nature

https://doi.org/10.1038/s41586-023-06854-3

Accelerated Article Preview

Computational prediction of complex cationic rearrangement outcomes

Received: 5 July 2022

Accepted: 8 November 2023

Accelerated Article Preview

Cite this article as: Klucznik, T. et al. Computational prediction of complex cationic rearrangement outcomes. *Nature* https://doi.org/10.1038/s41586-023-06854-3 (2023)

Tomasz Klucznik, Leonidas-Dimitrios Syntrivanis, Sebastian Baś, Barbara Mikulak-Klucznik, Martyna Moskal, Sara Szymkuć, Jacek Mlynarski, Louis Gadina, Wiktor Beker, Martin D. Burke, Konrad Tiefenbacher & Bartosz A. Grzybowski

Computational prediction of complex cationic rearrangement outcomes

Tomasz Klucznik[1,2†], Leonidas-Dimitrios Syntrivanis[3,4†*], Sebastian Baś[2,5], Barbara Mikulak-Klucznik[1,2], Martyna Moskal[1], Sara Szymkuć[1], Jacek Mlynarski[2], Louis Gadina[2], Wiktor Beker[1,2†*], Martin D. Burke[3,6,7,8,9*], Konrad Tiefenbacher[4,10*] & Bartosz A. Grzybowski[1,2,11,12*]

[1] Allchemy, Highland, IN, USA

[2] Institute of Organic Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, Warsaw 01-224, Poland

[3] Roger Adams Laboratory, School of Chemical Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, USA

[4] Department of Chemistry, University of Basel, 4058 Basel, Switzerland

[5] Faculty of Chemistry, Jagiellonian University, Gronostajowa 2, 30-387 Krakow, Poland

[6] Molecule Maker Lab Institute, University of Illinois at Urbana-Champaign, Urbana, IL, USA

[7] Molecule Maker Lab at the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA

[8] Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA

[9] Carle Illinois College of Medicine, University of Illinois at Urbana-Champaign, Urbana, IL, USA

[10] Department of Biosystems Science and Engineering, ETH Zurich, 4058 Basel, Switzerland

[11] IBS Center for Algorithmic and Robotized Synthesis, CARS, 50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan, 689-798, South Korea

[12] Department of Chemistry, UNIST, 50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan, 689-798, South Korea

[†] These authors contributed equally

*Correspondence to: l.syntrivanis@unibas.ch, wiktor.l.beker@gmail.com, mdburke@illinois.edu, konrad.tiefenbacher@unibas.ch, or nanogrzybowski@gmail.com

Recent years have seen revived interest in computer-assisted organic synthesis[1,2]. The use of reaction-network and neural-network algorithms which can plan multi-step synthetic pathways have revolutionized this field[1,3-7], including examples leading to advanced natural products[6,7]. Such methods typically operate on full, literature-derived "substrate(s)-to-product" reaction rules and cannot be easily extended to the analysis of reaction mechanisms. Here we show that computers equipped with a comprehensive knowledge-base of mechanistic steps augmented by physical-organic chemistry rules, as well as quantum mechanical (QM) and kinetic calculations, can use a reaction-network approach to analyze the mechanisms of some of the most complex organic transformations – namely, cationic rearrangements. Such rearrangements are a cornerstone of organic chemistry textbooks and entail dramatic changes in the molecule's carbon skeleton[8-12]. The algorithm we describe and deploy at https://HopCat.allchemy.net/ generates, within minutes, networks of possible mechanistic steps, traces plausible step sequences, and calculates expected product distributions. We validate this algorithm by three sets of experiments whose analysis would likely prove challenging even to highly trained chemists: (i) predicting the outcomes of Tail-to-Head Terpene (THT) cyclizations in which substantially different outcomes are encoded in modular precursors differing in minute structural details; (ii) comparing the outcome of THT cyclizations in solution or in a supramolecular capsule, and (iii) analyzing complex reaction mixtures. Our results support a vision in which computers no longer just manipulate

known reaction types[1-7] but will help rationalize and discover new, mechanistically complex transformations.

Reactions involving sequences of carbocation rearrangements are important in both biosynthesis[13,14] as well as organic synthesis, allowing for drastic reorganization of a complex scaffold hard to achieve by other methods. Such reactions have been used as key steps of some classic total syntheses[15,16] but their use remains limited as outcomes are not readily predictable, especially for unexplored substrates. To date, most approaches relied on quantum mechanical calculations – these can be highly accurate and can account for nuanced effects such as post-transition-state, p-TS, bifurcations[12] but can be computationally very costly[11,17] and have been used mostly to substantiate plausible pathways and products[11,17] rather than predict them *de novo* (but see [18,19]). Approximate, graph- and rule-based models are significantly faster but have often been limited in scope[20,21], prone to false-positive predictions, and not able to solve even moderate-complexity problems[22,23] (see Supplementary **Section S6**), though some proved useful in scaffold enumeration[24,25]. The difficulty in generalizing and applying these methods "on-the-fly," to arbitrary scaffolds appears two-fold. On one hand, quantum approaches may be "too-fine" as the number of degrees of freedom may be too large to consider the problem *a priori*, particularly for larger systems and multi-step pathways. On the other hand, coarse-grained models may be "too-crude" to properly define all intricacies (strain, stereochemistry, etc.) of individual mechanistic steps and apply them only in appropriate situations, to yield realistic intermediates (Supplementary **Sections S6.1** and **S6.2**). We reasoned that the problem may become tractable at an intermediate level of description, one in which an adequately broad yet chemically accurate set of mechanistic and physical-organic rules limits the solution space to a *network* of mechanistic steps on which finer, quantum-level analyses are then performed.

Curation of mechanistic steps. With the above considerations in mind, our first objective was to catalogue all or nearly all universally accepted mechanistic steps with which even very complex cationic rearrangements could be explained. Since mechanistic steps are not reported in repositories such as Reaxys or SciFinder (and cannot be just downloaded or machine-learned) we curated, over the years, a "training set" set of 715 solution-based (that is, not gas-phase) examples, mostly from advanced-level total syntheses and involving diverse scaffolds. For each of these 715 reactions, we performed detailed mechanistic analysis, mapping individual atoms and assigning each mechanistic step to a particular class (**Figure 1a** and all other examples posted at https://HopCatResults.allchemy.net/). Although the mechanisms are only chemically "plausible" (at the level of familiar "arrow-pushing"[26]), they allow us to ascertain that all 715 sequences could be written out using a finite set of commonly accepted mechanistic "transforms". Notably, this set was limited and, as we kept analyzing the 715 reactions, the numbers of newly added transforms steadily decreased (**Figure 1b**). Ultimately, our collection comprised 38 distinct ways of carbocation generation, 58 transforms for rearrangements and resonances, and 53 rules for cation quenching. We then followed the same strategy to curate two more collections – one set comprising 310 mechanistic "emergent" steps that are not (yet) commonly used in total syntheses (but enable retro alkyne *exo* cyclizations, 1,2 shifts for vinyl carbocations, etc.), and the other grouping 30 steps specific to biosynthetic carbocationic rearrangements. All transforms are detailed in Supplementary **Section S2**.

Rule encoding and additional constraints. Next, all transforms were encoded in the SMARTS notation as described in our previous works[1,3,6,27]. However, they encompass only a few atoms and are unaware of "broader" structures of (complex and diverse) scaffolds to which they may be applied – consequently, they can yield highly strained or structurally nonsensical products,

or may not account for subtler effects such as p-TS bifurcations. To remedy this, we implemented multiple constraints grounded in physical-organic considerations. The first group (Supplementary **Section S3.1**) evaluates transformation products and prevents the formation of, e.g., highly strained bridge carbocations, most primary carbocations (with exception of those formed in 1,3-olefin *exo* cyclization, retro-1,3-olefin *exo* cyclization, or allyl resonance), certain types of micro and macrocycles, as well as any other forbidden, structurally improbable motifs (e.g., 3-, 4-membered rings with unsaturations, or a tetrahedrane carbocation). The second group (Supplementary **Section S3.2**) applies to specific reaction types and inspects whether the substrate is properly preorganized for the reaction to occur (e.g., in 1,5-*H* shifts or 1,4-*oxa* cyclizations), whether participating rings are properly activated (e.g., in aryl cyclizations), or whether certain substituents meet bulkiness criteria (e.g., in silyl β-elimination). The third group (Supplementary **Section S3.3**) considers stereochemistry to prevent, e.g., cyclizations of ring substituents *trans* to each other, cyclizations from unsuitable E/Z configurations of double-bond systems, cyclizations "knitting though" larger rings, or hydride or alkyl migrations to different faces of a ring. Finally, the fourth group (**Methods** and Supplementary **Section S3.4)** regards p-TS bifurcations and aims to eliminate steps that cannot occur due to nonequilibrium or dynamic effects[12,28].

Generation of mechanistic networks. Together, the transforms and constraints enable the generation of networks of mechanistic steps within an Allchemy-based[29] "HopCat" webapp (https://HopCat.allchemy.net/; for user manual, see Supplementary **Section S1** and for illustration, Supplementary **Video 1**). In default settings, the program uses the "core" set of total-synthesis-oriented transforms and, optionally, also the 310 "emergent" and/or 30 biosynthetic ones. In a typical scenario, the software takes as input a starting carbocation (if a neutral molecule is input, possible starting carbocations are suggested to form generation $G_{n=0}$ of the network). Subsequently,

the algorithm considers the starting cation and its possible resonant structures, and applies to them matching rearrangement steps and constraints to produce the first generation, G_1, of "evolved" carbocations and their resonant forms. These molecules are subjected to the next round of rearrangements and also cation quenching steps to give G_2. The still "active," non-quenched carbocations can then be used to generate G_3 and the process is iterated until a user-specified limit of generations, G_{nmax}, is reached or all carbocations are quenched. At all stages, the structures are de-duplicated but if the same molecules can be reached via different routes, all such routes are kept. Of note, the user can optionally allow for the quenched cations to be regenerated (as in syntheses in refs. [30,31]); only one regeneration event is allowed per one route.

When the networks were propagated from the substrates of our 715 examples, all literature-reported products were identified within G_6 (95% within G_4 and ~50% within G_2). The networks varied in size from just teens to tens of thousands of nodes with calculations times on a multicore desktop ranging from seconds to several hours (typically, 1-10 min but, e.g., as long as 3 hrs for the 173,122-node network for homobrendane[32]). The network size did not correlate with the substrate's mass or the number of stereocenters but increased with the number of multiple non-aromatic bonds (**Figure 2a,b**) which allow for multiple resonance structures (which in turn, enable different rearrangements), and can also serve as source of electrons (e.g., in olefin cyclizations). Interestingly, irrespective of the substrate, the networks had a similar, branching structure: on average, each carbocation "branched out" into seven progeny carbocations (counting resonance structures) and three neutral/quenched products. These average branching factors did not depend on synthetic generation (**Figure 2c,d**). We also note that the application of the physical-organic constraints described in the previous section was absolutely essential as it reduced the network size at least several hundred times (and >1,300 times for the G_4 network of a cycloaraneosene

intermediate[33]). For many examples, networks without constraints were simply too large to be calculated on timescales of days.

Tracing of mechanistic pathways. Assuming the reaction product was found somewhere in the network, the mechanistic path(s) to the starting material were traced by a classic breadth-first search (in 312 out of 715 cases, these mechanistic solutions were unique). When the algorithm was tested on multiple examples from outside of the 715 literature set, it identified the routes proposed by the authors of these works (**Extended Figures 1,2** and Supplementary **Section S5.2**), provided solutions to problems that previously lacked mechanistic explanation (e.g., ref [34] and **Figure 3**), and suggested plausible mechanisms for biosynthetic rearrangements (**Extended Figure 3**).

Estimating product distributions. Next, we extended the algorithm to the challenging and practically important situation in which only the substrate and reaction conditions are given but the reaction outcome is unknown – that is, we wished to predict the main product and possible by-products, and also estimate their distributions under given reaction conditions. This required energetic and kinetic calculations for the network's molecules and reactions. While different theoretical approaches can be envisioned, we implemented methods which – without compromising accuracy too much – could perform network-wide calculations within times commensurate with the expectations of software's users (minutes) and would not require expensive licenses. These calculations are applicable to the "core" set of transforms for which many parameters can be adopted from previous studies. With the entire workflow detailed in the **Methods** section, its key stages were:

(i) Augmentation of the initial network of directed (i.e., irreversible) edges with edges corresponding to applicable reverse transformations. Such augmentation introduced reversibility and removed non-physical "sinks" from the network.

(ii) Generation of the Lowest Energy Conformers, LECs, for all nodes in the network.

(iii) Generation of Near Attack Conformers[35], NACs, for steps requiring pronounced conformational changes (cyclizations, long-range *H*-shifts).

(iv) Calculation of conformers' energies by a semi-empirical method (e.g., PM6) and calculation of substrate vs. product energy differences, ΔE's, including additional substrates if present (e.g., for eliminations, we included base and protonated base in energy calculations).

(v) Estimation of steps' contributions to activation barriers parametrized mostly on available literature data and QM calculations of model systems. The only free parameters used in all examples discussed later were Hammond parameters A_g and B_g for olefin protonation and water elimination; these values were fitted at a single temperature against experimental data for linalool (cf. below).

(vi) Coarsening of the network to group resonant structures into single nodes, identification of nodes corresponding to local energetic minima, calculation of the lowest-energy paths between the minima, and calculation of the rate constants for all steps using the standard Eyring equation.

(vii) Numerical integration of the set of kinetic equations.

Of note, we implemented (i-vii) for rearrangements taking place in solution and also under nanoscopic confinement. In the latter case, the analyses were limited to conformers that fit into either a spherical or an ellipsoidal enclosure of user-specified dimensions. Individual steps were allowed only when the reacting motifs were within a certain distance (**Methods**). The impact of the environment (e.g., slower deprotonation/quench process due to locking the conjugated base on

the capsule's wall[36]) could also be modelled by changing the protonation/deprotonation (quench) barriers (see user manual in Supplementary **Section S1**).

Experimental validations. The performance of the model thus constructed was tested against both existing data from the literature and new experiments.

