangement, followed by a later ring closure. Typical examples of **III** are reactions that pass through a multimolecular intermediate, where a significant conformational change must occur before executing the bond formation steps. By contrast, we also observe a high occurrence of concerted reactions (**IV**), defined here by reactions where all bond

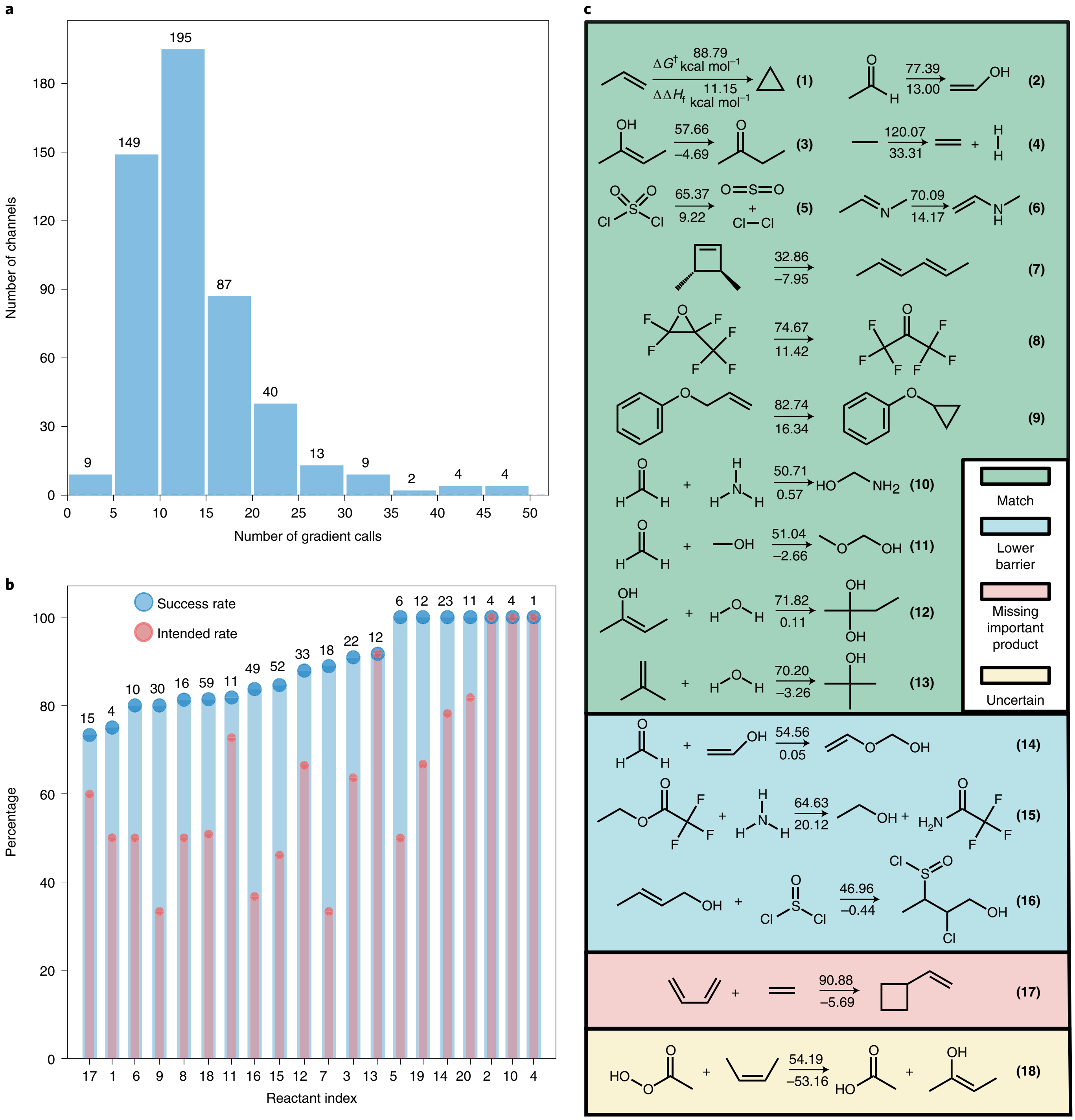


**Fig. 2 | Overview of YARP performance at predicting single-step organic reactions from the Zimmerman dataset. a**, The distribution of the number of gradient calls required to converge each intended reaction, with labels over each bar indicating the number of reactions in each bin. **b**, The success and intended rates of unique reactions involving each reactant. **c**, The minimum activation energy pathways discovered by YARP, classified on the basis of comparisons with Zimmerman's predictions.

rearrangements occur within ± 20 steps of the transition state. For example, in the hydride shift reaction shown in Fig. 3b, all bond changes occur within five steps of the transition state. Although the classifications of specific reactions are subject to how many images are included in the transition state region, these results demonstrate that in addition to being able to discover chemically diverse reactivities using simple ERSs, YARP also discovers a broad range of reaction mechanisms in the predicted pathways. The automatic classification of reaction mechanisms as performed here could also be extended to include more fine-grained mechanistic classification that could in turn inform ERS definitions.

**Unimolecular degradation network of 3-hydroperoxypropanal.** In addition to single-step reaction discovery, automated reaction prediction has potentially the largest relevance to elucidating complex multistep reaction networks. For resolving deep and/or broad

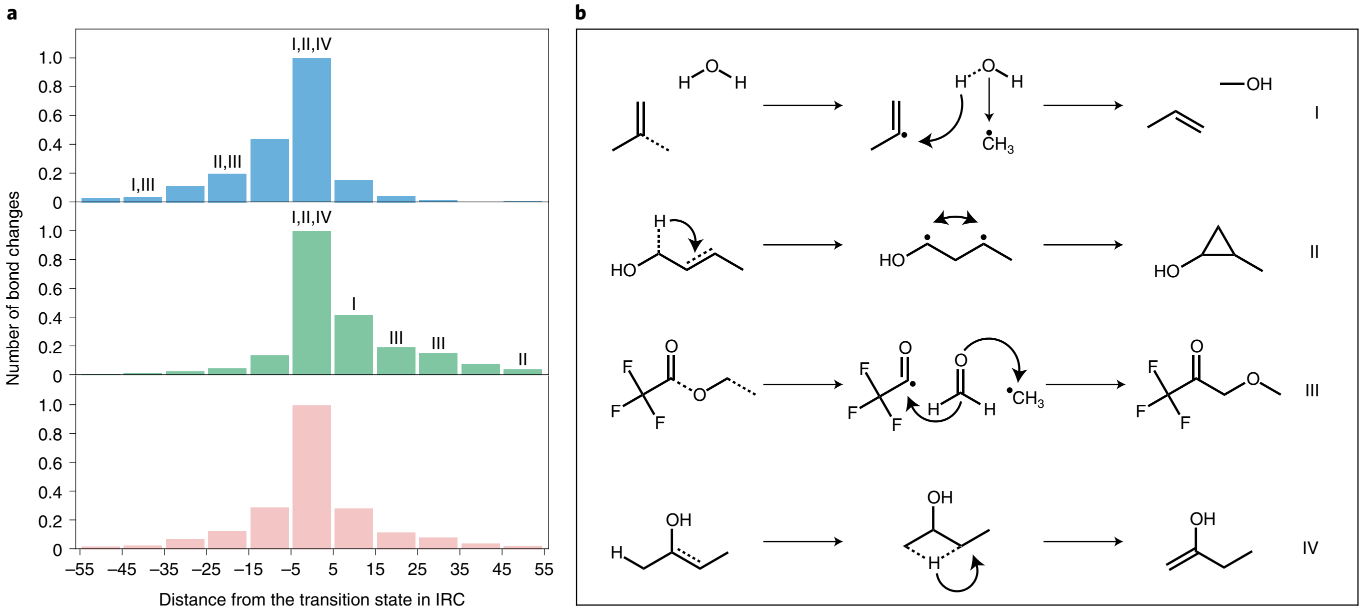


**Fig. 3 | Characterization of sequential and concerted reaction mechanisms discovered by YARP. a**, The histogram of when bond breaking (top), bond forming (middle), or any bond change (bottom), occur relative to the transition state based on analysis of the IRC minimum pathway. **b**, Representative sequential (I,II,III) and concerted reactions (IV) corresponding to the labels in **a**.

networks, reducing the computational cost while still robustly identifying all kinetically relevant pathways remains an outstanding problem within the field. A dearth of benchmark systems for which extensive reaction networks (including transition states, predicted products and gradient call statistics have been reported) also poses a major challenge to unequivocally establishing the performance of competing algorithms. Grambow and colleagues[30] recently published a valuable benchmark study on the unimolecular decomposition of 3-hydroperoxypropanal that provides a comparison of network predictions for five distinct reaction discovery algorithms. 3-Hydroperoxypropanal is a representative $\gamma$-ketohydroperoxide, an important class of molecules for auto-oxidation and preignition chemistries with complex reactions sequences and intermediates that are still under active investigation. In the context of the current work, 3-hydroperoxypropanal served as a useful benchmark system for evaluating the performance of YARP with respect to multistep reaction networks and computational cost.

