Article

Computational planning of the synthesis of complex natural products

https://doi.org/10.1038/s41586-020-2855-y

Received: 25 July 2020

Accepted: 6 October 2020

Published online: 13 October 2020

Check for updates

Barbara Mikulak-Klucznik[1], Patrycja Gołębiowska[1], Alison A. Bayly[2], Oskar Popik[1], Tomasz Klucznik[1], Sara Szymkuć[1], Ewa P. Gajewska[1], Piotr Dittwald[1], Olga Staszewska-Krajewska[1], Wiktor Beker[1], Tomasz Badowski[1], Karl A. Scheidt[2], Karol Molga[1]✉, Jacek Mlynarski[1]✉, Milan Mrksich[2]✉ & Bartosz A. Grzybowski[1,3,4]✉

Training algorithms to computationally plan multistep organic syntheses has been a challenge for more than 50 years[1–7]. However, the field has progressed greatly since the development of early programs such as LHASA[1,7], for which reaction choices at each step were made by human operators. Multiple software platforms[6,8–14] are now capable of completely autonomous planning. But these programs 'think' only one step at a time and have so far been limited to relatively simple targets, the syntheses of which could arguably be designed by human chemists within minutes, without the help of a computer. Furthermore, no algorithm has yet been able to design plausible routes to complex natural products, for which much more far-sighted, multistep planning is necessary[15,16] and closely related literature precedents cannot be relied on. Here we demonstrate that such computational synthesis planning is possible, provided that the program's knowledge of organic chemistry and data-based artificial intelligence routines are augmented with causal relationships[17,18], allowing it to 'strategize' over multiple synthetic steps. Using a Turing-like test administered to synthesis experts, we show that the routes designed by such a program are largely indistinguishable from those designed by humans. We also successfully validated three computer-designed syntheses of natural products in the laboratory. Taken together, these results indicate that expert-level automated synthetic planning is feasible, pending continued improvements to the reaction knowledge base and further code optimization.

Because purely data-oriented artificial intelligence (AI) approaches are not adequate to plan syntheses of complex targets (see Methods for discussion), we have long been developing a hybrid expert–AI system, called Chematica (or Synthia)[8,9,19–26]. Although Chematica has been effective in the design of syntheses that lead to high-value, medicinally relevant targets (validated by experiment[9,20]), its extension to complex natural products—for which the space of synthetic options to explore is orders of magnitude larger (Fig. 1)—has been challenging, and has required numerous improvements. Since the publication of ref. [9], the program has been taught an additional roughly 50,000 mechanism-based reaction rules (it now knows more than 100,000), especially stereoselective and scaffold-directed transformations (see Extended Data Fig. 1a and discussion in Methods). The applicability of these high-quality[27] rules to specific retrons has been further fine-tuned by the addition of various filters (Extended Data Fig. 1b), which evaluate site- or regio-selectivity, by using either machine-learning[21] or quantum-chemistry methods[9], and estimate reaction yields[28,29]. Sets of heuristic rules gauge whether synthons are prone to (unwanted) side reactions and rearrangements[8,9] (Extended Data Fig. 1c), whereas molecular-mechanics-derived heuristics help to evaluate the ability of select classes of synthons to cyclize[27]. During synthesis planning, decisions about each subsequent reaction move can be made by either heuristic or best-in-class neural-network (Extended Data Fig. 1d) scoring functions[22]. The exploration of synthetic space is guided by multiple beam-like searches performed simultaneously (Extended Data Fig. 1e), some that search 'wide' (to suggest diverse chemistries) and others that search 'deep' (to trace pathways to available substrates in the shortest possible time). In our experience, this search strategy worked more effectively for complex targets than did either standard A*[8,9] or Monte Carlo tree-search[10,11] algorithms, which we also tested. Finally, if large numbers of viable pathways are found to form a complex graph of solutions, then this graph is searched by another algorithm, which back-propagates from substrates to the product and, in doing so, assigns realistic cost estimates by which the pathways are ranked (see ref. [23] for details).

With these additions, Chematica began to find routes to some natural products, but produced no results (or only roundabout and unremarkable routes) for others, even when it knew all the individual reactions from which a viable synthesis could have been constructed. These problems were present irrespective of the scoring function used,

[1]Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw, Poland. [2]Department of Chemistry, Northwestern University, Evanston, IL, USA. [3]IBS Center for Soft and Living Matter, Ulsan, South Korea. [4]Department of Chemistry, UNIST, Ulsan, South Korea. ✉e-mail: karolmolga@gmail.com; jacek.mlynarski@gmail.com; milan.mrksich@northwestern.edu; nanogrzybowski@gmail.com