Retrospective data analyses. We first analyzed how many of the products experimentally observed in the 715 literature reactions discussed earlier were also amongst HopCat's top-*k* predictions. Such an analysis is obviously limited in that many, especially older, publications do not report the exact nature of the acid used to generate the cation, as many as 228 examples provide no information about reaction temperature or time, 202 have only temperature, and 109 only time; in addition, it is often not known if the isolated product was the major or just the desired one. In such cases, we assigned reasonable default values (*p*-toluenesulfonic acid, 298 K, 12h) or several such settings (e.g., 2 and 12 h from which we then took the "best" and "worst" result). Even with these uncertainties, with the theoretical approximations involved, and with no free parameters, the model did reasonably well, in ca. 70% of cases placing the literature-reported product within its top-10 predictions (higher for smaller networks and decreasing steadily with network size, **Figures 2e,f**).

Rearrangements of linalool and fenchol. Turning to our own experiments, we first focused on linalool (**1** in **Figure 4**) and fenchol (**2** in **Figure 4**), whose carbocationic rearrangements have been studied for decades and apparently in exhaustive detail[37-39]. Here, we were interested not only in whether network analyses would rediscover mechanisms leading to the known products but also whether they would (i) reproduce experimentally observed product distributions at different

temperatures and (ii) perhaps identify some additional products not identified in prior studies. They did.

For linalool, the G_4 network (**Figure 4a**) contained all experimentally observed products **3**-**10**. With only four Hammond A_g and B_g parameters fitted to experimental data only at a single temperature (T_{fit} = 80 °C), the model (1) rationalized the formation of trienes **11a** and **11b** not detected in previous studies (and in our experiments at T_{fit} seen in only trace amounts); (2) reproduced a switchover in product distribution around ~80 °C (from dominant **7** below this temperature to **5**, **10** above; **Figure 4b** and discussion of mechanism in Supplementary **Section S7**); and (3) did not predict false-positives among major products (at higher temperatures, *Z* isomer of **11b** – not detected in the experiment but likely a precursor to the *E* isomer of **11b** – was placed 5th, with abundance 0.6%; at lower temperatures, there were no false positives in the top-7 and down to ~ 0.1% abundance). For fenchol, experiments identified 12 products, all found within the network (**Figure 4c**) and with the first false-positive ranked 5th (with ~ 0.5% abundance). The exact abundances of products of longer paths were slightly underestimated at 50-170 °C but the agreement was better at higher temperatures (**Figure 4d**). Molecules **17** and **12** were not previously described as products of fenchol's rearrangement and, noteworthy, network analysis was instrumental in identifying **17** which was misidentified by the conventional Xcalibur/NIST assignment of the relevant GC-MS peak (see **Methods**).

Substrate-controlled terpene cyclizations. Arguably, the most challenging set of validations was to predict the outcomes of a series of novel tail-to-head terpene, THT, cyclizations. In biosynthetic settings, such cyclizations take place in a catalyst-controlled modality, whereby different terpene synthases can act on the same linear precursor to yield various complex terpene frameworks. As an alternative, we wished to encode different cyclization outcomes by structural

differences in the precursor molecules, with no individually tailored catalysts. Predictable conversion of structurally simple precursor molecules into complex scaffolds in this manner would represent a powerful approach for modular access to complex natural product structures. Here, information dictating the cyclization outcomes was pre-programmed into linear precursors, and, in turn, into building blocks from which those precursors are derived (**Extended Figure 5**) in a manner that might be amenable with modular automated synthesis[40,41]. The examples in **Figure 5a** are especially challenging to predict – even to experienced synthesis experts – because the precursors differ solely in the positions of one methyl group and/or one double bond. Moreover, we studied these cyclizations not only in solution but also in the supramolecular resorcinarene capsule (roughly ellipsoidal) previously shown to catalyze related rearrangements with a Brønsted acid co-catalyst[42] – rearrangements under such confinement are virtually intractable without computer's help because analysis of mechanistic pathways is coupled to the analysis of conformers that fit within the capsule's enclosure.

The networks originating from various precursors spanned 94 to 8284 nodes. Within these large spaces of potential outcomes, the algorithm performed well, as the products isolated in experiments were ranked, on average, 9th in solution and 7th in the capsules (for individual rankings, see **Figure 5a**; for mechanistic pathways, see Supplementary **Section S8**). Of note, the initial experiments yielded some products only in the capsule but not in solution (1 eq. HCl) – nonetheless, even with adjustments of protonation and deprotonation rate parameters, the algorithm persisted in suggesting that these products (e.g., **31**) or skeletons (e.g., **28**) are comparably likely to form in solution. These predictions turned out to be correct and the products were indeed observed when ionization of the precursors was effected with $BF_3 \times OEt_2$ instead of HCl. In the end, only product **29** was not formed

in solution and required the use of the capsule – this preference for the capsule was reflected by the algorithm's rankings.

Analyzing the poorest predictions (e.g., **28** ranked only 27th in solution and 20th in the capsule, **30** ranked 12th/6th or **32** ranked 12th/11th), we observed that their mechanistic pathways had several similar-probability "branchings" – i.e., several intermediates along these sequences could engage in "side" mechanistic steps having similar activation energies (**Figure 5b** and **Methods**). Naturally, energy calculations entail inherent error, which for the PM6 method used here, was estimated[43] at 4-8 kcal/mol. Therefore, we reasoned that the presence of "branchings" for which activation energies are within a few kcal/mol could be a rough measure of the uncertainty assigned to a given sequence and ranking prediction. For this metric, the Spearman correlation coefficient against predicted top-*k* values was ~0.65 and, when it was applied to THT cyclizations, it correctly assigned the highest uncertainties to **30**, **32** and **28** (**Figure 5b**).

Synthesis of rosadiene natural product. Finally, we tested a pair of seemingly very similar terpene precursors, **33** and **34** in **Figure 5c**. Although they differ only in the distal part of the molecule (colored light-gray in **34**), the outcomes were predicted to be markedly different. In particular, for compound **34**, the early 1,6-olefin *endo* cyclization was predicted to be followed by 1,2-*H* and 1,2-C shifts, before final elimination gives rosadiene **35** – a natural product previously obtained via 8,15-isopimaradiene rearrangement[44] or sclareol cyclization[45]. This prediction was ranked top-1 (solution) and top-1 (capsule) and the product was, indeed, obtained in both solution and capsule experiments in 33% yield.

Conclusions. Overall, these examples suggest that our "multiscale", network-QM algorithm can rapidly suggest plausible mechanisms and the most likely outcomes of non-trivial carbocationic rearrangements, and is capable of differentiating between minute structural

differences in the substrate molecules. We envision its unique applicability in three areas: (i) to guide spectral/chromatographic assignments of complex products and/or product mixtures (including authentication of such mixtures in food or fragrance industries); (ii) to study rearrangements under confinement (including microporous materials); and (iii) to systematically survey large numbers of automatically synthesized[40,41], linear precursors for productive THT cyclizations yielding unprecedented numbers of new scaffolds. In the future, the workflow could benefit from the use of more accurate QM methods (albeit at the expense of computing time) and could be adapted to other reaction classes, e.g., radical-based rearrangements.

Correspondence

Correspondence and requests for materials should be directed to l.syntrivanis@unibas.ch, wiktor.l.beker@gmail.com, mdburke@illinois.edu, konrad.tiefenbacher@unibas.ch, or nanogrzybowski@gmail.com

Acknowledgements

Development of all codes and algorithms described in this work was supported by internal funds of Allchemy, Inc. (to T.K., B.M.-K., M.M., S.S., W.B.). Experimental validations by S.B. and J.M. were supported in part by the Foundation for Polish Science (award TEAM/2017-4/38 to J.M.). Experimental validations by L.G. was supported by the National Science Center, Poland (grant Maestro, # 2018/30/A/ST5/00529). L.-D.S. received funding from the European Union's Framework Programme for Research and Innovation Horizon 2020 (2014-2020) under the Marie Skłodowska-Curie Grant Agreement No. 836024. M.D.B. gratefully acknowledges support from an NIH MIRA award (R35GM118185). K.T. gratefully acknowledges support from the NCCR

Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation. Analysis of pathways and writing of the paper by B.A.G. was supported by the Institute for Basic Science, Korea (Project Code IBS-R020-D1).

Author contributions

T.K. codified most of the mechanistic-step rules with the help from B.M.-K. and S.S. W.B. developed the network and kinetic codes and, with the help of T.K., the physical-organic constraints. L.-D.S., M.D.B. and K.T. conceived the substrate-controlled cyclization studies described in Figure 5. L.-D.S. carried out all synthesis and characterization for the substrate-controlled cyclization studies under supervision from M.D.B. and K.T. S.B. performed experiments under the supervision of J.M and B.A.G. L.G. helped with the identification and analysis of literature examples. M.M. developed the HopCat WebApp. B.A.G. conceived and supervised the project and wrote the paper with contributions of all co-authors.

Data availability

Mechanistic reaction rules, physical-organic methods, and the kinetic model are detailed in the main text, Methods, and the Supporting Information. All 715 atom-mapped mechanistic pathways from which the mechanistic steps were extracted are posted at https://HopCatResults.allchemy.net. Therein, networks propagated from the literature substrates are also deposited. Experimental details including spectroscopic data can be found in Supplementary Sections S7 and S8. We intend to update HopCat based on new literature findings; these improvements will be made available to software's users.

Code availability

The interactive HopCat web application allowing for calculations starting from arbitrary carbocations is freely available to academic users at https://HopCat.allchemy.net/ (given server capacity, to five concurrent academic users on a rolling basis and two-week slots). HopCat's pseudocode is provided in Supplementary Section S3. Code for the calculation of conformers under confinement is deposited at https://github.com/Nanotekton/ellipsoid_cavity.

Supplementary Information

Supplementary Information accompanies this paper and includes additional theoretical, synthetic, spectroscopic, and chromatographic details.

Competing interests

The authors declare the following competing interests: T.K., W.B., B.M.K., M.M., S.S. and B.A.G. are consultants and/or stakeholders of Allchemy, Inc. Allchemy software and its HopCat module are property of Allchemy, Inc., USA. All queries about access options to Allchemy, including academic collaborations, should be sent to saraszymkuc@allchemy.net.

References:

1. Szymkuć, S. et al. Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016).

2. Corey, E. J. & Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science* **166**, 178–192 (1969).

3. Klucznik, T. et al. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* **4**, 522–532 (2018).

4. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).

5. Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **51**, 1281–1289 (2018).

6. Mikulak-Klucznik, B. et al. Computational planning of the synthesis of complex natural products. *Nature* **588**, 83-88 (2020).

7. Lin, Y., Zhang, R., Wang, D., & Cernak, T. Computer-aided key step generation in alkaloid total synthesis. *Science* **379**, 453-457 (2023).

8. Tantillo, D. J. Biosynthesis via carbocations: Theoretical studies on terpene formation. *Nat. Prod. Rep.* **28**, 1035-1053 (2011).

9. Christianson, D. W. Structural biology and chemistry of the terpenoid cyclases. *Chem. Rev.* **106**, 3412-3442, (2006).

10. Olah, G. My search for carbocations and their role in chemistry (Nobel Lecture). *Angew. Chem. Int. Ed. Eng.* **34**, 1393–1405 (1995).

11. Reis, M. C., Lopez, C. S., Faza, O. N. & Tantillo, D. J. Pushing the limits of concertedness. A waltz of wandering carbocations. *Chem. Sci.* **10**, 2159-2170 (2019).

12. Hare, S. R. & Tantillo, D.J. Post-transition state bifurcations gain momentum – current state of the field. *Pure Appl. Chem.* **89(6)**, 679–698 (2017).

13. Breitmaier, E. *Terpenes*. (Wiley-VCH Verlag GmbH & Co. KGaA, 2006).

14. Hong, Y. J. & Tantillo, D. J. The taxadiene-forming carbocation cascade. *J. Am. Chem. Soc.* **133**, 18249–18256 (2011).

15. Surendra, K., Rajendar, G. & Corey, E. J. Useful catalytic enantioselective cationic double annulation reactions initiated a tan internal π-bond: Method and applications. *J. Am. Chem. Soc.* **136**, 642-645 (2014).

16. Jørgensen, L. et al. 14-Step synthesis of (+)-ingenol from (+)-3-carene. *Science* **341**, 878-882 (2013).

17. Pemberton, R. P., Hong, Y. J. & Tantillo, D. J. Inherent dynamical preferences in carbocation rearrangements leading to terpene natural products. *Pure Appl. Chem.* **85**, 1949–1957 (2013).

18. Hare, S. R., Pemberton, R. P. & Tantillo, D. J. Navigating past a fork in the road: carbocation− π interactions can manipulate dynamic behavior of reactions facing post-transition-state bifurcations. *J. Am. Chem. Soc.* **139(22)**, 7485-7493 (2017).

19. Gutta, P. & Tantillo, D. J. Proton sandwiches: nonclassical carbocations with tetracoordinate protons. *Angew. Chem. Int. Ed.* **44(18)**, 2719-2723 (2005).

20. Gordeeva, E. V., Shcherbukhin, V. V. & Zefirov, N. S. The ICAR program: Computer-assisted investigation of carbocationic rearrangements. *Tetrah. Comp. Meth.* **3**, 429-443 (1990).

21. Gund, T. M., Schleyer, P. R., Gund, P. H. & Wipke, W. T. Computer assisted graph theoretical analysis of complex mechanistic problems in polycyclic hydrocarbons. The mechanism of diamantane formation from various pentacyclotetradecanes. *J. Am. Chem. Soc.* **97**, 743-751 (1975).

22. Chen, J. H. & Baldi, P. No electron left behind: a rule-based expert system to predict chemical reactions and reaction mechanisms. *J. Chem. Inf. Model.* **49**, 2034-2043 (2009).

23. Kayala, M. A. & Baldi, P. ReactionPredictor: Prediction of complex chemical reactions at the mechanistic level using machine learning. *J. Chem. Inf. Mod.* **51**, 2526-2540 (2012).

24. Tian, B., Poulter, C. D., & Jacobson, M. P. (2016). Defining the product chemical space of monoterpenoid synthases. *PLOS Comput. Biol.* **12**, e1005053 (2016).

25. Chow, J. Y. *et al*. Computational-guided discovery and characterization of a sesquiterpene synthase from Streptomyces clavuligerus. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 5661-5666 (2015).