To provide an illustration of the reaction filtering that is common in network exploration, we first used YARP to recursively enumerate all b2f2 products of 3-hydroperoxypropanal out to two reaction steps. Before transition state characterization, this yields a putative reaction network consisting of 286 vertices (reactants and products) and 1,148 edges (distinct reactions). We then excluded products involving three- and four-membered rings due to the fact that these have a relatively high heat of formation and are unlikely to be kinetically relevant. After removing these compounds, the reaction network consisted of 174 vertices and 539 edges. After performing transition state characterizations, IRC calculations and retaining only products connected by intended channels, the final network was composed of 107 vertices and 173 edges (Fig. 4).

For comparison, in the earlier work by Grambow and colleagues[30], reaction products out to b4f4 type rearrangements were reported, consisting of 562 reaction channels. Of these, 75 intended channels were discovered involving 55 distinct products (aggregated across all five of the reported methods), which can be compared with the YARP results above. As applied in this case, YARP attempted to identify reaction pathways that yielded all degradation products of 3-hydroperoxypropanal up to b4f4-type rearrangements, but excluding compounds containing three-membered rings, four-membered rings or ionic species due to the selected ERS and filtering. That is, YARP was used to characterize a similar number of reactions and was not run to a deeper level of recursion to artificially inflate the number of discovered reactions and products. Differences between the YARP network and the methods benchmarked by Grambow (beyond those noted in the previous sentences) are therefore due to the improved localization of intended reaction channels by YARP.

Figure 4c shows the reaction network predicted by YARP. YARP automatically discovered all previously reported products, minus the small rings and ions excluded by filtering, and managed to connect them by intended channels to the 3-hydroperoxypropanal reactant. Furthermore, YARP discovered 77 products and 157 reactions that have not been previously reported. Constructing this network with YARP required only 8,364 DFT-level gradient calls. The distribution of number of gradient calls per intended channel is shown in Fig. 4a, which illustrates that more than 85% of the intended channels required less than 15 DFT gradient calls. By comparison, directly applying GSM at the DFT level as reported in the earlier benchmark required 756,227 gradient calls to explore 562 reaction channels, representing a nearly 100-fold reduction for YARP. The transition state success and intended rates averaged across all unique reactions involving each reactant (that is, 3-hydroperoxypropanal and all intended b2f2 products of 3-hydroperoxypropanal comprise the reactants for this network) are shown in Fig. 4b. We observe that the success rates for all reactants (apart from species **19**) are greater than 80%, whereas the intended rate varies between 33.3% and 72.7%. The average overall success and intended rates calculated on a unique reaction basis (see the Methods for the distinction) for this network are 88.8% and 54.7%, respectively. To compare with the earlier benchmark, we have also calculated the success and intended rates on a per-attempted-reaction basis (Supplementary Fig. 5), which yields overall success and intended rates of 81.4% and 41.6%, respectively, for YARP. By contrast, Grambow and colleagues[30] reported success and intended rates on a per-attempted-reaction basis for their GSM-based reaction discovery algorithm of 38% and 4%, respectively. On comparing these results, we want to

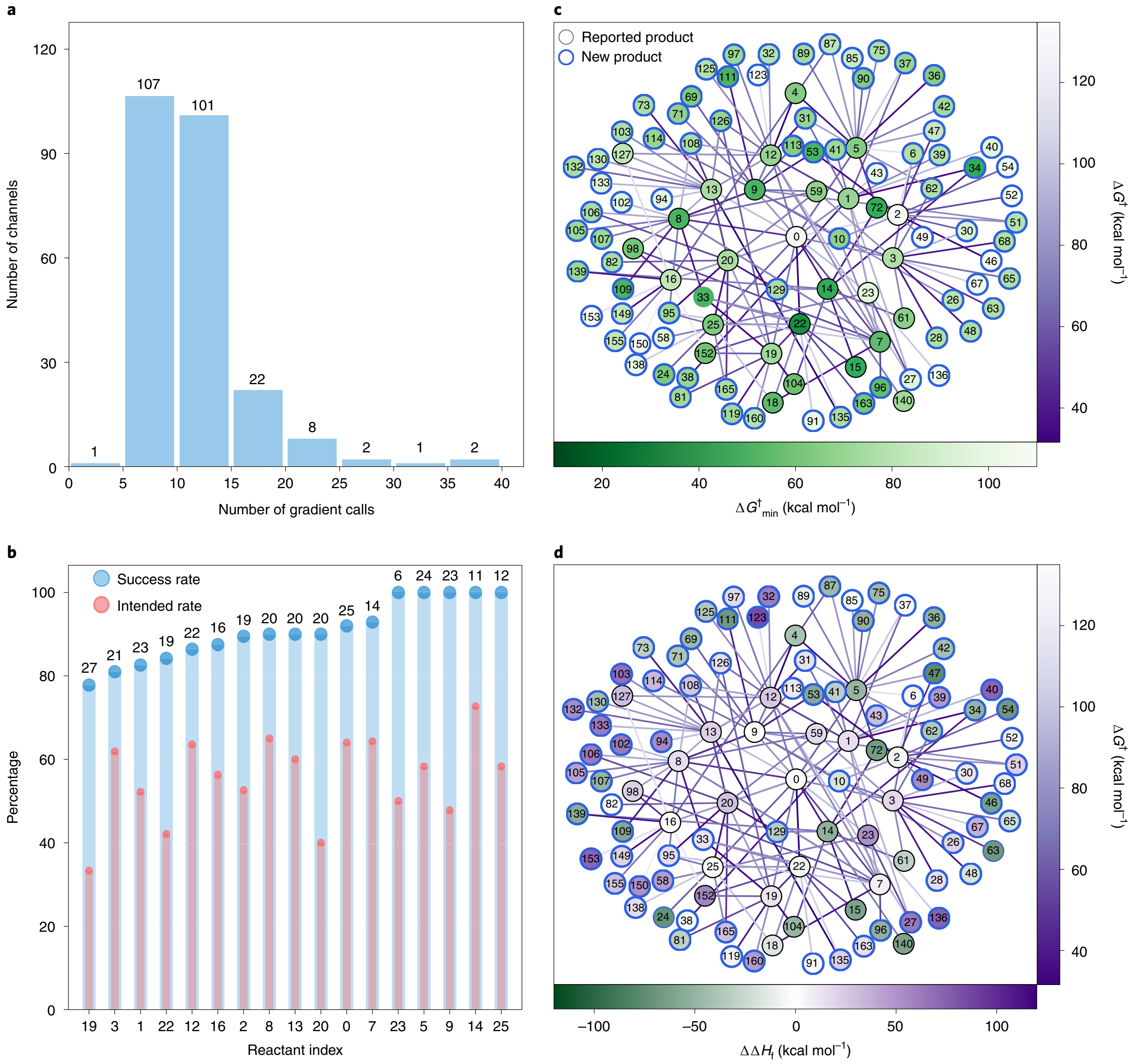


**Fig. 4 | Overview of YARP performance on predicting unimolecular degradation of 3-hydroperoxypropanal. a**, The distribution of gradient calls required to converge each intended reaction, with labels over each bar indicating the number of reactions in each bin. **b**, The success and intended rates of unique reactions involving each reactant. **c**, Degradation network with activation barriers for rate-limiting steps, which represents kinetic accessibility. **d**, Degradation network with heat of reaction, which represents thermodynamic accessibility.

emphasize that the Grambow statistics are for single-step reactions up to b4f4 and not exclusively b2f2, as reported here. Nevertheless, the dramatic difference in both the success and intended rates suggests that the b2f2 ERS enumeration provides realistic reaction candidates that are favorably conditioned for convergence. Moreover, as failed reactions still incur gradient calls, the higher success and intended rates play a major role in reducing the number of gradient calls per discovered reaction in YARP (Supplementary Table 2).

To evaluate whether the products and reactions discovered by YARP are of kinetic or thermodynamic significance, we also performed several further characterizations of the network and reaction pathways. Illustrations of the reaction network are provided in Fig. 4c,d, which display the activation energy associated with all reactions (edge color), as well as the activation energy of the rate-limiting step ($\Delta G^{\dagger}_{\mathrm{min}}$) and heat of reaction ($\Delta\Delta H_{\mathrm{f}}$) for each product. Although these illustrations compress a lot of information, it is apparent that several of the previously unreported products are both thermodynamically and kinetically favorable compared to previously reported products. For example, products **34**, **53**, **109** and **111** (among others) all exhibit relatively low-barrier rate-limiting steps ($\Delta G^{\dagger}_{\mathrm{min}} < 55$ kcal mol$^{-1}$) and strongly exothermic relationships to 3-hydroperoxypropanal ($\Delta\Delta H_{\mathrm{f}} < -50$ kcal mol$^{-1}$). YARP also identified intended channels for several products (**6**, **10** and **24**) that the previous benchmark failed to find intended channels for. The difference in these cases is that the previous benchmark only attempted a single-step reaction that failed to converge to an intended transition

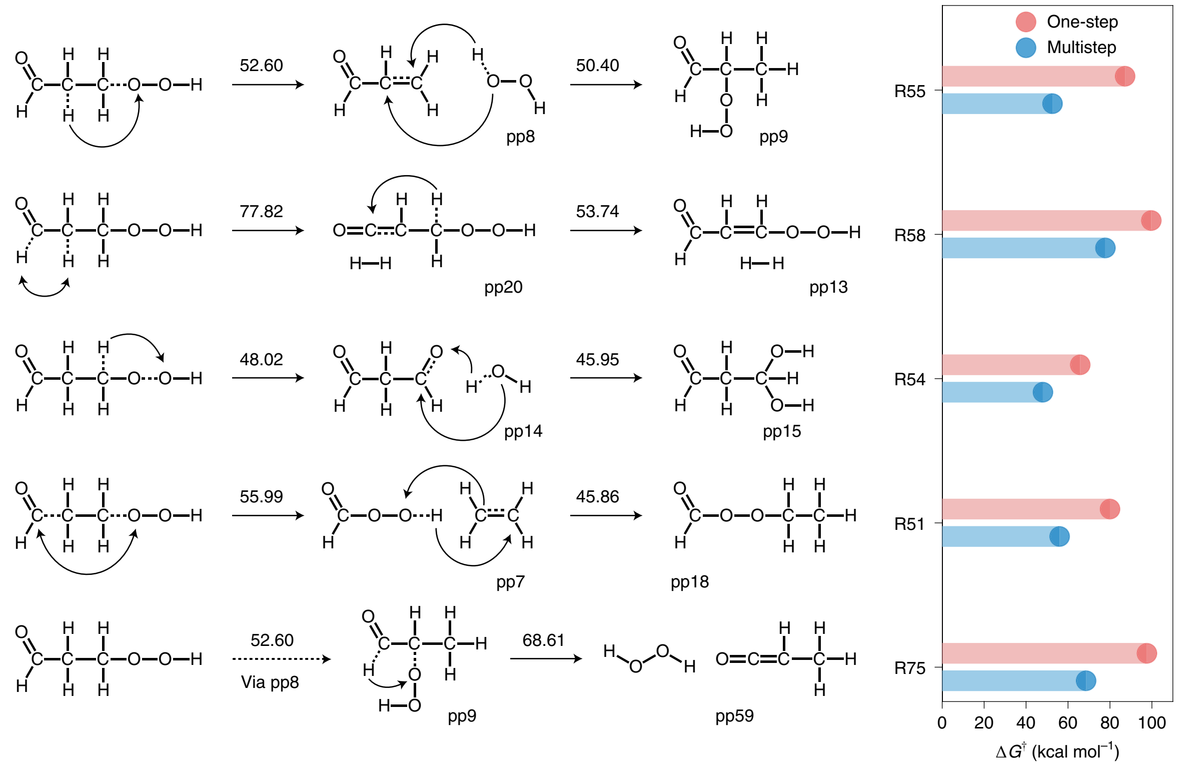


**Fig. 5 | Five multistep reaction pathways identified by YARP that exhibit more than 20 kcal mol$^{-1}$ reduction in activation energy compared with single-step reaction pathways.** Left: an illustration of transformations in multistep reaction pathways. ppN refers to *N*th product identified by YARP and the number above each arrow lists the activation energy for each reaction. All values are in kilocalories per mole. Right: comparisons between the activation energy for the multistep reactions identified by YARP (blue) and single-step reactions reported by Grambow and colleagues (red). RX refers to the corresponding index used by Grambow et al.

state, whereas YARP identified alternative multistep b2f2 pathways for these products. These results clearly demonstrate that the relatively low success and intended rates reported previously ultimately lead to incomplete reaction discovery and the neglect of potentially important reaction products. Documentation of all products, rate-limiting reaction barriers and reaction pathways predicted by YARP are provided in Supplementary Tables 5 and 6.

In addition to discovering new products, we also observe that YARP predicts lower barrier reaction pathways for ten previously reported products (that is, alternative reaction steps with a >5 kcal mol$^{-1}$ reduction in the activation barrier compared with previous reports). For example, in Fig. 5 we show five reaction pathways predicted by YARP that exhibit >20 kcal mol$^{-1}$ reduction in activation energy compared with the previous benchmark. In all of these cases, YARP identifies a favorable multistep reaction pathway, whereas Grambow and colleagues[30] report a single-step reaction. Thus, even in cases where a more complex ERS can be successfully converged to an intended channel, it may not be the most kinetically relevant pathway. By contrast, among the eight products for which YARP finds multistep pathways rather than single-step pathways, we only observe two cases (**R28** and **R69**, in the original benchmark) where Grambow and co-workers reported a lower barrier pathway than the multistep pathway identified by YARP (**0** → **14** → **61** and **0** → **14** → **72**, respectively, as listed in Supplementary Table 5). Both of these cases involve a b3f3 reaction that was not investigated by YARP.

The Korcek reaction is the lowest barrier unimolecular decomposition pathway for 3-hydroperoxypropanal discovered so far[45], in which a five-membered cyclic peroxide (1,2-dioxolan-3-ol) intermediate is first formed, followed by fragmentation to acetic acid and formaldehyde, or formic acid and acetaldehyde. YARP can identify all three Korcek intermediates, denoted as product **22**, **46** and **72**, respectively. Although the reaction pathway from 3-hydroperoxypropanal to 1,2-dioxolan-3-ol is reproduced, the subsequent fragmentation steps are not captured, as they are all concerted b3f3 reactions. YARP reports alternative reaction pathways to these two products with higher energy barrier (Supplementary Table 5). Even so, the formic acid and acetaldehyde product still exhibits among the lowest energy barrier reaction pathways of all discovered products (Supplementary Fig. 6).

**Reaction exploration on larger molecules.** In the previous two benchmarks we explored the performance of comprehensive b2f2 reaction searches in relatively small systems; however, in many scenarios a reactive subset of atoms, or a limited set of reactions are of interest for exploration. For large systems, such strategies are critical for limiting the scope of reactions to make reaction exploration tractable. As an example of applying YARP in this context, we also performed a case study of possible Diels–Alder ring-closures for a ketothioester that was previously studied by Yang and colleagues[39]. This system has 47 atoms and 11 rotatable bonds, posing a prohibitive challenge to a comprehensive reaction search. To limit the reaction space, we followed the previous study and restricted YARP to only explore reactions involving a subset of eight atoms, followed by comprehensive enumeration of all possible Diels–Alder channels. Under these restrictions, eight potential products were enumerated by YARP (Fig. 6a). To account for the additional conformational flexibility of this large system, the conformer–rotamer ensemble sampling tool (CREST)[46] was used to sample ten conformers for each attempted reaction. Without further discussion here, we note that conformer selection in large systems is a critical component of resolving meaningful transition states,

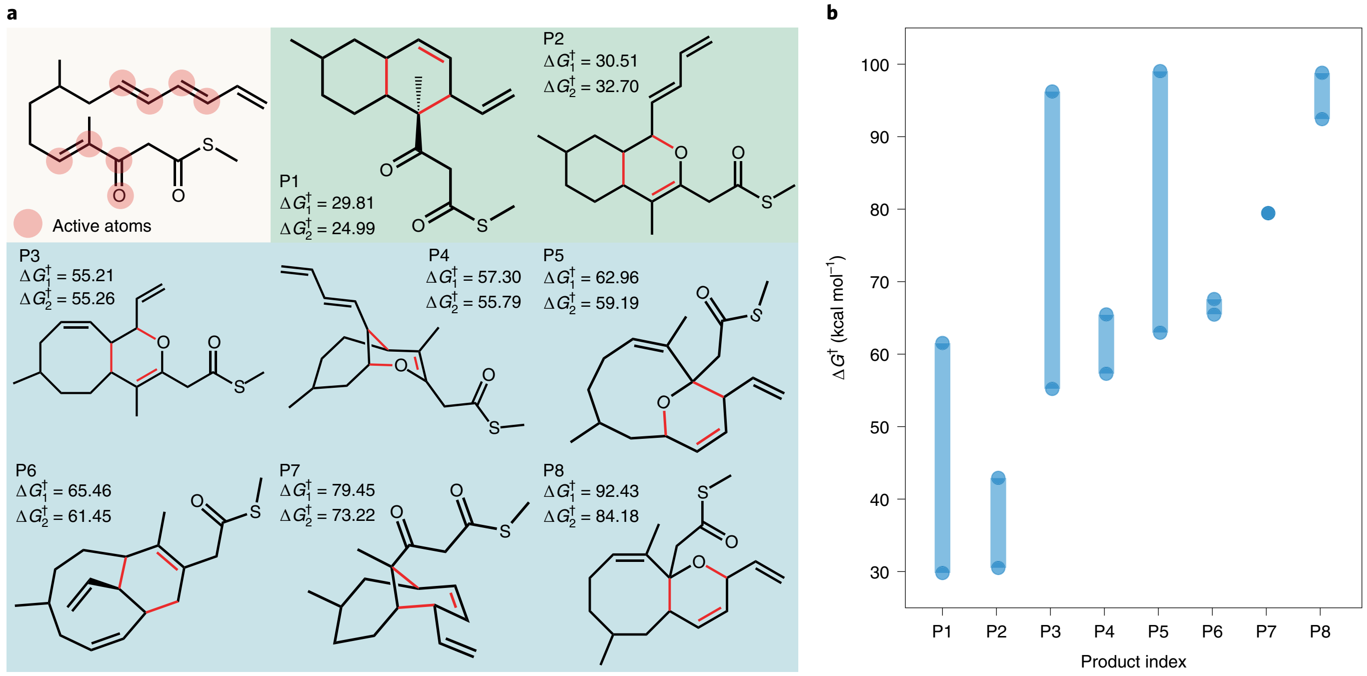


**Fig. 6 | Characterizing competing Diels–Alder ring-closures for a ketothioester. a**, Eight Diels–Alder products of the ketothioester (yellow) involving the active subset of atoms (red). The red bonds in products denote the newly formed bonds. $\Delta G_1^{\ddagger}$ and $\Delta G_2^{\ddagger}$ refer to the free energy of activation calculated at the B3LYP/6-31G and ωB97XD/TZVP levels, respectively (relative to the most stable conformer of the ketothioester in kilocalories per mole). **P1** and **P2** in the green boxes are two products reported in a previous work, whereas **P3**–**P8** in blue boxes are new products identified by YARP. **b**, The range of activation energies calculated at the B3LYP/6-31G level for distinct conformers of each intended reaction.

as evidenced by the large range of activation energies calculated for distinct conformers (Fig. 6b).