26. Levy, D. E. *Arrow-pushing in organic chemistry: An easy approach to understanding reaction mechanisms*. (John Wiley & Sons, 2017).

27. Molga, K., Gajewska, E. P., Szymkuć, S. & Grzybowski, B. A. The logic of translating chemical knowledge into machine-processable forms: a modern playground for physical-organic chemistry. *React. Chem. Eng.* **4**, 1506–1521 (2019).

28. Hare, S. R. & Tantillo, D. J. Dynamic behavior of rearranging carbocations–implications for terpene biosynthesis. *Beilstein J. Org. Chem.* **12(1)**, 377-390 (2016).

29. Wołos, A. et al. Computer-designed repurposing of chemical wastes into drugs. *Nature* **604**, 668-676 (2022).

30. Jonathan, H. G., Baldwin, J. E. & Adlington, R. M. Enantiospecific, biosynthetically inspired formal total synthesis of (+)-liphagal. *Org. Lett.* **12**, 2394-2397 (2010).

31. Duc, D. K. M., Fetizon, M. & Lazare, S. A short synthesis of (+)-isophyllocladene and (+)-phyllocladene. *J. Chem. Soc., Chem. Commun.* **8**, 282 (1975).

32. Kasturi, T. R. & Chandra, R. Rearrangement of homobrendane derivatives. Total syntheses of racemic copacamphor, ylangocamphor, and their homologues. *J. Org. Chem.* **53**, 3178-3183 (1988).

33. Michalak, M., Michalak, K., Urbanczyk-Lipkowska, Z. & Wicha, J. Synthetic studies on dicyclopenta[a,d]cyclooctane terpenoids: Construction of the core structure of fusicoccins and ophiobolins on the route involving a Wagner-Meerwein rearrangement. *J. Org. Chem.* **76**, 7497-7509 (2011).

34. Hosoyama, H., Shigemori, H. & Kobayashi, J. Further unexpected boron trifluoride-catalyzed reactions of toxoids with α- and β-4,20-epoxides. *J. Chem. Soc., Perkin Trans. 1* **3**, 449-451 (2000).

35. Hur, S. & Bruice, T. C. Enzymes do what is expected (chalcone isomerase versus chorismate mutase). *J. Am. Chem. Soc*. **125**, 1472-1473 (2003).

36. Merget, S., Catti, L., Piccini, G., & Tiefenbacher, K. Requirements for terpene cyclizations inside the supramolecular resorcinarene capsule: bound water and its protonation determine the catalytic activity. *J. Am. Chem. Soc*. **142**, 4400-4410 (2020).

37. Zhang, Q. & Tiefenbacher, K. Terpene cyclization catalysed inside a self-assembled cavity. *Nat. Chem.* **7**, 197-202 (2015).

38. Lossing, F. P. & Holmes, J. L. Stabilization energy and ion size in carbocations in the gas phase. *J. Am. Chem. Soc.* **106**, 6917–6920 (1984).

420 39. Pulkkinen, E. Vedenlohkaisussa, F. & Toisiintumisista, T. *Suom. Kemistil. A* **30**, 239-245
421 (1957).

422 40. Junqi L. et al. Synthesis of many different types of organic small molecules using one
423 automated process. *Science* **347**, 1221-1226 (2015).

424 41. Blair, D. J. et al. Automated iterative Csp^3-C bond formation. *Nature* **604**, 92-97 (2022).

425 42. Zhang, Q., Rinkel, J., Goldfuss, B., Dickschat, J. S. & Tiefenbacher, K. Sesquiterpene
426 cyclizations catalysed inside the resorcinarene capsule and application in the short synthesis of
427 isolongifolene and isolongifolenone. *Nat. Catal.* **1**, 609-615 (2018).

428 43. Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: more modifications
429 to the NDDO approximations and re-optimization of parameters. *J Mol. Model.* **19**, 1–32 (2013).

430 44. McCreadie, T. & Overton, K. H. The conversion of labdadienols into pimara-and rosa-dienes.
431 *J. Chem. Soc. C*, 312-316 (1971).

432 45. Ungur, N. D., Barba, A. N. & Vlad, P. F. Cyclization and rearrangement of diterpenoids. VII.
433 Composition of the hydrocarbon fraction of a mixture of the products of cyclization of manool and
434 sclareol by ordinary acids. *Chem. Nat. Compd.* **24**, 612-614 (1988).

435 46. Wang, M., Wu, A., Pan, X., & Yang, H. Total synthesis of two naturally occurring
436 bicyclo[3.2.1]octanoid neolignans. *J. Org. Chem.* **67**, 5405–5407 (2002).

437 47. Kobayashi, J. & Shigemori, H. Bioactive taxoids from the Japanese yew Taxus cuspidate. *Med.*
438 *Res. Rev.* **22**, 305-328 (2002).

439 48. Schneider, F., Pan, L., Ottenbruch, M., List, T. & Gaich, T. The chemistry of nonclassical
440 taxane diterpene. *Acc. Chem. Res.* **54**, 2347-2360 (2021).

FIGURE LEGENDS

Figure 1. Key aspects of a network-based algorithm to predict mechanisms and product distributions of complex carbocationic rearrangements. a, One of the literature examples (from total synthesis of methyl Kadsurenin C[46]) analyzed by expert chemists to assign individual mechanistic steps. Bonds and atoms colored in red span the "cores" of mechanistic transforms. **b,** The horizontal axis counts numbers of literature examples analyzed (715 in total, all examples deposited at https://HopCatResults.allchemy.net/) to derive these rules. The vertical axis plots the number of mechanistic rules identified by analyzing a given number of literature examples (please note that certain rules are grouped, e.g., *Addition of water 1* and *Addition of water 2* are counted as one). Light green curve = carbocation generation rules; dark green = rearrangements and resonances; blue = carbocation quenches. For each set, analysis was repeated 10,000 times, each time with a different and random ordering of literature examples. Solid line represents the median; the dark and light shaded areas delineate interpercentile ranges 0.25-0.75 and 0.05-0.95, respectively. All curves flatten out suggesting that our sets of mechanistic rules are nearly complete (for all rules, see Supplementary **Section S2**). **c,** The mechanistic steps thus derived are applied iteratively to propagate reaction networks commencing from arbitrary substrates ("parent" node at the very bottom). **d,** The networks are pruned according to physical-organic constraints (Supplementary **Section S3**) to reduce network size by up to ~1000 times. **e,** Subsequently, the algorithm can trace (*orange*) mechanistic pathway(s) between the substrate and some known product. Already at this stage, the algorithm can solve some complex mechanistic puzzles (see **Figure 3** and **Extended Figures 1-3** and Supplementary **Figures S41-S66**). **d,** Calculation of energies of all nodes/molecules and energetic barriers of all edges/steps (cf. **Methods**) yields kinetic rate constants (here, colored *blue*-to-*red* to indicate slow-to-fast steps). Solution of kinetic

equations then predicts the abundances of specific products, here indicated by the sizes of the nodes.

Figure 2. Statistics of the mechanistic networks and model performance. All analyses are based on the of networks generated for the "715" set. *Orange* lines indicate median, boxes envelop data between Q_1 and Q_3 quartiles; whiskers delineate the most spread pair of points within [Q_1 - 1.5·*IR*, Q_3 + 1.5·*IR*] range, where $IR = Q_3$-Q_1. **a,b,** Sizes of mechanistic networks do not correlate with, e.g., substrate's molecular weight or the number of stereocenters **(a)** but increase with the number of multiple, nonaromatic bonds **(b)**. **c,d,** Average branching factors remain similar irrespective of network's synthetic generation, G_n. Branching factor for carbocations in a given G_n = the number of carbocations in G_{n+1} divided by their number in G_n. Quench factor = the number of quenched/neutral products in G_{n+1} divided by the number of carbocations in G_n. Performance of the kinetic model is quantified in **e,f**. In **e,** the horizontal axis quantifies the absolute rank (best = zero) of the literature-reported product within the network. The vertical axis gives the percentage of the entire dataset (i.e., all "715" networks) for which the predicted rank is not larger than the corresponding value on the *x*-axis (top-*k* statistic). "Default" settings use the default quench and generation parametrization with time and temperature either taken from literature or set to 298 K and 12 h. "Best" settings modify these parameters within 30% from the default values (0.7, 0.8, 1.2 and 1.3) and missing time data (here, 2h and 12h) that give the best top-*k* statistics. The "worst" settings correspond to the worst result obtained with such modifications; **f,** Bars and the left axis quantify the dependence of the top-10 statistics on the number of synthetic generations. The right axis and *black* line plot the average network size – as the networks become very large, the accuracy in predicting the literature product within the top-10 decreases markedly.

Figure 3. HopCat's analysis of a carbocationic rearrangement in Kobayashi's synthesis of a taxanine derivative[34]. a, Reaction of β-4(20)-epoxy-5-*O*-triethylsilyltaxinine A, T1, which upon treatment with $BF_3 \times OEt_2$ gives compound T3 containing a cyclobutane ring marked for clarity with green dotted bonds. Although the authors explained the mechanism for the **T1 → T2** step (Lewis-acid-promoted liberation of formaldehyde followed by elimination of silyl ether and formation of 1,3-dioxane), step **T2 → T3** was referred to as "curious"[47] and had no mechanistic explanation even in a recent review[48] (published 20 years after the original paper). **b,** Screenshot of a fragment of the 3,404-node network – generated within few minutes – illustrates HopCat's analysis starting with LA-activated T2. Resonances correspond to horizontal connections in each synthetic generation G_n; rearrangements are connections between generations. *Blue* nodes = carbocationic intermediates; *green* = quenched molecules. The mechanistic route to the experimentally observed product is traced by *purple* lines. Miniatures showing the intermediates are overlaid on the network (see also Supplementary **Figures S1-S9**). Note: an alternative but longer pathway replacing 1,4-olefin *exo* cyclization with a sequence of 1,5-olefin endo and 1,2-C shift, and then replacing carbonyl resonance and elimination of Lewis Acid with elimination and enolization was also found. For additional problems solved by HopCat on similar time scales, see **Extended Figures 1** and **2** and Supplementary **Figures S41-S43** and **S50-S66**. The Reader is encouraged to use https://HopCat.allchemy.net/ to solve other mechanistic riddles, starting from substrates/carbocations of their choice.

Figure 4. Experimental vs. predicted product distributions emerging from rearrangements of linalool and fenchol at different temperatures. a,b, are data for linalool for which

experimental conditions were: linalool **1** (0.65 mmol), TsOH•H_2O (0.65 mmol), MS 4Å, dry benzene, Ar, 16 h. **c,d,** are data for fenchol for which experimental conditions were: fenchol **2** (1.95 mmol), $KHSO_4$ (1.95 mmol), MS 4Å, neat, Ar, 16 h. For all experimental details, see Supplementary **Section S7. a,** and **c,** are screenshots of HopCat's networks, both propagated up to G_4. *Purple* nodes are products observed experimentally and node sizes correspond to relative abundances of the products (see also the second part of Supplementary **Video 1**). Previously unreported products are in red frames. **b,** and **d,** provide comparisons of experimental vs. predicted product distributions at different temperatures. Vertical axes quantify percentages of specific products in the reaction mixture (whenever applicable, in the model and in experiment, these percentages are sums of values for enantiomers and diastereoisomers). In the experiments, the crude mixture was analyzed by GC-MS (see **Methods** and Supplementary **Section S7** for details).

Figure 5. Experimental vs. predicted outcomes for the Tail-to-Head Terpene cyclizations and uncertainty of theoretical predictions. a, Table illustrating the building blocks (*light-blue* and *red* fragments) and the corresponding cyclized products, if observed. Methyl groups and the double bonds differing between the fragments are highlighted in gray. Within each "tile," *dark blue* = yields for solution experiments (determined by GC); *orange* = yields for experiments performed in the capsule (isolated yields). The algorithm-predicted "top-k" rankings are listed below the structures. Values in parentheses are rankings of the correct skeleton being formed (i.e., with all stereochemical information and all double bonds removed). Unless otherwise noted, capsule reaction conditions were 20 mol% supramolecular capsule, 3 mol% HCl, $CHCl_3$ solvent, and 40 °C. [a] GC yield. [b] 10 mol% of capsule used, and reaction carried out at 30 °C. [c] Product isolated after preparative scale reaction of alcohol substrate; in all these cases, the same major product is

observed in the reaction of the acetate substrate. [d] Substrate is an equimolar mixture of diastereomers. [e] HCl (1.0 equiv.) and [f] $BF_3 \times OEt_2$ (1.0 equiv.) solution conditions. **b,** Scheme of a similar-probability and thus high-uncertainty "branching" along a hypothetical fragment of a mechanistic route. In the graph, the vertical axis quantifies such kinetic uncertainty as the number of branchings (within 4 kcal/mol) normalized by the maximum value within the set. The horizontal axis plots the normalized numbers of products in each network whose energies are within 4 kcal/mol from the "correct" one; this thermodynamic measure of uncertainty is not predictive. **c,** Mechanistic routes from two precursors differing only in the distal part (marked in gray in the lower structure). For the lower precursor, rosadiene was predicted as the top-1 (solution)/top-1 (capsule) outcome – indeed, it was obtained in experiments in 33% yield. Note that despite seemingly similar precursors, the two mechanistic routes are markedly different.

Methods.

Physical-organic constraints. All constraints are discussed in detail in Supplementary **Section S3**.

Energy and kinetic calculations. The overall relationships between the HopCat's mechanistic networks and the Potential Energy Surface (PES) can be phrased as follows: each node corresponds to a single valence bond (VB) structure of a molecule, whereas the (directed) edges represent minimum energy paths (MEPs) between these points. Compared to high-level QM calculations, (i) PES's we consider do not take into account non-statistical dynamic effects directly (though certain patterns of this kind are included in a heuristic manner by elimination of edges corresponding to dynamically improbable routes, see section on post-transition state bifurcations below), and (ii) the concatenation of multiple "arrow-pushing" steps into concerted processes is performed in a simplified manner (see ***Graph transformation*** paragraph below) though these concatenations are not manifest in the WebApp which displays individual mechanistic steps.