For direct comparison with the results of Yang and colleagues[39], all transition states were optimized and IRC calculations were performed at the B3LYP/6-31G level ($\Delta G_1^{\ddagger}$ in Fig. 6a). For comparison with a more modern functional, the transition states of the lowest barrier pathway for each reaction were also reoptimized at the ωB97XD/TZVP level ($\Delta G_2^{\ddagger}$ in Fig. 6a). The two lowest barrier products, **P1** ($\Delta G_1^{\ddagger} = 29.81\,\mathrm{kcal\,mol^{-1}}$) and **P2** ($\Delta G_1^{\ddagger} = 30.51\,\mathrm{kcal\,mol^{-1}}$), discovered by YARP match the predicted products and barriers of the previous study (27.9 kcal mol$^{-1}$ and 30.4 kcal mol$^{-1}$, respectively). All other products have activation energies over 50 kcal mol$^{-1}$, which validates the conclusion that **P1** and **P2** are the kinetically important products. Meanwhile, from the comparison between B3LYP/6-31G and ωB97XD/TZVP, we observe that B3LYP/6-31G accurately predicts the ranking of the eight products according to energy barriers, albeit with a trend to overestimate the activation energies by ~4 kcal mol$^{-1}$.

## Discussion

The presented benchmarks have focused on the extent to which exhaustive graph-based reaction searches of particular reactants can be sped and made more accurate, while leaving the distinct challenges of exploring deep reaction networks mostly unaddressed. In particular, even with the improved scaling of characterizing individual reactants, exhaustive searches of reaction networks beyond a few layers is impractical given the growth in potential reactants as the network grows. To address this challenge, it will be necessary to implement data-driven or physics-based approaches to prioritize reaction characterization for individual reactants and to select among network branches that warrant deeper exploration. We also note that when characterizing deep networks, it is typically unnecessary to explore all branches to an equal depth, which is the common assumption in estimates of quadratic scaling of the number of reactions in a network with respect to number of intermediates. As has been pointed out by Lu, Van de Vijver and others, real reaction networks typically exhibit linear scaling in the number of kinetically relevant reactions with respect to number of intermediates[40,41]. Thus, combining systematic reaction prediction, on a per reactant basis, with kinetic modeling to prioritize which branches to explore to increased reaction depth is a promising pathway towards achieving linear scaling of deep network characterizations, without sacrificing reaction discovery.

In the present work we have focused on the extent to which the b2f2 ERS is capable of recapitulating known reactions and discovering new ones, while also illustrating cases in which it is necessary to include the b3f3 ERS for YARP to find the lowest activation energy channel. The presented benchmarks suggest that simpler ERSs lead to better convergence and more physically relevant pathways compared with more complex reactions such as b3f3 and b4f4. The general conditions under which this is true is interesting to consider, and we have emphasized the cases where the b2f2 ERS missed important chemistry. Thus, more complex ERS(s) are definitely needed in some cases, although empirically these situations seem to be rare for full-octet neutral organics. To avoid the side-effects of generically attempting b3f3 reactions (that is, attempting a large number of spurious or poorly conditioned reactions) it is possible to implement b3f3 reactions as reaction templates when known, albeit at the cost of restricting reaction discovery. Likewise, for addressing larger systems, hybrid strategies that selectively apply multiple ERSs or use reactive subsets of atoms are probably necessary to balance reaction exploration against computational cost. Further methods development to add sophistication to reaction enumeration is clearly a critical avenue of research.

Activities are also underway to further improve YARP's performance and extend its applicability. First, we have focused on b2f2 and b3f3 ERSs due to their relevance to full-octet organics, but more general graph-based ERS definitions are compatible with YARP (for example, b0f1, b1f2 and so on) and will be elucidated in future work as they apply to ions, radicals and organometallics. Second, conformational sampling of transition states and reactant geometries is an obvious extension of the joint-optimization procedure that could improve the quantitative details of discovered transition states. In particular, although we have demonstrated that success and intended rates can be dramatically improved using the joint-optimization procedure, it is possible that some of the failed reactions that are still observed might be converged with better initial conformation generation. Third, prioritizing reaction channels based on enthalpies of reaction is a potentially cost-efficient approach for screening infeasible reactions prior to high-level characterizations. When extending YARP to larger systems, we anticipate that semi-empirical heat of formation models being developed by our group might be employed for this purpose[47]. Finally, as the data generated by high-throughput physics-based approaches like YARP mature, many opportunities will also be created to utilize machine learning to further accelerate transition state characterization and refine reaction enumeration. In combination, these avenues of research make it likely that automated reaction prediction will develop into a routine research tool for elucidating currently intractable reaction network problems.

## Methods

**Graph-based product enumeration.** Identifying transition states is the most time-consuming step in the reaction prediction process. The transition state is located in a $3N-6$ dimensional space, where $N$ is the number of atoms in the system. For all but the smallest systems, brute force optimization is impossible and it is critical to reduce the size of this search. Thus, several heuristics have been developed to identify transition states, either based on reaction templates, artificial forces, or using the local curvature of the potential energy surface[24,25,32,33,48–51]. Separately, several double-ended searching algorithms have been developed for situations where the start and end points (that is, reactant and product states, respectively) are known in advance. In these cases, the transition state search can be recast as a one-dimensional search, whereby a string of states connecting reactant(s) and product(s) is optimized to locate the transition state. In the context of reaction prediction, product enumeration is inexpensive compared with transition state characterization, so one strategy that has been utilized is to enumerate putative products that are in turn efficiently characterized by one or more double-ended transition state algorithms. In this scenario, the double-ended transition state characterization algorithm carries the responsibility of discriminating between plausible and implausible reactions. In practice, however, transition state convergence failure is extremely common (for example, ~60% being recently reported)[30], especially in scenarios with multiple transition states between the reactant and product structures[22,28], suggesting that strategies to improve product enumeration are critical to ensure transition state convergence.

*Elementary reaction step (ERS).* YARP uses the concept of an ERS to enumerate potential products and reduce the occurrence of products separated by more than one transition state[52]. For full-octet neutral organic systems, the simplest reaction step that yields non-trivial products is b2f2. For instance, a form-one-bond step (f1) is inapplicable to full octet organics. Similarly, a b1f1 step simply reproduces the reactant(s), and a b2f1 step results in an unstable Lewis structure; b2f2 thus yields the most elementary set of bond rearrangements amongst full octet organic reactant molecules that can produce distinct products. In recent work, other groups have also included b3f3 and b4f4 steps during enumeration[28–30,53,54]. We note that all b3f3 and b4f4 products can be obtained by repeated applications of b2f2 steps, although the reaction pathways may be distinct. In the case of a sequential reaction, b2f2 reaction pathways are better conditioned to double-ended searches as they reduce the possibility of transition state convergence failure due to multiple intervening transition states. For instance, out of the converged b2f2 reactions investigated here, 98.5% exhibit only one transition state due to the relative rarity of forming a stable intermediate with a single broken bond. However, in the case of a b3f3 concerted reaction (for example, Diels–Alder), sequential application of b2f2 steps would discover the b3f3 product but fail to identify the transition state corresponding to the concerted b3f3 reaction. We note that the relevant type of ERS is system specific, and for metal-containing, ionic, or radical systems, additional steps such as b2f1, and charge transfer steps may be applicable ERSs.

*Bond-electron matrix formalism.* Product enumeration is implemented in YARP using the bond-electron matrix formalism developed by Ugi and colleagues[55]. This approach provides a machine-readable grammar for expressing the Lewis structure of products and reactants, and encoding reactions as matrices. A bond-electron matrix, **A**, for reactant(s) with $N$ atoms is $N$-dimensional and symmetric. The diagonal elements $A_{\mathrm{ii}}$ correspond to the number of lone valence electrons on each atom and the non-diagonal elements $A_{ij}$ correspond to the number of covalent bonds between atoms $i$ and $j$. We will refer to the bond-electron matrices corresponding to reactants and products as **A** and **B**, respectively. Multiple reactant or product molecules can be combined in one bond-electron matrix, with a separate connected subgraph for each molecule. For molecules with multiple resonance structures, YARP stores each structure as a distinct **A** matrix. Furthermore, the matrix formalism can be equivalently recast using lists to avoid handling sparse matrices, but here the matrix formalism will be retained for clarity of description.

The matrix $\mathbf{R}=\mathbf{B}-\mathbf{A}$ describes the changes in bond order and electron transfers associated with the reaction $\mathbf{A}\rightarrow\mathbf{B}$. **R** can be decomposed into a summation of two matrices, $\mathbf{R}=\mathbf{U}+\mathbf{V}$, where **U** is a diagonal matrix that describes electron transfer associated with the reaction, and **V** is a symmetric matrix that describes bond formation and dissociation[56]. In the context of the current work, the b2f2 ERS is used to generate all **R** matrices for a given set of full-octet organic reactants defined by an **A** matrix. For the specific case of the b2f2 ERS, distinct resonance structures do not lead to distinct products; however, in the general case, resonance-equivalent **A** matrices must be considered when generating **R** matrices. All **R** matrices corresponding to b2f2 steps are enumerated by looping over the unique pairs of bonded atoms in **A** (that is, the atoms involved in a double bond do not constitute a valid pair on their own). Without loss of generality, assuming the pair is composed of atoms (a,b) and (c,d), a bond is broken between each pair of atoms by setting $V_{\mathrm{ab}}=V_{\mathrm{ba}}=-1$ and $V_{\mathrm{cd}}=V_{\mathrm{dc}}=-1$ (hereafter, we do not explicitly show the symmetric terms). These broken bonds create the possibility for forming two new pairs of bonds between either (a,c) and (b,d), or (a,d) and (b,c), respectively. Thus, for each pair of two broken bonds, two distinct reaction matrices are formed with $V_{\mathrm{ac}}=V_{\mathrm{bd}}=+1$ or $V_{\mathrm{ad}}=V_{\mathrm{bc}}=+1$, respectively (Extended Data Fig. 1a). In the case where an atom is shared between the two bonds (that is, (a,b) and (a,c) compose the unique bonded pairs of atoms), we only generate a reaction matrix if atom a is bonded to another atom d that possesses at least one lone pair of electrons. In this case, the b2f2 step yields an effective bond rearrangement of $V_{\mathrm{ab}}=V_{\mathrm{ac}}=-1$ and $V_{\mathrm{bc}}=V_{\mathrm{ad}}=+1$ with a transfer of valence electrons corresponding to $U_{\mathrm{aa}}=+2$ and $U_{\mathrm{dd}}=-2$, and the resulting product will have formal charges on the a and d atoms. One example of such a reaction is the decomposition of formaldehyde to carbon monoxide and hydrogen (Extended Data Fig. 1b). The matrix **B**, corresponding to the products of each reaction, is obtained by adding the **R** matrix generated at each iteration to the reactant matrix **A**.

Finally, we note that the matrix representation is non-unique due to graph isomorphism. This leads to potential redundancy during recursive enumeration, where multiple identical reactants or products are separately subjected to reaction evaluation while in fact representing identical molecular structures. To avoid this issue, YARP implements a canonicalization procedure to sort the indexing of the bond-electron matrix based on a connectivity hash function for each atom. This hash function incorporates the elemental identity of each atom, its bonded neighbors and their bonded neighbors, out to an arbitrary depth (ten in the present work). This procedure ensures that the same reactants or products, regardless of atom indexing, resolve to the same bond-electron matrix and thus avoids redundant calculations. Furthermore, symmetry equivalent atoms evaluate to an identical hash value, which can in principle be used to further reduce the number of reactions performed. In the current study, symmetry was not used to reduce the number of reactions; thus in some cases, formally equivalent reactions (for example, a reaction involving one of two equivalent hydrogens on a methylene) were performed multiple times.

*Geometry Initialization.* The bond-electron formalism provides an effective machine-readable grammar for describing reactions in terms of molecular graphs, however these graphs must be converted to three-dimensional structures for determining the transition states and thermodynamics of the reactions. Moreover, double-ended transition state algorithms are highly sensitive to the geometry of the initial structures, with documented convergence failures if the reactant and product structures are poorly aligned[17,57]. We have also observed that even in cases where transition state convergence occurs, the initial geometry can strongly impact whether the discovered transition state corresponds to the intended reaction after inspection by an IRC calculation. Within YARP, this issue is addressed by jointly optimizing reactant and product geometries for each investigated reaction. This is accomplished by first generating an optimized product geometry using the universal force-field (UFF)[58] as implemented in Open Babel[59]. The reactant geometry is then generated using the product geometry as a starting point, but with optimization occurring on the UFF potential energy surface corresponding to the bond-electron matrix of the reactant. The UFF optimized reactant and product geometries are then optimized at the GFN2-xTB level. Finally, root mean square deviation minimization, as implemented in the atomic simulation environment (ASE)[60], is applied to each product geometry by rotating and translating the center-of-mass to align with the reactant[61]. This procedure yields high overall transition state convergence and a high rate of discovering transition states corresponding to intended reactions. The only deviation from this protocol in the current study is that for the large ketothioester system, the joint-optimization procedure was applied to the distinct conformers generated from CREST[46] sampling as opposed to a single conformer generated from Open Bable/UFF as in the other cases.

**Reaction network construction.** YARP incrementally builds the reaction network for a given set of reactants by alternating between product enumeration and transition state characterization steps. For exploring deep networks, products from one generation may serve as reactants for a subsequent iteration of reaction enumeration. Depending on the application, it may also be necessary to include bimolecular reactions between products produced at different levels of the reaction network. In the current benchmark systems (described below), we have used YARP to perform comprehensive enumeration of b2f2 products up to two iterations. Comprehensive b2f2 sampling scales as $\binom{N}{2}$ where $N$ is the number of unique bonded pairs of atoms in the reactant matrix. In cases with symmetry equivalent atoms the actual unique number of reactions may be substantially reduced. This scaling yields a tractable number of reactions to investigate at each iteration of product enumeration, even for relatively large reactant matrices; however, the actual performance is ultimately determined by the ability to localize accurate transition states for each reaction.

After the product enumeration phase, YARP uses one or more double-ended search algorithms to localize transition states for each reaction, followed by Berny optimization and an IRC calculation to identify and characterize the final transition state. Recently, Grambow and colleagues[30] compared several double-ended search methods, including the freezing string method, the GSM and the single-ended growing string method (SSM). Based on their study, GSM achieves a balance between computational cost and convergence rate. Although GSM is faster than SSM, directly applying it at the DFT level is still prohibitively costly for exploring deep networks or reactions involving large molecules. To decrease the computational cost in YARP, the current set of reaction pathways were localized using the GSM algorithm with the semi-empirical GFN2-xTB model chemistry prior to searching for the final DFT-level transition states via Berny optimization. This choice was motivated in part by the recent demonstration of Dohm and colleagues[62] that combining GFN2-xTB with GSM to localize organometallic transition states achieved convergence for approximately 90% of the investigated reactions. We investigated the possibility of using the GFN2-xTB transition state energies directly and thus circumventing the need for DFT calculations; however, substantial deviations are observed between the GFN2-xTB values after GSM localization and the final DFT optimized energies (Supplementary Fig. 3). Thus, after convergence of the GSM pathway, YARP automatically selects the highest energy node as a preliminary transition state candidate that is then used as a starting structure for Berny transition state optimization. After convergence, an IRC calculation is performed to determine whether the detected saddle point corresponds to the input reaction channel. If the end nodes obtained by the IRC match the input reactant and product bond-electron matrices, the reaction channel is classified as intended, otherwise it is classified as unintended. The latter reactions are removed from the final network, since they represent a failure of the algorithm to identify a relevant transition state either due to the fact that the attempted reaction is unphysical or because the double-ended search was poorly conditioned.

**Reaction network analysis.** The large number of reactions generated by automated algorithms creates an interpretation bottleneck for extracting mechanistic information about reactions and competing pathways. In particular, ad hoc evaluation of pathways and transition states is both error-prone and too costly to comprehensively evaluate large networks. To address this, we have implemented three analysis routines within YARP as a starting point to automate the extraction of semantic information from algorithmically generated reaction networks. First, rate-limiting steps are evaluated for all products identified by YARP to estimate the kinetic relevance of various reaction pathways discovered in the network. Here, the rate-limiting step is defined as the reaction with the maximum Gibbs free energy of activation in a pathway connecting the starting reactants to a given product (Fig. 1d). For a network composed of multiple enumeration steps, multiple distinct pathways will potentially exist that yield a specific product. To compare these competing pathways, YARP performs a breadth-first search from each product, using the directed graph of reactions in the network to identify distinct pathways and keeping all pathways that terminate at the reactant node. For each product, all pathways out to $m+n$, are enumerated, where $m$ is the depth of the product in the network (that is, the reactant is considered as depth zero, and the product depth is defined based on the smallest number of reactions connecting it to the reactant) and $n$ is a user-defined parameter that controls the maximum number of reactions in a pathway ($n=2$ in this work). The pathway with the lowest Gibbs free energy of activation (denoted throughout this work as the activation energy) rate-limiting step is considered the dominant pathway, and this barrier ($\Delta G^{\dagger}_{\mathrm{min}}$) is reported for all products in the network. For bimolecular reactions the reported activation energies correspond to the difference between the transition state energy and the summation of individual reactant energies. In addition, we also report the heat of reaction for each product ($\Delta\Delta H_{\mathrm{f}}$) calculated with respect to the reactants to characterize the thermodynamic relevance of the various products identified by YARP. The heat of formation of each product was calculated at the DFT level based on difference in thermal enthalpies between the reactants and the products. Finally, we have implemented an algorithm to characterize the occurrence of bond-breaking and bond-formation steps relative to the transition state. This algorithm uses a distance-based criteria to define changes in bond configuration. Specifically, bond breaking (formation) is defined to occur when the observed separation between a pair of atoms is greater (less) than 1.2 times the summation of the UFF radii of the atoms. This algorithm is applied to the minimum energy pathway generated by the IRC calculation to determine which bond-breaking or bond-forming events are associated with the transition state and whether they occur sequentially or in a concerted fashion. The specific position within the minimum energy path of the identified bond changes is relatively insensitive to the choice of scaling factor, but it cannot be set too large without detecting spurious bonds. In the case of double-bonds, the distance based criteria is inapplicable and the algorithm assigns bond-breaking to occur at the same time as a bond-forming event involving either of the double-bonded atoms.