In order to model the reaction kinetics within the network (and, thus, the distribution of products), one needs to determine a) energy function of nodes, and b) energy function of directed edges. Regarding (b), we define such a function as a non-negative estimate of the highest energy along the MEP between the starting and ending nodes, taken with respect to the energy of the starting node. For instance, for MEP containing a transition-state, *TS*, structure between nodes *a* and *b*, the corresponding edge energy is defined as $E_{TS} - E_a$, whereas for an "interpolating" (without *TS*) MEP, the edge energy is taken as $max(E_b - E_a, 0)$.

Choice of energy function. First, as a benchmark, we computed SCS-MP2/aug-cc-pVDZ energy profiles for a set of model mechanistic steps: 1,2-C shift, 1,2-*H* shift, addition of water to a

carbocation, and addition of a carbocation to an alkene. In the case of 1,2-shifts, we considered six cases differing in the order of the starting and ending carbocations. The results of these model calculations suggest that a) addition of a nucleophile (be it a lone pair of oxygen in water or double bond in alkene) follows the potential energy profile akin to bond dissociation (**Extended Figure 4d)** and b) that C/*H* shifts are effectively 'barrierless' when the stability difference between initial and ending carbocations is high (bottom panels in **Extended Figure 4b and 4c**). The stability difference between carbocations of consecutive orders was in the of range 10-20 kcal/mol, in agreement with experimental stabilities reported in ref. [49] (**Extended Figure 4e)**. In the case of symmetric (that is, when the orders of initial and final carbocations are the same), *H*-shift in 2,3-dimethylbutyl system (top right panel in **Extended Figure 4e)**, the barrier ~4 kcal/mol coincides with the value from[50].

Next, since DFT and *ab initio* calculations are too computationally demanding for high-throughput applications, in particular for energy calculations of hundreds or thousands of nodes within a single network, we tested energy profiles for these model systems using several semiempirical quantum mechanical (SQM) methods: xTB[49], PM6[50] and RM1[51] (as implemented in Open Mopac[52,55]). We also considered application of the neural network described in ref. [53], but it turned out to accept only neutral molecules as input. As can be seen in **Extended Figure 4**, all SQM methods either overestimate barriers or produce non-physical 'Transition States' along the paths, likely due to limitations originating from their underlying approximations (the minimal basis set and tight-binding or NDDO assumptions, which seem to work well in the vicinity of VB structures but fail in-between). The only exception is the xTB model, which performs well for our model *C*-shifts (**Extended Figure 4c**), but fails on some *H*-shifts (**Extended Figure 4b**).

Because of the abovementioned limitations, we could not use semiempirical calculations to detect possible TS points between connected VB structures. Instead, we proceeded with the construction of a phenomenological model, in which the edge energy function was defined as

$$E_{ab} = \max(E_b - E_a, P_{ab}) \quad (1)$$

where E_{ab} denotes edge energy, P_{ab} is a non-negative edge-specific energy penalty (accounting for either an energy barrier or energy related to conformational change required for the step to occur), and E_a and E_b correspond to energies of nodes *a* and *b*, respectively.

Since the edge energy function defined above explicitly depends on the node energy function, the next step was to benchmark SQM methods with respect to energy difference between the lowest energy conformers of the starting and ending carbocations of a mechanistic step. To do so, we selected 127 representative mechanistic steps involving diverse scaffolds (in particular, 33 reverse olefin cyclizations, 29 *oxa*-cyclizations, 28 *C*-shifts, 25 olefin cyclizations, 5 *H*-shifts, 4 alkyne cyclizations and 3 aryl cyclizations), For this set, we used SCS-MP2/cc-pVDZ method as a reference and compared it with corresponding SQM energies (**Extended Table 1**), which revealed the xTB and PM6 families of methods to provide the most accurate step energies.

Choice of parameters. With the overall objective to keep the number of adjustable parameters as small as possible, we assigned the following values (reasonable in the light of prior studies):

H- and C-shifts. Here, we assign energy penalties P_H and P_C to 6 kcal/mol which is close to the average of the values reported in ref. [50] for 1,2-*H*, 1,3-*H* and 1,4-*H* (4.96, 4.63 and 8.5

612 kcal/mol, respectively; see also **Extended Figure 4**). Furthermore, we assign separate P_{C3} and P_{Me}
613 parameters (set to 8 and 9 kcal/mol, respectively, based on experimental results obtained for
614 fenchol system) to 1,3-C and 1,2-methyl shifts.

615 ***Cyclizations.*** As addition of a nucleophile to a carbocation is expected to follow the bond
616 dissociation curve when the corresponding atoms are in proximity, the remaining factor is the
617 energy penalty related to bringing the substrate into a near-attack conformation (NAC). The NAC
618 energy relative to the substrate constitutes the cyclization penalty P_{ab}^{NAC}. For reverse cyclizations,
619 we assume no penalty.

620 ***Generation steps.*** This process is assumed to be generally endoenergetic, and equation (1)
621 is no longer suitable. Instead, we compute activation barriers for generation steps based on the
622 Hammond postulate, which assumes a linear relationship between reaction energy and reaction
623 activation energy, $\Delta E_{act} = A_g \cdot (E_b - E_a) + B_g$, where the Hammond parameters A_g and B_g are
624 assigned to each generation rule g. We set the following A_g and B_g values: (0.6, -20) for olefin
625 protonation and (0.2, -10) for water elimination from alcohol (fitted against distribution of linalool
626 products at 80 °C). For other generation rules (e.g., carbonyl protonation) we simply estimate ΔE_{act}
627 by the reaction energy $E_b - E_a$ – in other words, we set (A_g, B_g) = (1, 0).

628 ***Quench steps.*** As the profiles of addition of several nucleophiles to carbocations
629 (**Extended Figure 4d**) suggest that such an association process does not have an energy barrier by
630 itself (in the sense of the Transition State Theory), its rate should be only influenced by diffusion,
631 desolvation effects, and concentration of base. We incorporated these effects into parameter P_q,
632 whose value we set to 8 kcal/mol.

633 ***Other cases.*** A double bond adjacent to the formal carbocation may change its E/Z
634 regiochemistry due to delocalization effect (allyl resonance), which makes rotation around this

bond possible. We estimate these barriers by 1) generation of an approximate transition-state conformer (corresponding to the dihedral angle of 90 degrees); 2) computing an energy barrier using a semiempirical method; and 3) linear transformation of this barrier. The coefficients for the last step were obtained by linear regression with B3LYP results for a set of allylic systems with different substitution patterns (see Supplementary **Section S3.5**). For the irreversible oxidation of dienes to aryls, we assigned P_{arom} = 26 kcal/mol (as reported in ref. [54]).

Network initialization. Preparation of Allchemy's mechanistic network for kinetic calculation entailed four steps: a) assignment of structures (LEC and NAC calculations) followed by removal of improper nodes; b) node energy calculations, c) assignment of edge energies (according to equations 1 and 2); and d) transformation to a kinetic graph.

LEC calculations. With RDKit's implementation of the distance matrix algorithm and MMFF94 force field, we generated and optimized 50 conformers (with force tolerance set to 0.01) for each node in the network and took the one with the lowest energy as LEC, saving its coordinates and MMFF94 energy. For calculations in capsules, 100 LECs were considered (100 is the maximum number of LECs that can be set in HopCat).

NAC calculations. For each transformation, the substrate's conformer resembling the product is calculated by bringing the reacting, non-hydrogen atoms to the distance of 3 Å and relaxing the rest of the molecule while keeping this distance fixed. Specifically, having identified substrate's atoms that should be in close proximity but prior to conformer generation, we set coordinates of these atoms to (0,0,0) and (0,0,3). Using RdKit, we generated 50 such conformers for every relevant node molecule (specifically, only substrates to cyclizations), and selected the lowest-energy one as the NAC. The MMFF94 energy with respect to the substrate was taken as

NAC energy (note: in some cases studied, e.g., for a model limonene molecule, MMFF94 energy value of 4.6 kcal/mol was closer to B3LYP/6-31+G** result of 3.5 kcal/mol than PM6, which provided only 2 kcal/mol). For calculations in capsules, 100 NAC's were considered (100 is the maximum number of NACs that can be set in HopCat).

Detection of 'improper' conformers. When the distance matrix algorithm failed to produce a valid conformer, it was assumed that the structure was impossible to construct without breaking bonds or flipping stereocenters. We detected two failure modes: either the function did not produce any conformer or the generated structure was entirely flat (one of xyz coordinates was zero for all atoms) despite having sp^3-hybridized atoms. All nodes and edges with such "improper" structures were collected and scheduled for removal. In a typical network, about 2% of nodes were removed.

Removal of 'improper' nodes. All "improper" nodes were removed from the node list. All edges (both incoming and outgoing) that were either beginning or ending in any of the removed nodes were also removed. A BFS algorithm was then ran starting from the initial carbocation (root of the network) to detect and remove any disconnected components (nodes that lost their connections to the rest of the network because of the removal of their ancestors).

Removal of "improper" edges. Edges for which NAC calculations of the substrates failed were removed from the edge list. Then, detection and removal of disconnected components was performed as described in the previous point.

SQM calculations. Single point calculations were performed with COSMO model of solvent (with dielectric constant set to 2.27, corresponding to benzene). When testing different methods, NDDO methods (PM6, RM1, etc.) were computed in OpenMopac[55]; whereas xTB was provided with its own package[51]. In the final version of HopCat, PM6 model was used.

Graph transformation. First, all nodes connected by resonance transformations were combined, as they represent the same physical entity. For E/Z isomerization, we checked dihedral angles to group together nodes representing the same conformation of an allyl system. Then, the "resonance" nodes within a given group were assigned to a single "supernode" representing the group's structure. The supernode inherits all connections from the constituent nodes. Next, we identified local minima (MIN) in the graph (this operation is required to properly define the system of kinetic equations). These minima are defined as nodes for which all outgoing edges have non-zero value of edge energies E_{ab}. Then, in order to define the rate constants for transitions between different MINs, corresponding lowest energy paths (LEPs) have to be found (as they give a leading contribution to the rate constant). MINs and LEPs define a kinetic graph, in which nodes correspond to MINs and edges to LEPs, with LEP energies related to the corresponding rate constants by the Eyring equation, $k = \frac{k_B T}{h} \exp\left(-\frac{\Delta E_{act}}{RT}\right)$. Since edge energies in the mechanistic graph are non-negative (eq. 1), the maximum energy along LEP – an approximation of Transition State between MINs – is simply the sum of individual edge energies. However, in order to avoid double-counting, the LEP connecting two MINs cannot pass through another MIN. Therefore, the algorithm to detect LEPs proceeded as follows:

(1) All edges *outgoing* from any of MINs were (temporarily) removed from the graph and stored elsewhere.

(2) For each MIN node *m*:

(2.1) the edges outgoing from *m* were reintroduced into the graph,

(2.2) LEPs connecting *m* to any other MIN node in the graph were found with Dijkstra algorithm,

(2.3) LEP energies for thus detected pairs of MIN nodes were transformed into rate

constants via the abovementioned Eyring equation and stored,

(2.4) The edges outgoing from *m* were once again removed from the graph.

Finally, for eliminations, we multiplied the corresponding rate constants by the number of symmetrically-equivalent hydrogens that can be abstracted leading to the same product. For instance, abstraction of a proton from a terminal methyl group will be three times as fast as abstraction from tertiary carbon atoms, e.g., C(CH)(C)[CH+]C>>CC(C)=CC.

Kinetic calculations. Once the kinetic graph is defined, we perform numerical integration[56] (using SciPy implementation of the backward differentiation method) with initial concentration vector set to 1 for the initial carbocation/substrate and 0 for other nodes in the network.

Calculations under confinement. For calculations mimicking nanoconfinement, e.g., within the supramolecular capsule we used in some of our experiments, we imposed geometry-specific constraints on the generated conformers. If the confinement could be approximated as spherical, we 1) restricted the upper bound of the initial distance matrix to the diameter of the sphere, then 2) applied a distance-geometry algorithm with such modified bounds, and finally 3) optimized the resulting geometry using MMFF94 force field with harmonic distance constraint applied to every atom, so as to keep the entire molecule inside the sphere (if possible). This last constraint was imposed by i) addition of a fixed reference point X_0 in the current geometric center of the molecule and ii) addition of an energy penalty of the form $k \sum_{a \in atoms\ outside\ sphere}(|X_a - X_0| - R)^2$, where R is the sphere's radius. After generation and optimization, we additionally removed conformers which a) exceeded the confinement boundary by more than 1 Å (this also means that we accepted conformers for which, for instance, one hydrogen atom was *only slightly* outside the target sphere/ellipsoid) or b) contained valence angles centered at each atom and deviating by >10 degrees

from the unconstrained structure. This pruning step was intended to remove unsuccessful optimization attempts as well as the structures that were unphysically squished or twisted during optimization.

In the case of ellipsoidal confinement, the first two steps were the same as in the spherical case, with the largest axis taken as an upper bound of the distance matrix. The third step was conceptually identical – we optimized the geometry with harmonic constraint with respect to the ellipsoid's surface and applied to the atoms outside the ellipsoid – but the mathematical complexity of the problem required certain modifications to the optimization procedure. First, we computed the minimum volume enclosing the ellipsoid (MVEE) of the generated conformer and used the result to rotate the molecule so as to align the axes of MVEE with the coordinate system (as we choose our confinement ellipsoid to be in standard orientation, that is, with principal axes oriented along x, y, and z axes in order of their length). Second, we expressed the MMFF94 energy with confinement penalty as a function of rotatable dihedral angles instead of atomic coordinates, thus reducing the overall number of variables. We then optimized this target function with SciPy implementation of COBYLA algorithm[57-59]. After generation and optimization, the set of conformers was pruned just as in the spherical case. Finally, we prohibited mechanistic steps that are geometrically disfavored under confinement – that is, those in which bond-forming atoms are separated by more than 4.6 Å in all generated conformers.