**Details on performance statistics.** The number of quantum chemistry gradient calls is a commonly used surrogate for the computational cost of a reaction exploration method, as the transition state calculations represent the most expensive step in reaction characterization[20,29,30,44]. We report timing (Extended Data Fig. 2) and scale-up data (Supplementary Fig. 1 and Supplementary Tables 1 and 2) that confirms that the most expensive step in YARP is DFT-based transition state optimization, with negligible computational costs (that is, <1−5% of walltime depending on system size) associated with the GFN2-xTB gradient calls, geometry initialization, and product enumeration steps. Thus, we report the distribution of DFT gradient calls associated with all reaction channels explored with YARP as a measure of the overall computational costs that can be compared with other approaches. The gradient calls associated with IRC calculations are excluded from these totals to be consistent with the gradients statistics reported by Grambow and colleagues as the IRC calculations are a validation activity and not directly associated with finding the transition states.

Additionally, there is a minor ambiguity in how the success and intended rates have been defined in previous work, depending on whether one uses the total number of reactions or the total number of unique reactions in calculating these rates. The distinction is that the same product may be obtained by multiple single-step reactions (for example, if there are symmetry equivalent atoms in the reactant or in the product) and we are here defining ‘unique’ reactions to be reactions that yield distinct products. We adopt the convention that if at least one reaction out of a set of equivalent reactions is classified as successful (intended), then that unique reaction is classified as successful (intended). Our rationale is that since at least one transition state was successful (intended) for such a reaction, any discrepancy in the rates calculated using unique reactions versus total reactions reflects the sensitivity of converging transition states rather than the physical relevance of the reaction. We report rates on a unique reaction basis within the main text for each distinct reactant investigated in this study. We also report statistics on a per reaction basis (Supplementary Fig. 5).

**Benchmark systems.** Three benchmarks are reported in the main text. The first is predicting a set of single-step organic reactions curated by Zimmerman[20]. This dataset includes 25 reactants and 105 single-step reactions. We excluded zwitterionic and ionic species ($NH_3BH_3$ and the taxadiene carbocation) and reactants with incomplete octets ($NH_2BH_2$ and $SiH_2$) as they require a more general set of ERSs beyond b2f2 and b3f3. We also excluded alanine dipeptide from reaction benchmarking, as Zimmerman only reported a conformational rearrangement rather than a reaction for this compound. The final dataset evaluated here was composed of 20 distinct reactants and 61 distinct reactions. We used YARP to characterize all b2f2 reactions for each reactant, resulting in a total of 533 distinct reactions. The only molecule where all b2f2 reactions were not performed was phenyl ether, for which the default setting of YARP is to exclude breaking bonds in benzene rings. b3f3 reaction searches were also performed on two reactants for comparison with the b2f2 results (Extended Data Fig. 3) on two of the Zimmerman reactants. Berny optimizations and IRC calculations were performed at the B3LYP/6-31G** level to be consistent with the Zimmerman dataset.

YARP was also used to characterize the thermal decomposition network of 3-hydroperoxypropanal, recently benchmarked by Grambow and colleagues[30]. In this earlier work, the reaction networks predicted by five distinct reaction discovery algorithms were compared, in total consisting of 55 distinct products and 75 intended reactions. We used YARP to perform two iterations of b2f2 reactions for this system (that is, all b2f2 reactions for 3-hydroperoxypropanal and the first generation of products) Cycloalkyne, cycloallene, and three and four membered rings were excluded as potential products during these steps, since they were not identified among the kinetically important structures in the previous study. b2f2 reactions were only applied to products of 3-hydroperoxypropanal whose transition state calculations converged to the intended reaction as determined from IRC calculations. The decision to investigate all b2f2 reactions, save the noted exceptions, out to the second generation for the 3-hydroperoxypropanal network was made to provide a direct comparison with the products discovered by the earlier study, which reported single-step reaction channels up to b4f4. This distinction is important for comparing the results in the two cases, since YARP finds two-step b2f2 pathways for many products that were previously only attempted by single-step reactions. Berny optimizations and IRC calculations in this benchmark were performed at the B3LYP/6-31+G* level to be consistent with the Grambow dataset.

YARP was applied in a third benchmark for characterizing competing Diels–Alder reactions in a 47 atom ketothioester that was previously studied by Yang and colleagues[39]. In the earlier study, a comparison of two out of the eight possible Diels–Alder reactions was reported. Here, YARP was applied YARP to converge all eight of the possible reactions.

**Computational details.** In the present study, YARP used Gaussian 16 as the reference quantum chemistry engine for the DFT calculations associated with the Berny optimizations and IRC characterizations[63]. The GSM calculations were performed by interfacing YARP with the pyGSM package[64,65] using default convergence hyperparameters (nine nodes, climbing image method and translation-rotation-internal coordinate system). All GFN2-xTB calculations were performed with the xTB program (v.6.2.3) maintained by the Grimme group[43]. UFF-based geometry optimizations were performed with Open Babel (v.2.4.1)[59]. All simulations were run on a 580 node commodity cluster composed of two Intel Haswell CPUs (2.60 GHz), 20 effective cores, and 128 GB of memory per node. DFT calculations were performed with 20-core parallelization, while all other calculations were performed as bundled single-core jobs.

## Data availability

The authors declare that the data supporting the findings of this study are available within the paper and its supplementary information files. Source data for Figs. 3–6 and Extended Data Figs. 2 and 3 are available in Source Data. Data referenced from other studies were scraped from the manuscripts or supporting information of the indicated publications, including the Zimmerman[20] and KHP decomposition datasets[30]. Further raw data sources generated by this work are available at https://doi.org/10.6084/m9.figshare.14766624 (ref. [66]), including raw output files and molecular geometries.

## Code availability

The version of YARP used in this study and a guide to reproducing the results is available through GitHub under the GNU GPL-3.0 License (https://github.com/zhaoqy1996/YARP). The specific version of the package used to generate the results in the current study can be found at https://doi.org/10.5281/zenodo.4947195 (ref. [67]).

Received: 11 December 2020; Accepted: 18 June 2021;
Published online: 22 July 2021

## References

1. Westbrook, C. K., Mizobuchi, Y., Poinsot, T. J., Smith, P. J. & Warnatz, J. Computational combustion. *Proc. Combust. Inst.* **30**, 125–157 (2005).
2. Sarathy, S. M. et al. Comprehensive chemical kinetic modeling of the oxidation of 2-methylalkanes from C7 to C20. *Combust. Flame* **158**, 2338–2357 (2011).
3. Rodrigo, G., Carrera, J., Prather, K. J. & Jaramillo, A. DESHARKY: automatic design of metabolic pathways for optimal cell growth. *Bioinformatics* **24**, 2554–2556 (2008).
4. Wu, D., Wang, Q., Assary, R. S., Broadbelt, L. J. & Krilov, G. A computational approach to design and evaluate enzymatic reaction pathways: application to 1-butanol production from pyruvate. *J. Chem. Inf. Model.* **51**, 1634–1647 (2011).
5. Stine, A. et al. Exploring de novo metabolic pathways from pyruvate to propionic acid. *Biotechnol. Prog.* **32**, 303–311 (2016).
6. Jalan, A., Allen, J. W. & Green, W. H. Chemically activated formation of organic acids in reactions of the Criegee intermediate with aldehydes and ketones. *Phys. Chem. Chem. Phys.* **15**, 16841–16852 (2013).
7. Rousso, A. C., Hansen, N., Jasper, A. W. & Ju, Y. Identification of the Criegee intermediate reaction network in ethylene ozonolysis: impact on energy conversion strategies and atmospheric chemistry. *Phys. Chem. Chem. Phys.* **21**, 7341–7357 (2019).
8. Corey, E. J. & Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science* **166**, 178–192 (1969).
9. Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **51**, 1281–1289 (2018).
10. Schwaller, P. et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).
11. Simm, G. N., Vaucher, A. C. & Reiher, M. Exploration of reaction pathways and chemical transformation networks. *J. Phys. Chem. A* **123**, 385–399 (2018).
12. Green, W. H. *Computer Aided Chemical Engineering* Vol. 45, 259–294 (Elsevier, 2019).
13. Vernuccio, S. & Broadbelt, L. J. Discerning complex reaction networks using automated generators. *AIChE J.* **65**, e16663 (2019).
14. Coley, C. et al. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370–377 (2019).
15. Schreck, J. S., Coley, C. W. & Bishop, K. J. Learning retrosynthetic planning through simulated experience. *ACS Cent. Sci.* **5**, 970–981 (2019).
16. Henkelman, G., Uberuaga, B. P. & Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **113**, 9901–9904 (2000).
17. Zimmerman, P. M. Growing string method with interpolation and optimization in internal coordinates: method and examples. *J. Chem. Phys.* **138**, 184102 (2013).
18. Birkholz, A. B. & Schlegel, H. B. Path optimization by a variational reaction coordinate method. I. Development of formalism and algorithms. *J. Chem. Phys.* **143**, 244101 (2015).
19. Behn, A., Zimmerman, P. M., Bell, A. T. & Head-Gordon, M. Efficient exploration of reaction paths via a freezing string method. *J. Chem. Phys.* **135**, 224108 (2011).
20. Zimmerman, P. M. Reliable transition state searches integrated with the growing string method. *J. Chem. Theory Comput.* **9**, 3043–3050 (2013).
21. Martínez, T. J. Ab initio reactive computer aided molecular design. *Acc. Chem. Res.* **50**, 652–656 (2017).
22. Dewyer, A. L., Argüelles, A. J. & Zimmerman, P. M. Methods for exploring reaction space in molecular systems. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **8**, e1354 (2018).
23. Unsleber, J. P. & Reiher, M. The exploration of chemical reaction networks. *Annu. Rev. Phys. Chem.* **71**, 121–142 (2020).
24. Luo, Y., Maeda, S. & Ohno, K. Automated exploration of stable isomers of $H^+(H_2O)_n$ ($n = 5$–$7$) via ab initio calculations: an application of the anharmonic downward distortion following algorithm. *J. Comput. Chem.* **30**, 952–961 (2009).
25. Maeda, S., Taketsugu, T. & Morokuma, K. Exploring transition state structures for intramolecular pathways by the artificial force induced reaction method. *J. Comput. Chem.* **35**, 166–173 (2014).
26. Maeda, S., Harabuchi, Y., Takagi, M., Taketsugu, T. & Morokuma, K. Artificial force induced reaction (AFIR) method for exploring quantum chemical potential energy surfaces. *Chem. Rec.* **16**, 2232–2248 (2016).
27. Shang, C. & Liu, Z. P. Stochastic surface walking method for structure prediction and pathway searching. *J. Chem. Theory Comput.* **9**, 1838–1845 (2013).
28. Zimmerman, P. M. Automated discovery of chemically reasonable elementary reaction steps. *J. Comput. Chem.* **34**, 1385–1392 (2013).
29. Suleimanov, Y. V. & Green, W. H. Automated discovery of elementary chemical reaction steps using freezing string and Berny optimization methods. *J. Chem. Theory Comput.* **11**, 4248–4259 (2015).
30. Grambow, C. A. et al. Unimolecular reaction pathways of a $\gamma$-ketohydroperoxide from combined application of automated reaction discovery methods. *J. Am. Chem. Soc.* **140**, 1035–1048 (2018).
31. Broadbelt, L. J., Stark, S. M. & Klein, M. T. Computer generated pyrolysis modeling: on-the-fly generation of species, reactions, and rates. *Ind. Eng. Chem. Res.* **33**, 790–799 (1994).
32. Gao, C. W., Allen, J. W., Green, W. H. & West, R. H. Reaction mechanism generator: automatic construction of chemical kinetic mechanisms. *Comput. Phys. Commun.* **203**, 212–225 (2016).
33. Van de Vijver, R. & Zádor, J. KinBot: automated stationary point search on potential energy surfaces. *Comput. Phys. Commun.* **248**, 106947 (2020).
34. Bergeler, M., Simm, G. N., Proppe, J. & Reiher, M. Heuristics-guided exploration of reaction mechanisms. *J. Chem. Theory Comput.* **11**, 5712–5722 (2015).
35. Puripat, M. et al. The Biginelli reaction is a urea-catalyzed organocatalytic multicomponent reaction. *J. Org. Chem* **80**, 6959–6967 (2015).
36. Ludwig, J. R., Zimmerman, P. M., Gianino, J. B. & Schindler, C. S. Iron(III)-catalysed carbonyl–olefin metathesis. *Nature* **533**, 374–379 (2016).
37. Dewyer, A. L. & Zimmerman, P. M. Simulated mechanism for palladium-catalyzed, directed $\gamma$-arylation of piperidine. *ACS Catal.* **7**, 5466–5477 (2017).
38. Jacobson, L. D. et al. Automated transition state search and its application to diverse types of organic reactions. *J. Chem. Theory Comput.* **13**, 5780–5797 (2017).
39. Yang, M., Zou, J., Wang, G. & Li, S. Automatic reaction pathway search via combined molecular dynamics and coordinate driving method. *J. Phys. Chem. A* **121**, 1351–1361 (2017).
40. Lu, T. & Law, C. K. Toward accommodating realistic fuel chemistry in large-scale computations. *Prog. Energy Combust. Sci.* **35**, 192–215 (2009).
41. Van de Vijver, R. et al. Automatic mechanism and kinetic model generation for gas-and solution-phase processes: a perspective on best practices, recent advances, and future challenges. *Int. J. Chem. Kinet.* **47**, 199–231 (2015).
42. Grimme, S., Bannwarth, C. & Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (Z = 1–86). *J. Chem. Theory Comput.* **13**, 1989–2009 (2017).

43. Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB? An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).
44. Maeda, S. & Harabuchi, Y. On benchmarking of automated methods for performing exhaustive reaction path search. *J. Chem. Theory Comput.* **15**, 2111–2115 (2019).
45. Jalan, A. et al. New pathways for formation of acids and carbonyl products in low-temperature oxidation: the Korcek decomposition of $\gamma$-ketohydroperoxides. *J. Am. Chem. Soc.* **135**, 11100–11114 (2013).
46. Pracht, P., Bohle, F. & Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **22**, 7169–7192 (2020).
47. Zhao, Q. & Savoie, B. M. Self-consistent component increment theory for predicting enthalpy of formation. *J. Chem. Inf. Model.* **60**, 2199–2207 (2020).
48. Tsai, C. J. & Jordan, K. D. Use of an eigenmode method to locate the stationary points on the potential energy surfaces of selected argon and water clusters. *J. Phys. Chem.* **97**, 11227–11237 (1993).
49. Maeda, S. & Ohno, K. Global mapping of equilibrium and transition structures on potential energy surfaces by the scaled hypersphere search method: applications to ab initio surfaces of formaldehyde and propyne molecules. *J. Phys. Chem. A* **109**, 5742–5753 (2005).
50. Maeda, S., Ohno, K. & Morokuma, K. Systematic exploration of the mechanism of chemical reactions: the global reaction route mapping (GRRM) strategy using the ADDF and AFIR methods. *Phys. Chem. Chem. Phys.* **15**, 3683–3701 (2013).
51. Martínez-Núñez, E. An automated method to find transition states using chemical dynamics simulations. *J. Comput. Chem.* **36**, 222–234 (2015).
52. Yoneda, Y. A computer program package for the analysis, creation, and estimation of generalized reactions? GRACE. I. Generation of elementary reaction network in radical reactions? GRACE (I). *Bull. Chem. Soc. Jpn.* **52**, 8–14 (1979).
53. Zimmerman, P. M. Navigating molecular space for reaction mechanisms: an efficient, automated procedure. *Mol. Simul.* **41**, 43–54 (2015).
54. Kim, Y., Kim, J. W., Kim, Z. & Kim, W. Y. Efficient prediction of reaction paths through molecular graph and reaction network analysis. *Chem. Sci.* **9**, 825–835 (2018).
55. Ugi, I. et al. New applications of computers in chemistry. *Angew. Chem. Int. Ed.* **18**, 111–123 (1979).
56. Di Maio, F. P. & Lignola, P. G. KING, a kinetic network generator. *Chem. Eng. Sci.* **47**, 2713–2718 (1992).
57. Baker, J., Kessi, A. & Delley, B. The generation and use of delocalized internal coordinates in geometry optimization. *J. Chem. Phys.* **105**, 192–212 (1996).
58. Rappé, A. K., Casewit, C. J., Colwell, K. S., Goddard III, W. A. & Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024–10035 (1992).
59. O'Boyle, N. M. et al. Open Babel: an open chemical toolbox. *J. Cheminf.* **3**, 33 (2011).
60. Larsen, A. et al. The atomic simulation environment? A Python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).
61. Melander, M., Laasonen, K. & Jonsson, H. Removing external degrees of freedom from transition-state search methods using quaternions. *J. Chem. Theory Comput.* **11**, 1055–1062 (2015).
62. Dohm, S., Bursch, M., Hansen, A. & Grimme, S. Semiautomated transition state localization for organometallic complexes with semiempirical quantum chemical methods. *J. Chem. Theory Comput.* **16**, 2002–2012 (2020).
63. Frisch, M. J. et al. *Gaussian 16 Revision C.01* (Gaussian, 2016).
64. Wang, L. P. & Song, C. Geometry optimization made simple with translation and rotation coordinates. *J. Chem. Phys.* **144**, 214108 (2016).
65. Aldaz, C., Kammeraad, J. A. & Zimmerman, P. M. Discovery of conical intersection mediated photochemistry with growing string methods. *Phys. Chem. Chem. Phys.* **20**, 27394–27405 (2018).
66. Zhao, Q., Savoie, B. *YARP Dataset* (FigShare, 2021); https://doi.org/10.6084/m9.figshare.14766624
67. Zhao, Q., Savoie, B. *YARP: Yet Another Reaction Program (YARP)* (Zenodo, 2021); https://doi.org/10.5281/zenodo.4947195

## Acknowledgements

The work performed by Q.Z. and B.M.S was made possible by the Office of Naval Research (ONR) through support provided by the Energetic Materials Program (MURI grant no. N00014-21-1-2476, Program Manager: C. Stoltz). B.M.S also acknowledges partial support for this work from the Dreyfus Program for Machine Learning in the Chemical Sciences and Engineering and the Purdue Process Safety and Assurance Center. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

## Author contributions

Q.Z. and B.M.S conceived and designed the study. Q.Z developed tools, performed analysis and wrote the paper. B.M.S. oversaw the project and wrote the paper. All authors reviewed the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

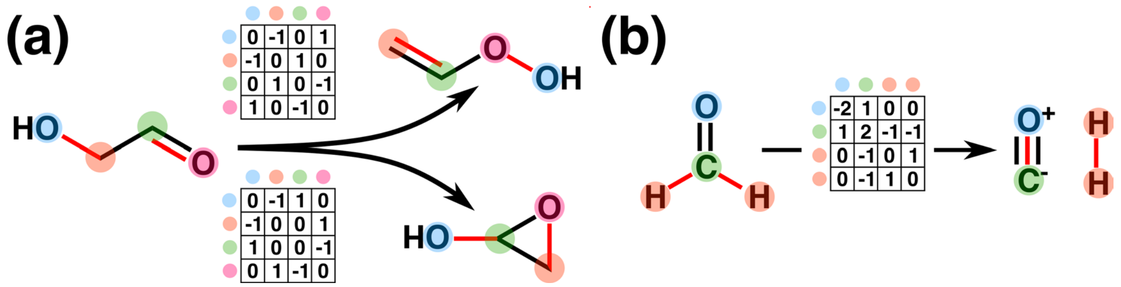
**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s43588-021-00101-3.