Treatment of post-transition state bifurcations (p-TSB). The post-Transition State bifurcation (p-TSB) involve a pair of mechanistic steps sharing a common transition state (TS) after which the minimum energy path "bifurcates" towards different products without additional barrier. In such a scenario, nonequilibrium or dynamic effects may effectively exclude one of these transitions in favor of the other. Typically, modelling of such phenomena requires high-level QM calculations

(possibly even *ab initio* molecular dynamics) – clearly, an approach that cannot be applied effectively to reaction networks consisting of thousands of nodes. Instead, we aimed to capture p-TSB effects at an approximate level of knowledge-based rules reflecting various energetic and/or structural features. First, based on the available literature (i.e., our set of 715 reactions spanning 4174 mechanistic steps, all deposited at https://HopCatResults.allchemy.net), we identified three types of mechanistic steps that i) were reported or postulated to proceed through a non-classical carbocation and ii) are present in our literature dataset in numbers allowing for meaningful analysis: olefin cyclization[60,61] (347 examples), 1,2-C shift starting from cyclobutylcarbinyl cation[62,63] (53 examples), and retro-1,3-olefin cyclization originating from cyclopropylcarbinyl cation[63] (80 examples). In the case of olefin cyclization – in which the two products arise from attacks on different atoms of the same C=C bond – we performed PM6 calculations that evidenced that in 96% of cases, a product with lower energy was preferred. This observation was encoded as a heuristic to eliminate higher-energy products; some additional sub-rules were applied to the remaining 4% of exceptions. For the remaining two classes, we also identified and encoded structural criteria (e.g., based on the difference in connectivities of carbon atoms in β-position to carbocation in cyclopropylcarbinyl and cyclobutylcarbinyl carbocations) dictating the preferred product. For the remaining types of steps/systems (e.g., pimarenyl cation, pimar-8-en-15-yl cation), we encoded the preferred products verbatim. All these rules are discussed in detail in Supplementary **Section 3.4**.

GC-MS analyses. In experiments with linalool and fenchol, the products were analyzed by GC-MS and assignments were proposed by Thermo Scientific Xcalibur 2.1 with NIST 08 MS Library software. These assignments were further corroborated by comparison against reference

compounds (purchased or synthesized separately) and available literature data, and by additional control experiments (for all experimental and spectroscopic details see Supplementary **Section S7**).

In the analysis of products of fenchol's rearrangements, Xcalibur/NIST's assignment of GC-MS peaks was correct for **12** (a tricyclene with a retention time of 8.65 min), but this software incorrectly suggested **17** (at 8.01 retention time) to be 1,5,5-trimethyl-3-propan-2-ylidenecyclohexane. However, the formation of this molecule was mechanistically unlikely as it would require an unrealistically long sequence of seven mechanistic steps (1,2-C shift, retro-1,5-olefin *exo* cyclization, 1,2-C shift, 1,2-*H* shift, 1,3-olefin *exo* cyclization, and retro-1,3-olefin *exo* cyclization, quench by elimination). Also, no products that branched off this path were experimentally observed which suggested a wrong signal assignment. This made us reconsider some of the reference compounds we synthesized, identifying the same retention time for bornylene. For this compound, HopCat (i) suggested a concise three-step sequence (1,2-C shift, 1,2-C shift, elimination), and (ii) indicated that the same network branch leads not only to **17** but also to experimentally-observed compounds ***p*-6** and **12** (see Supplementary **Sections S7.1** and **S7.2** for, respectively, the scheme of the network and for HopCat's screenshots of mechanistic pathways starting from linalool and fenchol). We highlight this example because it illustrates the benefits of combining conventional (GC-MS, Xcalibur) and mechanistic network analyses.

THT cyclization studies.

Synthesis of the cyclization substrates. Details of all syntheses are included in Supplementary **Section S8**. Briefly, the substrates were prepared by a Negishi coupling between an organozinc reagent (generated from the corresponding alkyl bromide) and a vinyl iodide. The alkyl bromide (3.0 equiv.) was dissolved in THF (0.4 M) and *t*-BuLi (1.6 M in hexanes, 6.0 equiv.) was added at −78 ºC. The solution was stirred at −78 ºC for 30 min, then cannulated (rinsed with

THF, 2 mL) to a flask containing $ZnCl_2$ (3.0 equiv.) suspended in THF (2 M). The resulting solution was stirred at rt for 20 min, then cannulated to a flask (rinsed with THF, 2 mL) containing the vinyl iodide (1.0 equiv.) and $Pd(PPh_3)_4$ (0.1 equiv.) dissolved in THF (0.6 M with respect to the vinyl iodide). The reaction mixture was shielded from light and stirred at rt for 18 h. Brine and EtOAc were then added, the layers separated and the aqueous layer further extracted with EtOAc. The combined organic layers were dried over Na_2SO_4 and the solvent removed *in vacuo*.

The crude residue was dissolved in THF (0.2 M) and TBAF was added (2.2 equiv.). The reaction mixture was stirred at rt until complete consumption of the starting material was observed by TLC (3 – 18 h). Brine and EtOAc were then added, the layers separated, and the aqueous layer further extracted with EtOAc. The combined organic layers were dried over Na_2SO_4, the solvent removed *in vacuo*, and the crude residue purified via flash chromatography (hexanes/EtOAc 95:5 $\rightarrow$ 9:1) to give the pure alcohol.

For the preparation of the corresponding acetate substrates, this alcohol (1.0 equiv.) was dissolved in DCM (0.2 M), and Et_3N (2.5 equiv.), N,N-dimethyl 4-aminopyridine (0.4 equiv.), and acetic anhydride (2.0 equiv.) were added. The reaction mixture was stirred at rt until complete consumption of the starting material was observed by TLC (1 – 3 h). HCl 1M aqueous solution was then added, the layers separated and the aqueous layer extracted with DCM. The combined organic layers were washed with brine, dried over Na_2SO_4 and the solvent removed in vacuo. The crude residue was purified via flash chromatography (hexanes/EtOAc 97:3) to give the pure acetate.

Cyclizations using the resorcinarene capsule catalyst. To the substrate dissolved in $CHCl_3$ (30 mM) was added the specified amount of resorcinarene capsule catalyst followed by an

HCl stock solution in $CDCl_3$ (3 mol%), and the mixture was stirred at the specified temperature. Once the reaction was judged to be complete by GC analysis, the solvent was partially removed in vacuo (350 mbar at 40 ºC for short timeframes of 10-15 min – longer evaporation times could lead to significant loss of the volatile sesquiterpene products) and the mixture was passed through a column of silica (eluting with pentane) to remove the capsule catalyst and polar byproducts (for reactions utilizing 20 mol% of the capsule catalyst, two such passages may be required), followed by column chromatography using $AgNO_3$-impregnated silica to isolate the product of the reaction. Procedures for each compound and characterization data are available in Supplementary **Section S8**.

Solution cyclizations using HCl. To the substrate (16.7 μmol, 1.00 eq) dissolved in $CDCl_3$ (480 – X μL, where X is the amount in μL of HCl stock solution in $CDCl_3$ to be added to the reaction, as determined after titration, *vide infra*) was added a *n*-decane stock solution in $CDCl_3$ (20 μL, 167 mmol l^{-1}, 3.34 μmol, 0.2 eq). An aliquot (approximately 10 μL) of the reaction mixture was diluted with 0.25 mL of hexane (containing 0.08% DMSO) and subjected to gas chromatographic analysis (initial sample). An HCl stock solution in $CDCl_3$ (X μL, 1.0 equiv.) was then added and the mixture was stirred in a closed vial at the specified (internal) temperature, and further samples were taken at the indicated times and analyzed by gas chromatography. Conversions and yields were calculated as described in our previous work[64].

Preparation and titration of HCl stock solution in $CDCl_3$. HCl stock solution in $CDCl_3$ was prepared by passing HCl gas, generated by the dropwise addition of concentrated H_2SO_4 to dry NaCl, through $CDCl_3$ for approximately 30 mins. The concentration of HCl in the resulting solution was determined as follows: HCl stock solution in $CDCl_3$ (100 μL) was added to a solution of phenol red in EtOH (0.002 wt%, 2.5 mL) via a Microman M1 pipette equipped with plastic tips. Upon addition, the solution turned from yellow (neutral) to pink (acidic). The resulting solution

was then titrated with a 0.100 M ethanolic solution of triethylamine. At the equivalence point, the solution turned from pink to yellow. The HCl stock solution was kept in the fridge, and the titration was repeated immediately before each use.

Solution cyclizations using $BF_3{\times}OEt_2$. To the substrate (16.7 μmol, 1.0 equiv.) dissolved in $CDCl_3$ (310 μL) was added a *n*-decane stock solution in $CDCl_3$ (20 μL, 167 mmol l^{-1}, 3.34 μmol, 0.2 equiv.). An aliquot (approximately 10 μL) of the reaction mixture was diluted with 0.25 mL of hexane (initial sample) and subjected to gas chromatographic analysis. A $BF_3{\times}OEt_2$ stock solution in $CDCl_3$ (0.1 M, 170 μL, 170 μmol, 1.0 equiv.) was then added and the mixture was stirred in a closed vial at the specified (internal) temperature, and further samples were taken at the indicated times and analyzed by gas chromatography. Conversions and yields were calculated as described in our previous work[64].

Methods' references:

49. Lossing, F. & Holmes, J. Stabilization energy and ion size in carbocations in the gas phase. *J. Am. Chem. Soc.* **106**, 6917-6920 (1984).

50. Vrček, I., Vrček, V. & Siehl, H. Quantum chemical study of degenerate hydride shifts in acyclic tertiary carbocations. *J. Phys. Chem. A* **106**, 1604-1611 (2002).

51. Bannwarth, C. et al. Extended tight-binding quantum chemistry methods. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **11**, e1493 (2021).

52. Stewart, J. J. P. Optimization of parameters for semiempirical methods. V. Modification of NDDO approximations and application to 70 elements, *J. Mol. Model.* **13,** 1173-213 (2007).

53. Atz, K., Isert, C., Böcker, M., Jiménez-Luna, J., & Schneider, G. (2021). Open-source Δ-quantum machine learning for medicinal chemistry. *ChemRxiv*. DOI 10.26434/chemrxiv-2021-fz6v7-v2

54. Cristiano, M. et al. Investigations into the mechanism of action of nitrobenzene as a mild dehydrogenating agent under acid-catalysed conditions. *Org. Biomol. Chem.* **1**, 565-574 (2003).

55. The modern open-source version of the Molecular Orbital PACkage (MOPAC); https://github.com/openmopac/mopac

56. Shampine, L. F., Reichelt, M. W. The Matlab Ode Suite. *SIAM J. Sci. Comput.* **18**, 1-22 (1997).

57. Powell, M J D. *A direct search optimization method that models the objective and constraint functions by linear interpolation.* (Springer, Netherlands 1994).

58. Gomez, S. & Hennart, J-P. *Advances in Optimization and Numerical Analysis.* (Springer Science & Business Media, 2013).

879 59. Powell, M. J. A view of algorithms for optimization without derivatives. *Math. Today Bull.*
880 *Inst. Math. Appl.* **43**, 170-174 (2007).

881 60. Gutierrez, O. et al. Carbonium vs. carbenium ion-like transition state geometries for carbocation
882 cyclization–how strain associated with bridging affects 5-exo vs. 6-endo selectivity. *Chem. Sci.* **4**,
883 3894-3898 (2013).

884 61. Pemberton, R. P. & Tantillo, D. J. Lifetimes of carbocations encountered along reaction
885 coordinates for terpene formation. *Chem. Sci.* **5**, 3301-3308 (2014).

886 62. Olah, G. A., Jeuell, C. L., Kelly, D. P., & Porter, R. D. Stable carbocations. CXIV. Structure
887 of cyclopropylcarbinyl and cyclobutyl cations. *J. Am. Chem. Soc*. **94**, 146-156 (1972).

888 63. Barkash, V. A. & Shubin, V. G. *Contemporary Problems in Carbonium Ion Chemistry I/II*.
889 (Springer, 1984).

890 64. Zhang, Q. & Tiefenbacher, K. Terpene cyclization catalysed inside a self-assembled cavity.
891 *Nat. Chem.* **7**, 197-202 (2015).

892 65. Yokoo, K., Sakai, D. & Mori, K. Highly stereoselective synthesis of fused tatrahydropyrans
893 via Lewis-acid-promoted double C(sp^3)-H bond functionalization. *Org. Lett.* **22**, 5801-5805
894 (2020).

895 66. Cui, C. et al. Total synthesis and target identification of the curcusone diterpenes. *J. Am. Chem.*
896 *Soc*. **143**, 4379-4386 (2021).

897 67. Sato, H., Takagi, T., Miyamoto, K. & Uchiyama, M. Theoretical study on the mechanism of
898 spirocyclization in spiroviolene biosynthesis. *Chem. Pharm. Bull*. **69**, 1034-1038 (2021).

68. Lauterbach, L., Rinkel, J. & Dickschat, J. S. Two bacterial diterpene synthases from *Allokutzneria albata* produce bonnadiene, phomopsene, and allokutznerene. *Angew. Chem. Int. Ed.* **57**, 8280–8283 (2018).

69. Qin, B. *et al.* An unusual chimeric diterpene synthase from *Emericella variecolor* and its functional conversion into a sesterterpene synthase by domain swapping. *Angew. Chem. Int. Ed.* **55**, 1658–1661 (2016).

EXTENDED FIGURE LEGENDS

Extended Figure 1. **HopCat's mechanistic analysis of a reaction yielding a fused tetrahydropyran**. An example of a problem not "seen" by the machine during training on 715 literature examples. In the original publication[65], the authors focused on the double 1,5-*H* shifts as key steps and did not consider the full mechanism. HopCat's calculations ran up to $n = 4$ generations and traced a complete and unique mechanistic pathway. This pathway starts with a series of carbonyl and allyl resonances placing positive charge at the position available for 1,5-*H* shift followed by 1,6-olefin *exo* cyclization. Subsequently, the sequence of 1,5-*H* shift and 1,6-olefin *exo* cyclisation steps is repeated to afford tetrahydropyran's bicyclic scaffold. The last two mechanistic steps along the pathway are: (i) carbonyl resonance to form oxocarbenium species and (ii) elimination of Lewis acid yielding the final, quenched product. The software's solution agrees with the partial mechanism postulated in the original publication. **a,** A screenshot showing a simplified network (without stereochemistry). In reality, the network was generated with full stereochemistry and comprised of ~28,000 nodes that cannot be clearly visualized as a miniature. **b,** Details of all mechanistic steps (for raw screenshots from HopCat, in traditional and atom-mapped visualization modalities, see Supplementary **Section S5**). Additional examples are also provided in Supplementary **Section S5**.