**Correspondence and requests for materials** should be addressed to B.M.S.

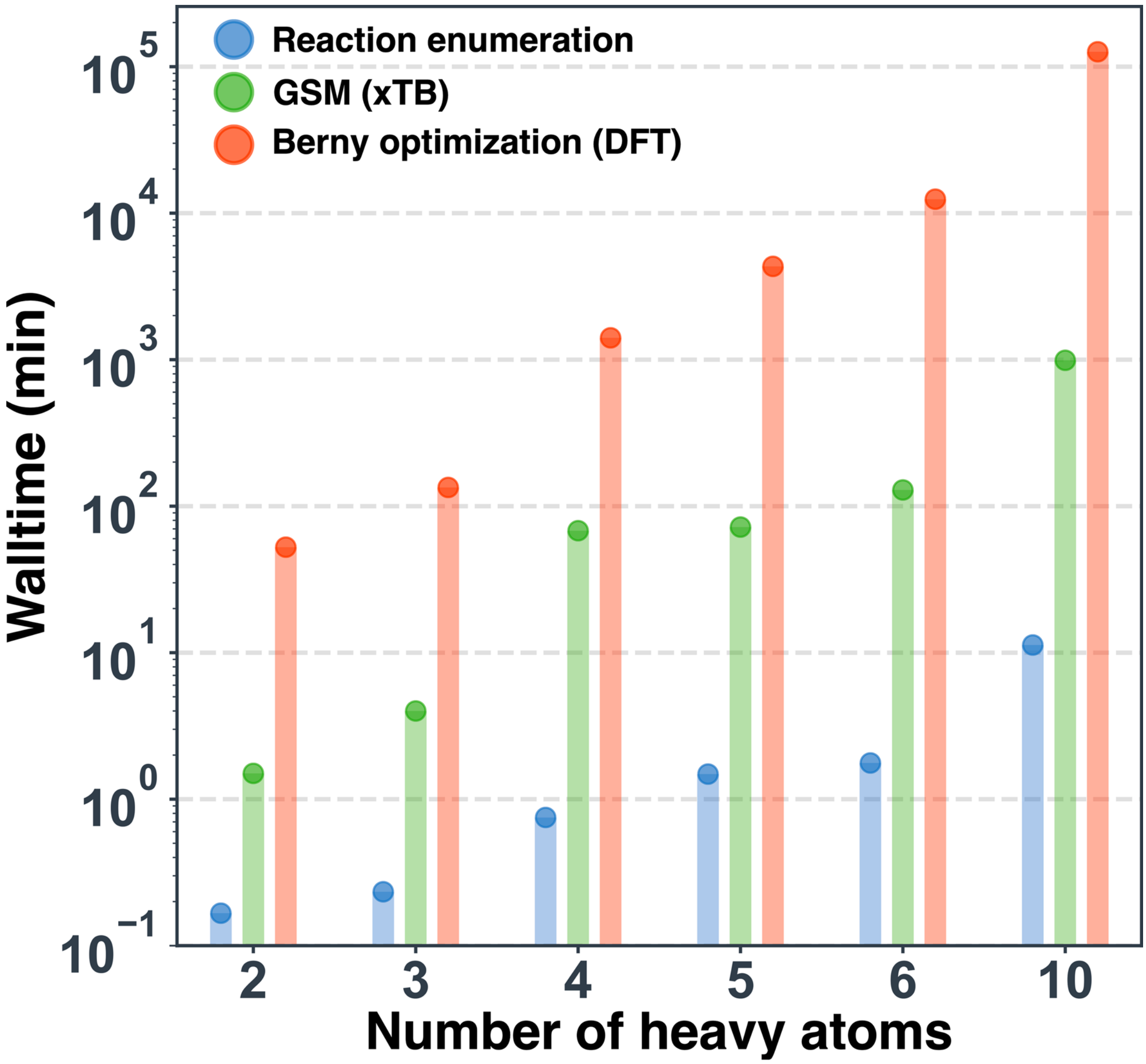
**Reviewer recognition statement** *Nature Computational Science* thanks Cyrille Lavigne, Andreas Hansen and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Handling editor: Jie Pan, in collaboration with the *Nature Computational Science* team.

**Reprints and permissions information** is available at www.nature.com/reprints.

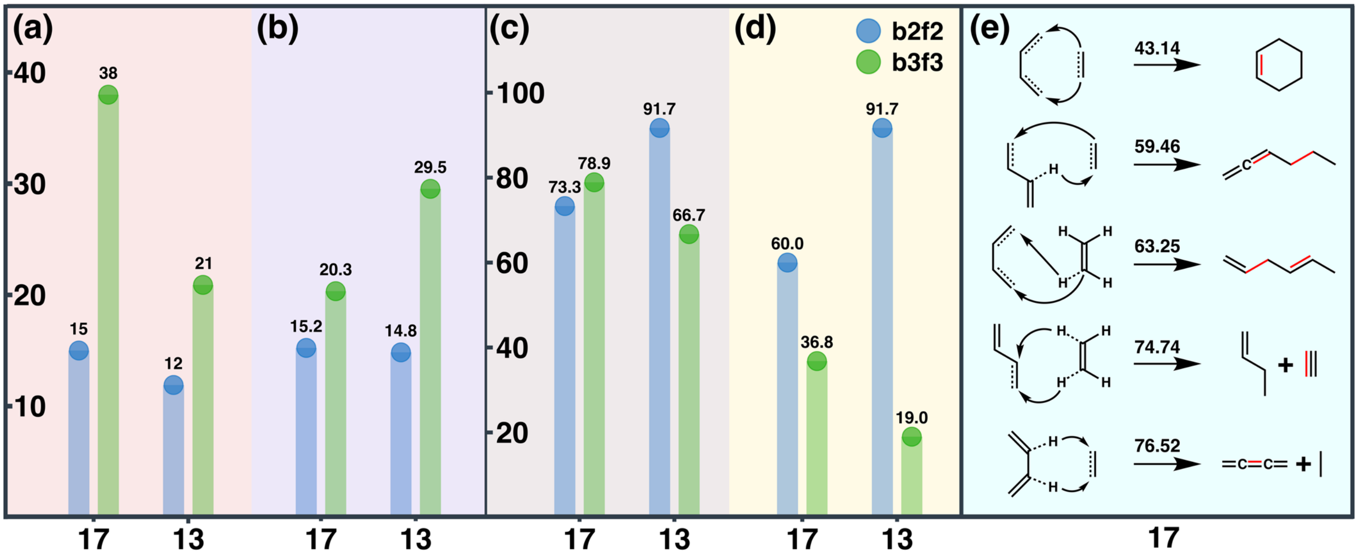
**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Extended Data Fig. 1 | Illustration of Elementary Reaction Steps.** Two cases of the ‘break two bonds and form two bonds’ (b2f2) elementary reaction step (ERS). **a**, The two bonds involved in the ERS connect four different atoms. **b**, An atom is shared between the two bonds involved in the ERS.



**Extended Data Fig. 2 | Timing comparisons for YARP.** Wall times for reaction enumeration, GFN2-xTB/GSM, and Berny optimization with respect to the number of heavy atoms in the reactant. The cases shown here are drawn from the Zimmerman dataset. The computational cost of Berny optimization occupies 95% to 99% of the total cost while the GSM at most contributes ~ 5%. All walltimes are reported without parallelization (that is, single-core equivalent walltimes). Additional timing details are reported in Section 1 of the Supporting Information.



**Extended Data Fig. 3 | Comparison of b2f2 and b3f3 reaction searches and performance statistics.** Comparison of b2f2 and b3f3 reaction enumeration for the reactants 1,3-butadiene and ethene (**17**), and isobutene and water (**13**) from the Zimmerman dataset. **a**, Number of potential products, **b**, average number of DFT gradient calls per successful channel, **c**, the success rates of unique reactions and **d**, the intended rates of unique reactions. **e**, Five b3f3 reactions for **17** that exhibit lower activation barriers compared with the lowest barrier b2f2 reaction, including the Diels-Alder reaction (top). Activation energies are reported in kcal/mol and additional technical details for this comparison are reported in Section 2 of the Supporting Information.