Extended Figure 2. **HopCat's mechanistic analysis of a reaction leading to a tricyclic dienone.** HopCat solves another problem not "seen" in the 715 training set. The dienone is an intermediate used in the recent synthesis of curcusone diterpenes[66]. In the original publication, authors included a plausible arrow-pushing scheme of electron movements for the double deprotection-aldol

sequence but did not support it with a more detailed mechanistic analysis. HopCat identifies the reactions product in G_4 and proposes a plausible and unique mechanistic route. Starting from a carbocation generated via elimination of substrate's tertiary alcohol (bottom row of the network), this intermediate undergoes two consecutive resonances (allyl and carbonyl) that result in the formation of an oxocarbenium cation. Subsequent retro *oxa*-cyclization followed by ring closure constructs a seven-membered, central ring of the molecule. The last two steps describe deprotection of the enol ether. Formation of the oxocarbenium cation via carbonyl resonance makes the alkyl group on the oxygen a good leaving group, enabling its subsequent elimination and formation of the final product. The overall movement of electrons is consistent with the one proposed by the authors. **a,** A screenshot showing the network comprised of ~2,000 nodes. **b,** Details of all mechanistic steps (for raw screenshots from HopCat, in traditional and atom-mapped visualization modalities, see Supplementary **Section S5**). Additional examples are also provided in Supplementary **Section S5**.

Extended Figure 3. **A contested and only recently resolved[67] biosynthesis of spiroviolene relies on a macrocyclization step (1,11-olefin *endo* cyclization), which does not occur in abiotic set of carbocation transformations**. Identifying the mechanistic pathway for the biosynthesis of spiroviolene has proven a computationally challenging problem – in fact, the pathway was not found within G_7 and expansions to higher generations exceeded computing power. Accordingly, we implemented a "mixed" strategy search in which 7 generations were expanded from the substrate in the forward direction and 6 generations from the product in the retrosynthetic direction (using "reversed" mechanistic rules). This strategy considerably reduces the computational cost as the number of nodes in two smaller networks, each propagated to *n* generations and with branching

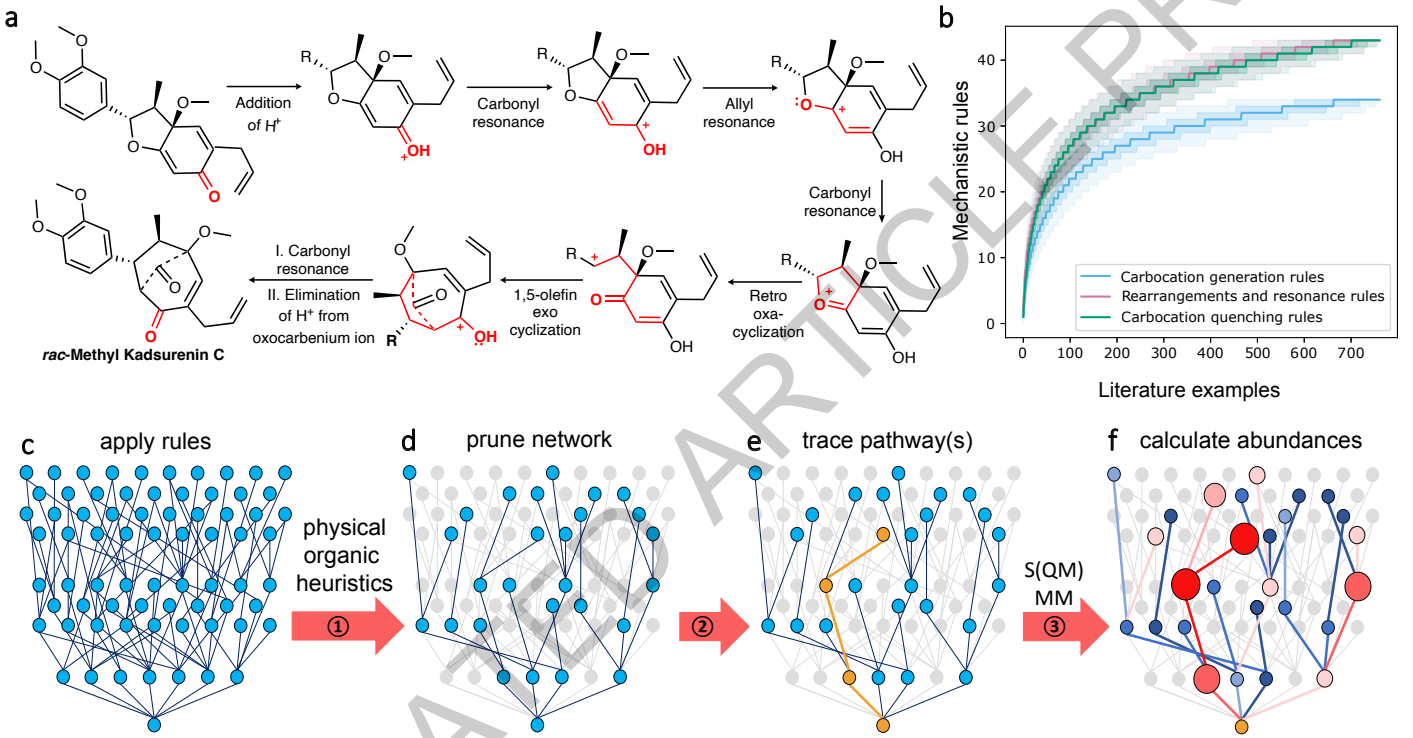
factor m, scales as $2m^n$ vs. m^{2n} for one forward network expanded to $2n$ generations (for $n = m = 7$, the difference is $m^n/2 \sim 400{,}000$ times). The algorithm then searched for common node(s) in the two networks and, when they were found, was able to concatenate a 10-step route. **a,** HopCat's screenshot showing a grossly simplified network generated by a mixed forward-retro search. In reality, the network comprised of 909,937 nodes that could not be clearly visualized as a miniature. HopCat's shortest route is marked with *purple* lines and agrees with the recently revised pathway[67]. Also, in the same network, rearrangement sequences leading to three other natural products were found – phomopsene[68] (*red* lines and frame), allokutznerene[68] (*orange*) and variediene[69] (*green*); **b,** Details of all mechanistic steps for spiroviolene's mechanistic route. For raw HopCat's screenshots of the sequences leading to all four natural products, in traditional and atom-mapped visualization modalities, see Supplementary **Section S5**). Note: Akin to Figure 4 and Extended Figures 1,2, none of the biosyntheses shown in this figure were considered when extracting mechanistic steps from literature examples.

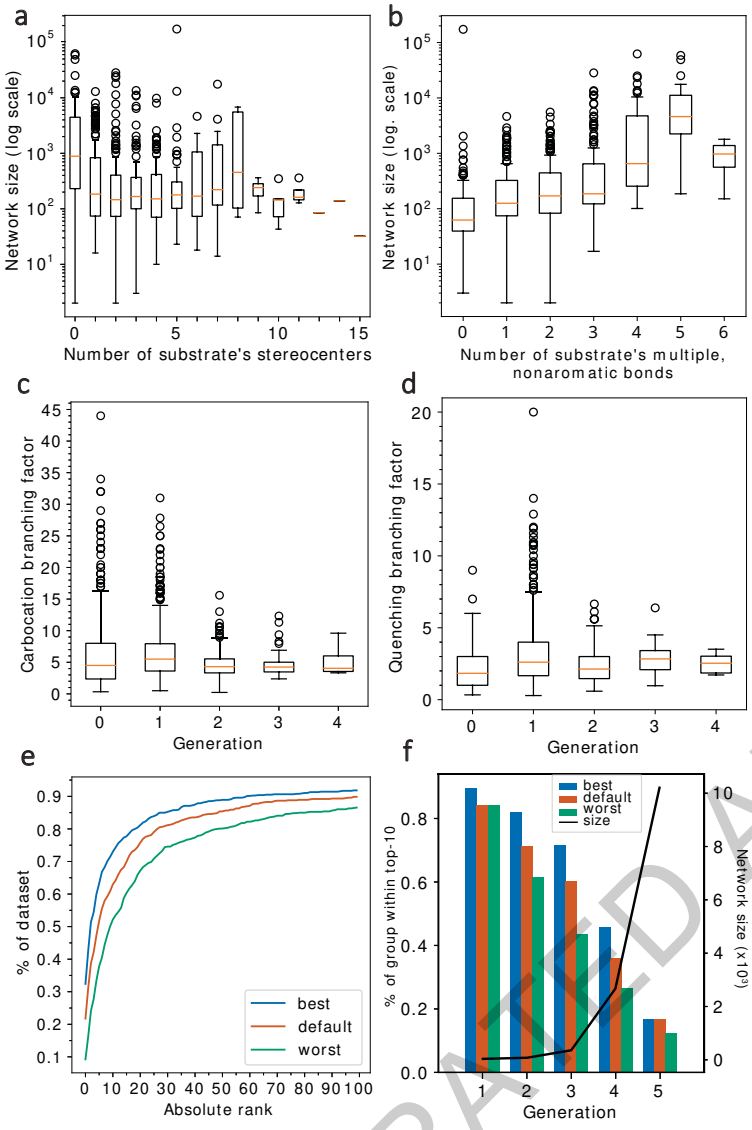
Extended Figure 4. Theoretical studies of model *H*- and *C*-shifts. a, System setup. For all unique configurations of substituents *R1*-*R4* (-H and -Me were considered), atom *X* was dragged along distance vector *r* so as to simulate the shift. Initial geometry was chosen such that the C-X bond was approximately perpendicular to the plane of the carbocation. All trajectories were subsequently verified by visual inspection. **b,** *H*-shifts ($X = H$). Top three panels represent symmetric shifts (such that the orders of initial and resulting carbocations are the same), with the order of carbocation increasing from the left to the right. In the bottom row, two leftmost panels represent shifts in which the carbocation changes order by one, whereas the rightmost panel represents an extreme example of transition between first-order and tertiary carbocations. **c,** C-shifts (X=Me). Top three panels

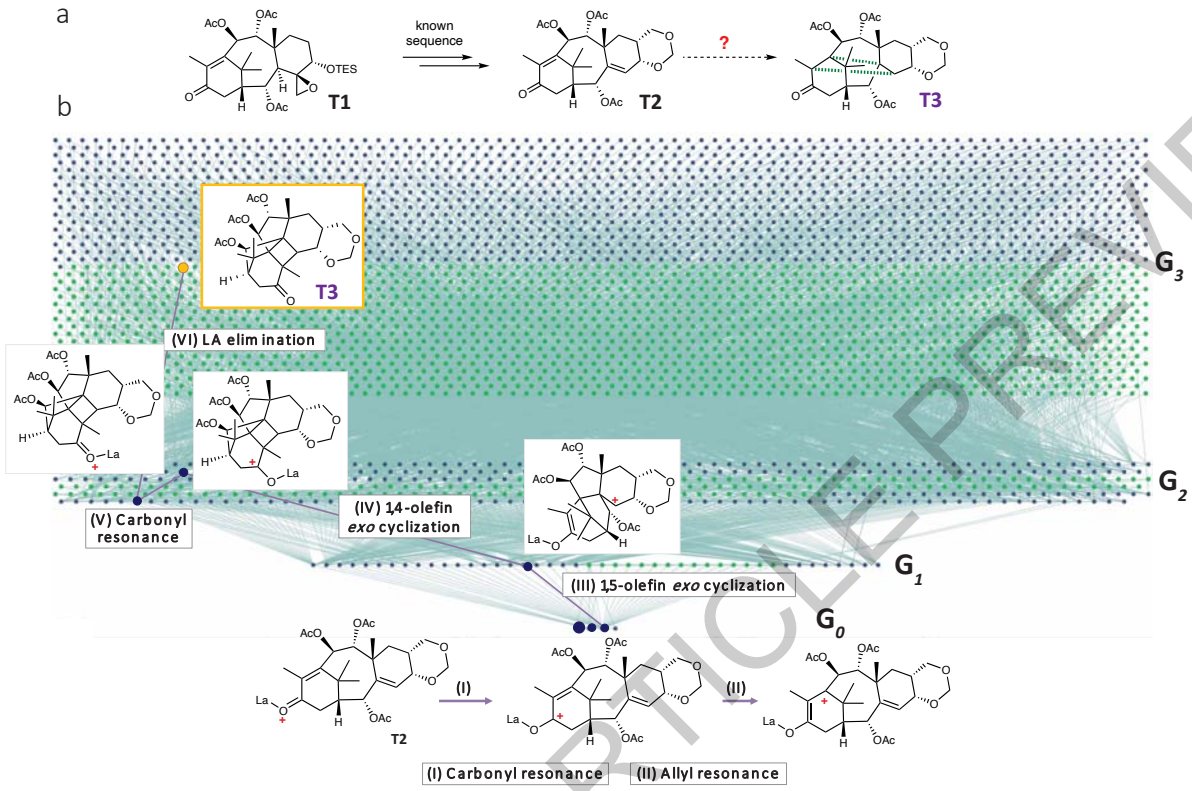
represent symmetric shifts (such that the orders of initial and resulting carbocations are the same), with the order of carbocation increasing from left to right. In the bottom row, two leftmost panels represent shifts in which the carbocation changes order by one, whereas the rightmost panel represents an extreme example of transition between first-order and tertiary carbocations. **d,** Theoretical studies of carbocation association process. Each curve represents the SCS-MP2/aug-cc-pVDZ energy profile with PCM model of water, modelling approach of four nucleophiles (formaldehyde, water, methanol and ethene) towards CH_3^+ along vector *R* (scheme inserted in the top left of the panel). **e,** Boxplot representing experimental stabilities of carbocations taken from[49] with respect to the CH_3^+ cation. The data was grouped according to the order of a carbocation (number of non-hydrogen atoms directly connected to the formally charged carbon atom), showing the general trend in the stability: increasing the order of a carbocation lowers the energy, on average, by 10-20 kcal/mol.

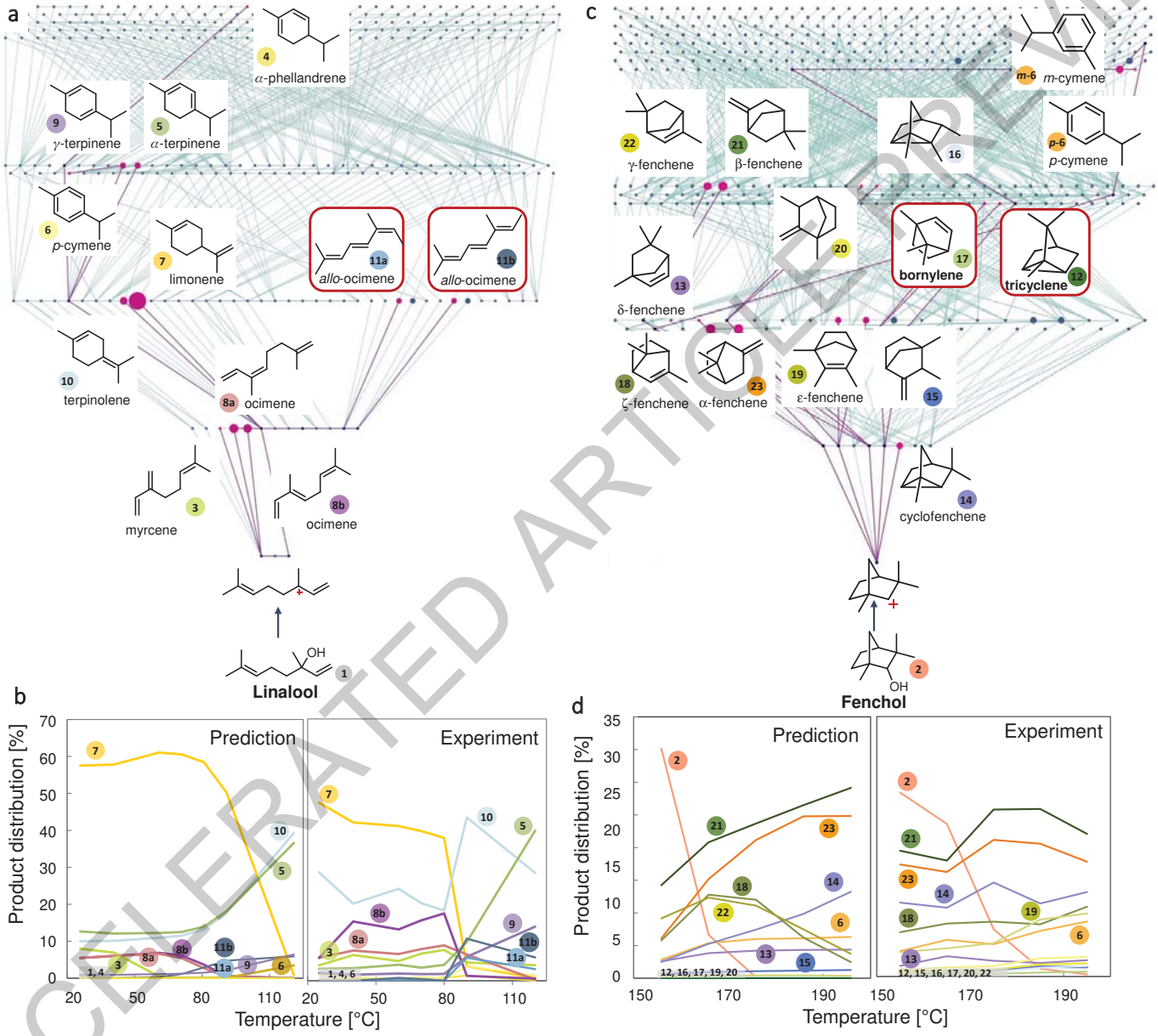
Extended Figure 5. General synthetic scheme for the preparation of the precursors employed in Figure 5. An alkyl bromide is converted into the corresponding organozinc reagent by sequential treatment with *t*-BuLi and $ZnCl_2$. This reagent is then used in a Negishi coupling with a vinyl iodide bearing a protected alcohol group. The coupling product is then deprotected to give the free alcohol, and the corresponding acetate is prepared by acetylation of this alcohol.

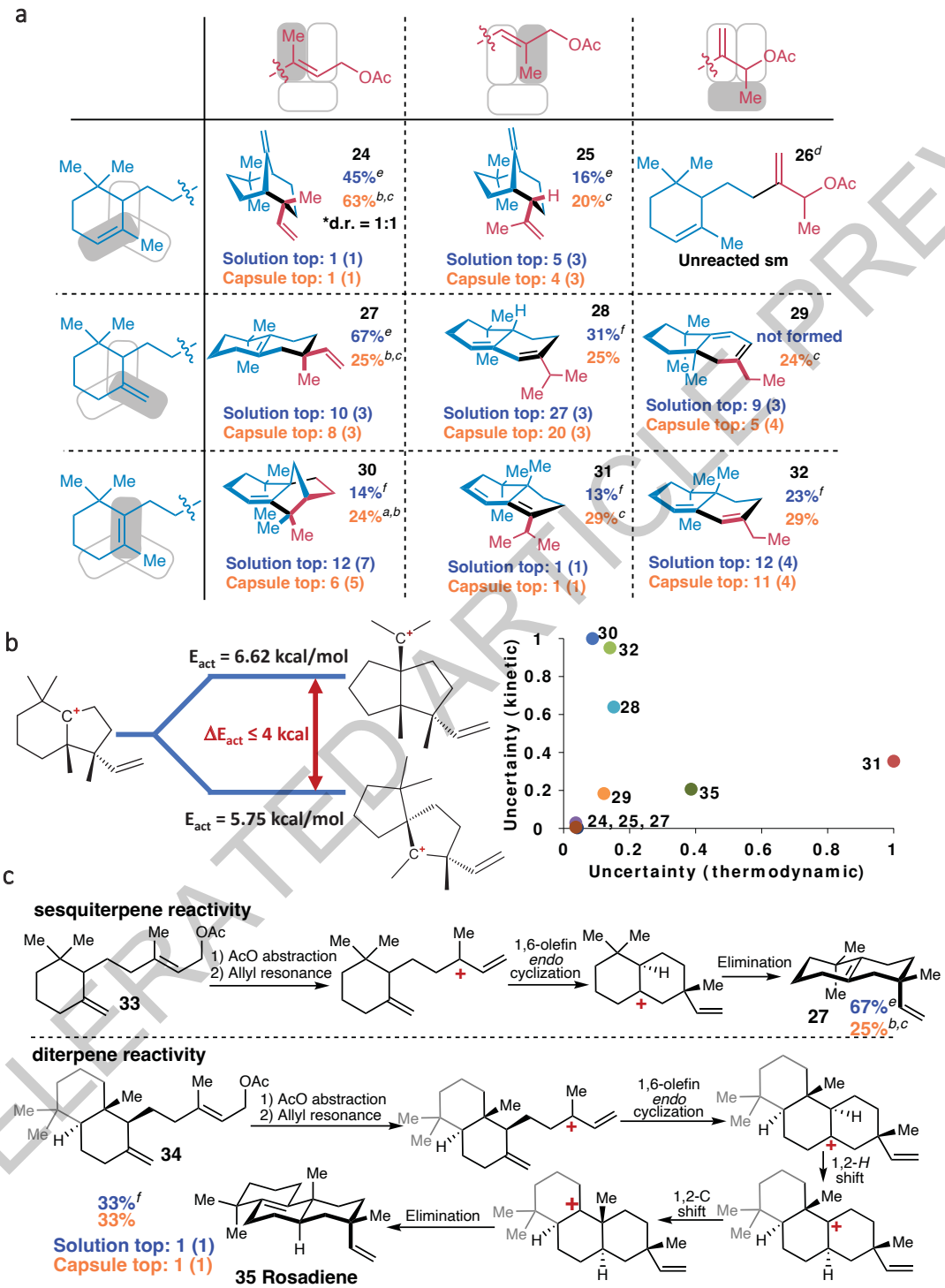
Extended Table 1. Accuracy of ΔE computed for representative rearrangement steps with semiempirical methods. SCS-MP2/aug-cc-pVDZ is taken as reference, ρ_R is Pearson correlation coefficient, ρ_S denotes Spearman rank correlation coefficient, and MAE is mean absolute error.

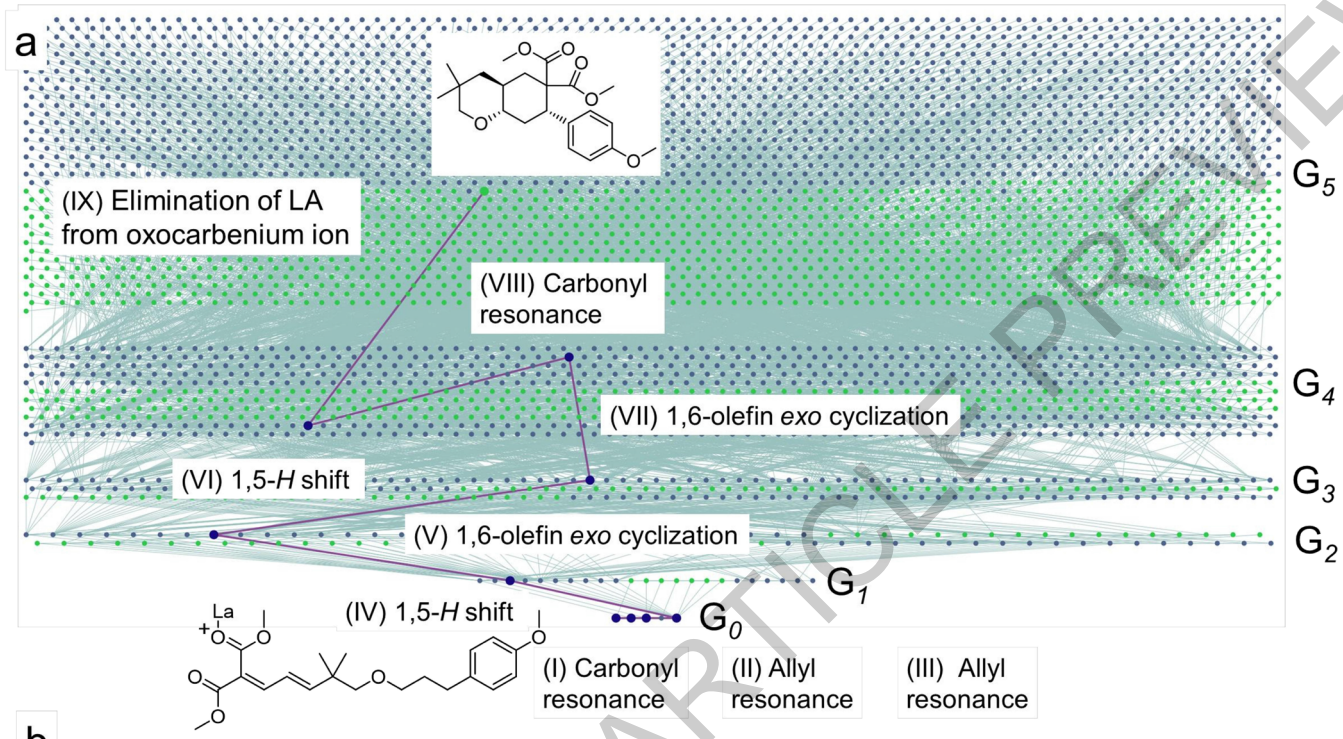


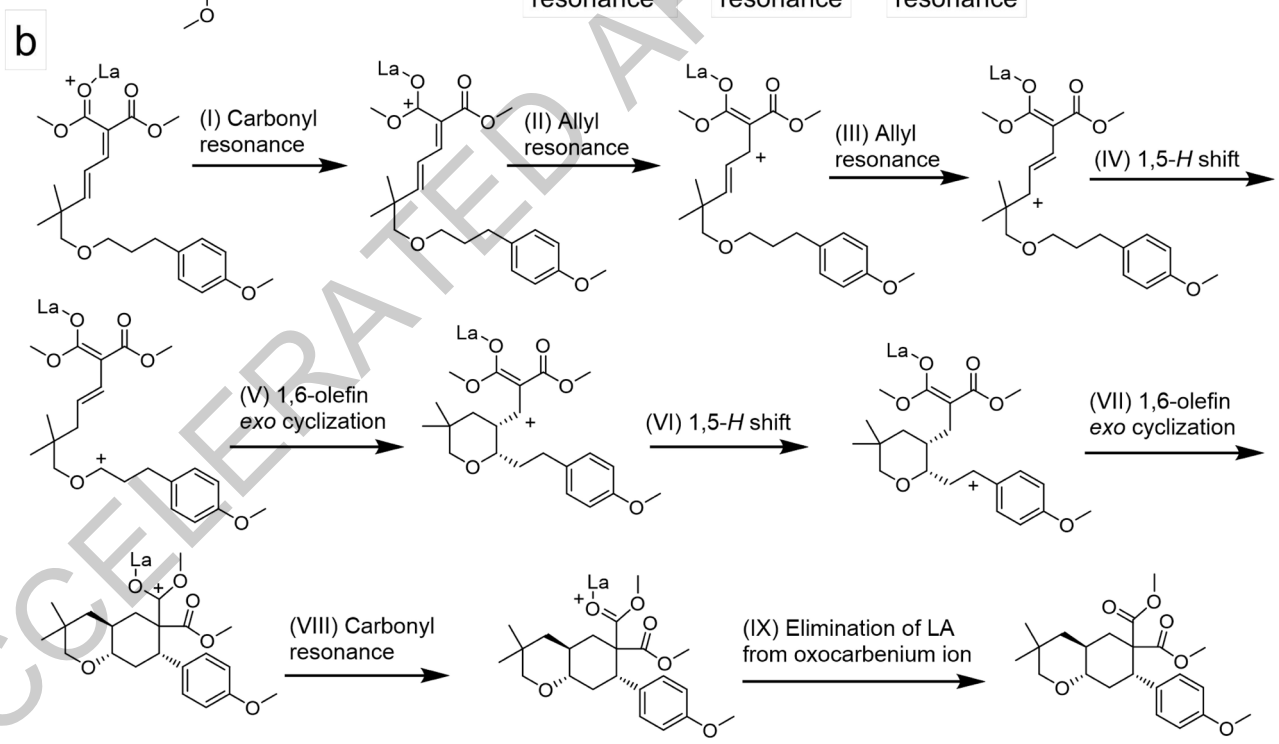




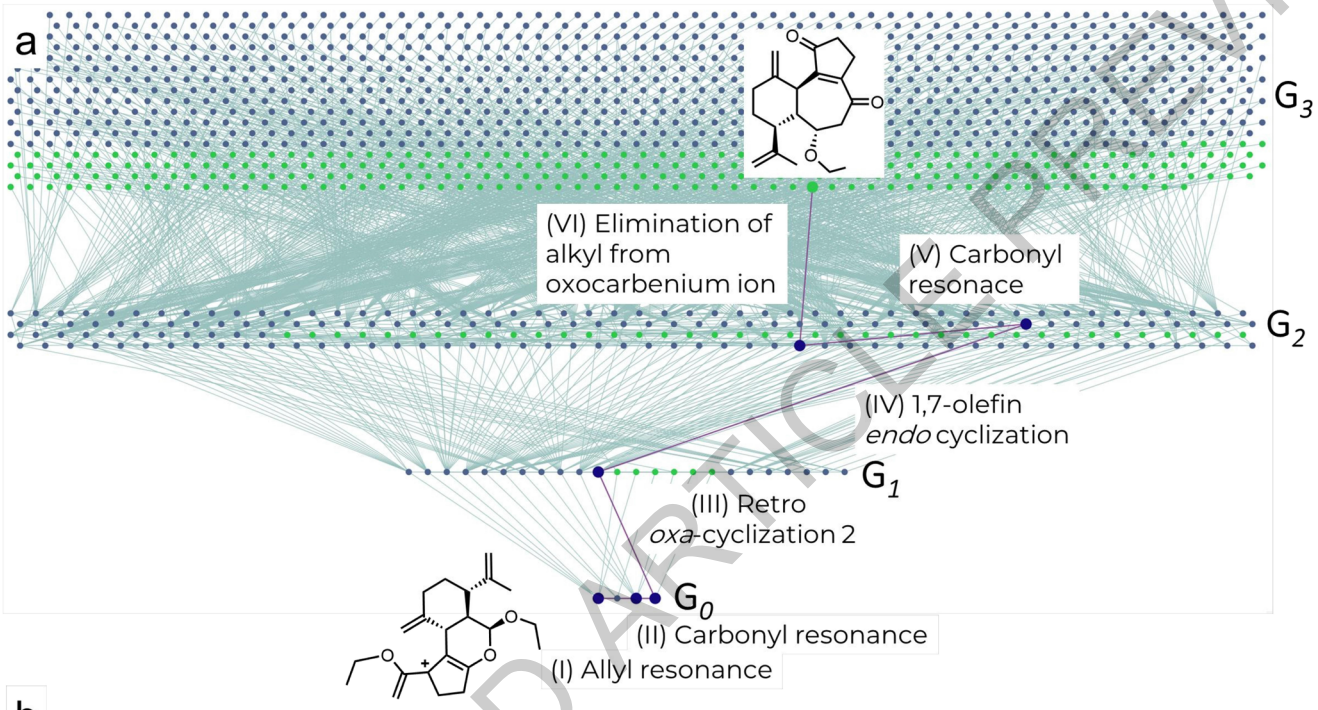


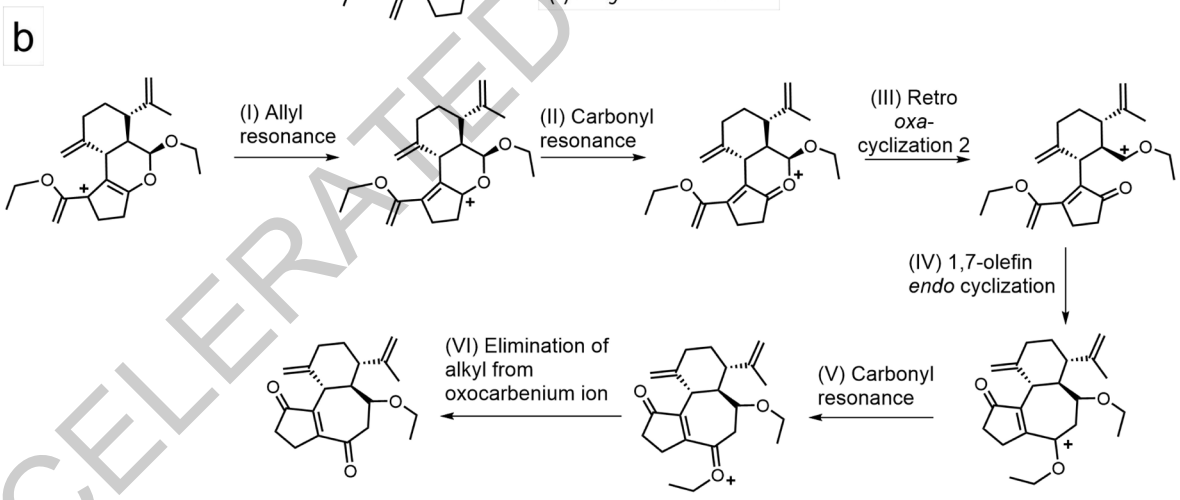




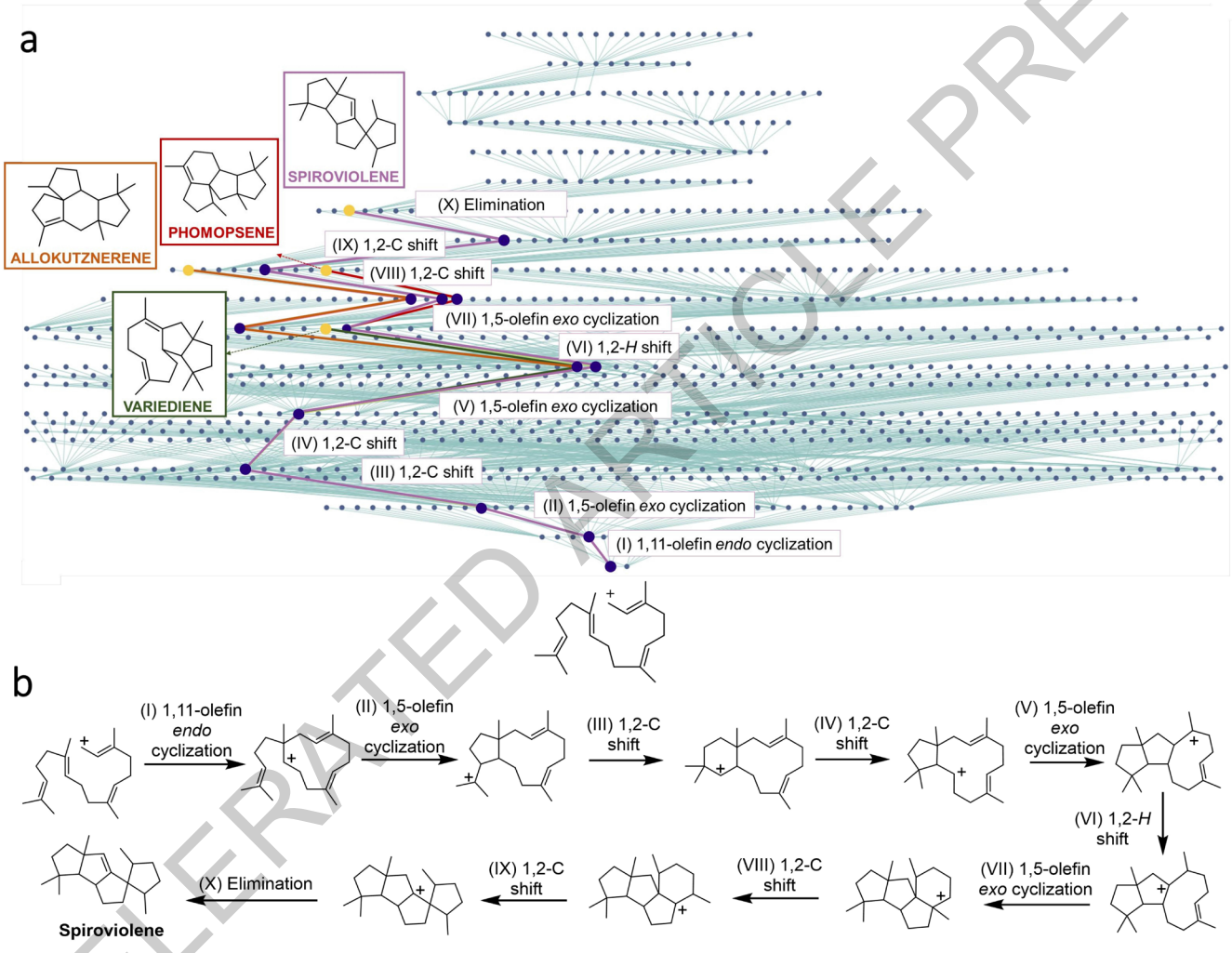


Extended Data Fig. 1

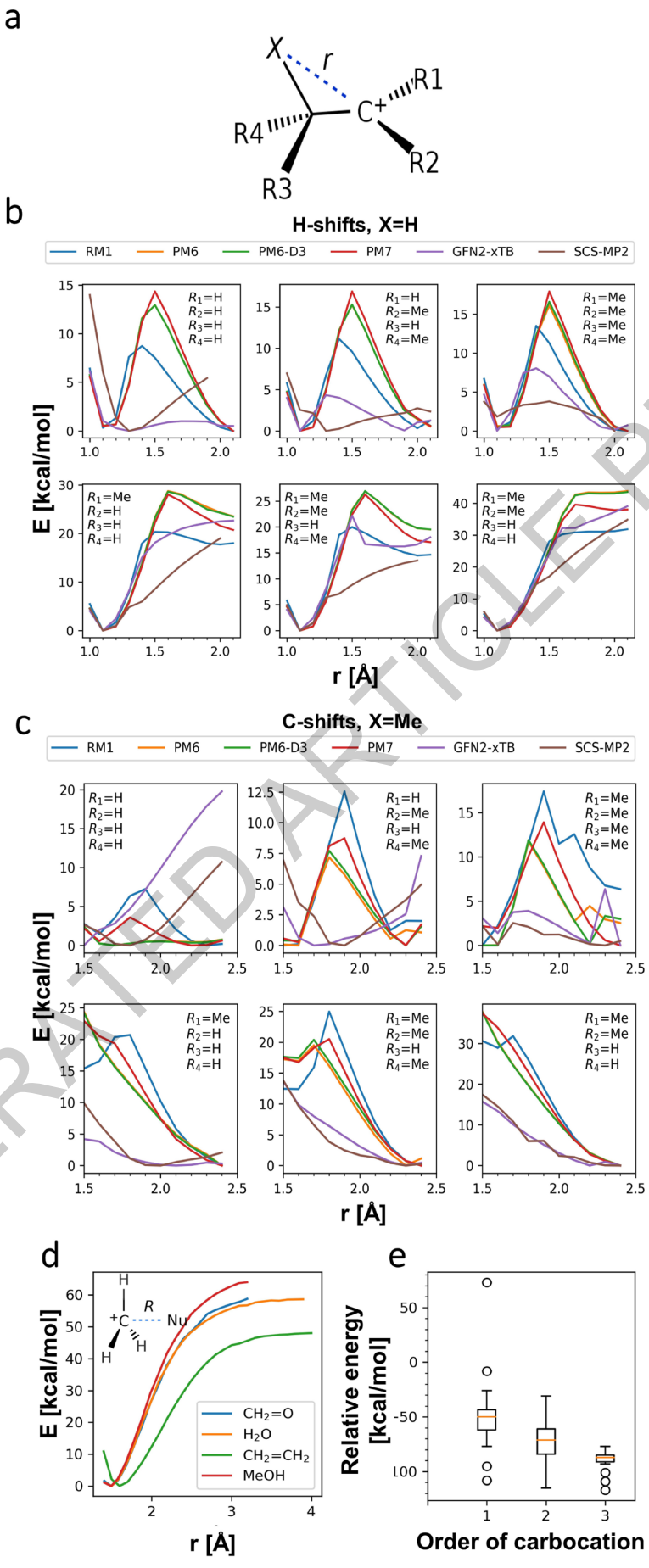




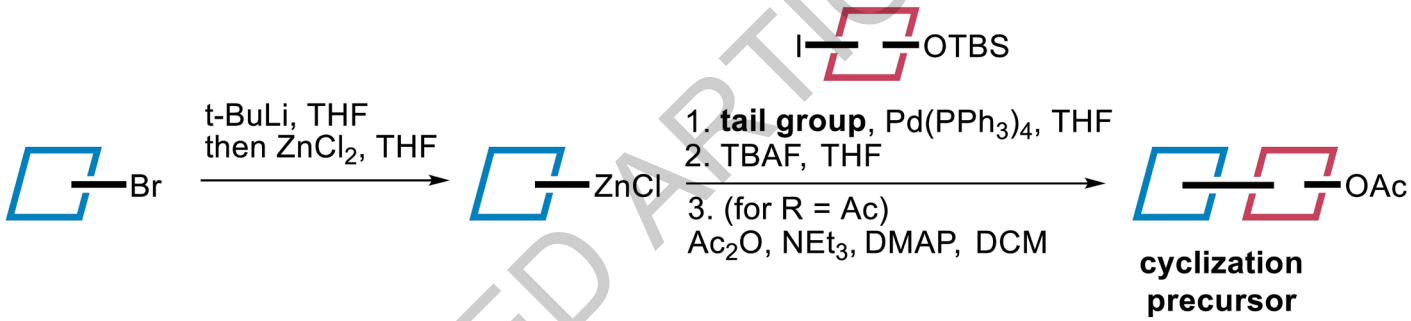
Extended Data Fig. 2



Extended Data Fig. 3



Extended Data Fig. 4



Extended Data Fig. 5

Method	ρ_R	ρ_S	MAE
xTB	0.892	0.910	8.28
PM6-D3	0.894	0.848	8.215
PM6	0.891	0.841	8.384
PM7	0.853	0.789	9.685
RM1	0.821	0.759	9.872
PM3	0.774	0.67	10.932

Extended Data Table 1