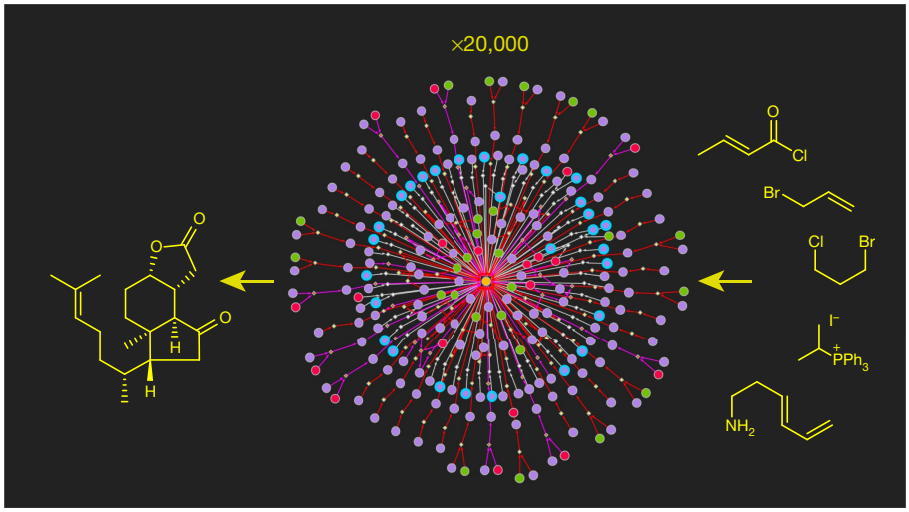


Fig. 1 | Automatic synthesis planning over large networks of possible reactions. This screenshot from Chematica illustrates the synthetic possibilities the machine considers for just one intermediate en route to the natural product aplykurodinone-1 (shown on the left). When designing the full pathway (Extended Data Fig. 6) that traces to the simple starting materials (shown on the right), the program explored and evaluated around 20,000 such graphs connected into a very large network of synthetic options. Each graph comprises one-reaction-step options (white reaction arrows) and multistep sequences (FGIs; red arrows). Nodes correspond to specific molecules: orange, current retron; violet, unknown substances; green, literature-reported substances; red, commercially available chemicals; blue halos, protection needed.

including those that prioritize certain types of disconnections according to Corey's rules of synthesis logic[30]. Analyses of the reaction networks that Chematica generated during such unsuccessful searches revealed that the program was particularly unlikely to follow a series of steps that individually offered little or no structural simplification but could lead to an efficient disconnection downstream. The program showed marked improvement only after the inclusion of relatively few but carefully chosen heuristics that prescribe how to strategize over multiple steps, taking into account how certain reaction choices imply succession (or elimination) of other transformations. These causal relationships—essential for artificial generalized intelligence, which is believed to mimic human reasoning better than models based solely on data-derived correlations[17,18]—were largely inspired by the logic of classic total syntheses designed by human experts, and were of four types.

(1) Two-step sequences are those in which the first step (in the retrosynthetic direction) complexifies the structure but, by doing so, enables a disconnection that offers a high degree of structural simplification (Extended Data Fig. 1f). If the first reaction matches the retron, then the second is executed automatically, allowing Chematica to overcome local maxima of structural complexity and suggest elegant and counterintuitive synthetic strategies. A systematic, big-data analysis[26] allowed us to identify millions of such reaction combinations (compared to the roughly 500 known before); we incorporated of the order of 100,000 of the most useful ones into Chematica.

(2) Functional-group interconversions (FGIs; Extended Data Fig. 1g) are two- or three-step reaction sequences often used in the syntheses of natural products[31], which convert (in the retrosynthetic direction) highly reactive groups into more stable ones and adjust the oxidation level of carbon. If a retron comprises a pattern of functional groups that signals a potential FGI, then the entire sequence is executed, provided none of the individual steps entails any chemical incompatibilities. On the basis of a thorough analysis of several thousand classic total syntheses, we selected around 100 common FGI sequences.

(3) Bypasses (Extended Data Fig. 1h) resolve intermittent reactivity conflicts for otherwise very promising reactions. At a given step, Chematica may encounter a reaction that could lead to a substantial structural simplification (that is, is ranked by the scoring function within the top 10% of possible choices) but is unfeasible because some group(s) within the molecule is incompatible with this proposed reaction. When this happens, the algorithm first checks whether it can execute another reaction (or FGI sequence) that removes the conflict, and only then retries the original, very promising transformation.

(4) Simultaneous and tandem reactions are combinations (pairs, triplets or quadruplets) of reactions types that, under given reaction conditions, should be applied in one rather than multiple reaction steps. Inclusion of this algorithm is important, for instance, for the management of protecting groups, so that several of them can be introduced or removed in one step (Extended Data Fig. 1i).

With these improvements, Chematica became adept in constructing plausible and original routes to targets such as callyspongiolide (Fig. 2a; for discussion and examples of even more complex targets, see Methods and Extended Data Figs. 1–8). Neither the previous versions of Chematica nor other AI tools[11,12] could de-novo construct any sensible routes to such complex targets. In many of its designs, Chematica combined strategies (1)–(4) to effectively probe logically coherent sequences that reach as far as four or five steps downstream. We evaluated these syntheses by using a human-versus-machine Turing-like test and through synthetic validation.

Turing test

We compiled a collection of 40 total syntheses: 20 from the organic-chemistry literature (for example, from *Organic Letters*, *Journal of Organic Chemistry*, *Angewandte Chemie*, *Journal of the American Chemical Society* or *Synlett*) and 20 designed by Chematica. The chosen literature covered the period 1999–2019, and the syntheses were chosen to be representative of these journals. The targets chosen for Chematica were of similar complexity in terms of average mass, number of atoms, stereocentres or rings (see Supplementary Fig. 1 for statistics). For Chematica's designs, we wished to mimic how the program is used by the general chemical audience, so all searches used its default scoring function. Stop points were either commercially available chemicals or simple molecules with known syntheses (but that could be further traced by Chematica, to even simpler substrates, if desired). For each target, the searches (several hours per target; Methods) were run on the newest version of Chematica (which explicitly performs all protection and deprotection steps rather than suggesting the most suitable protecting groups only at steps requiring protection, as in previous versions

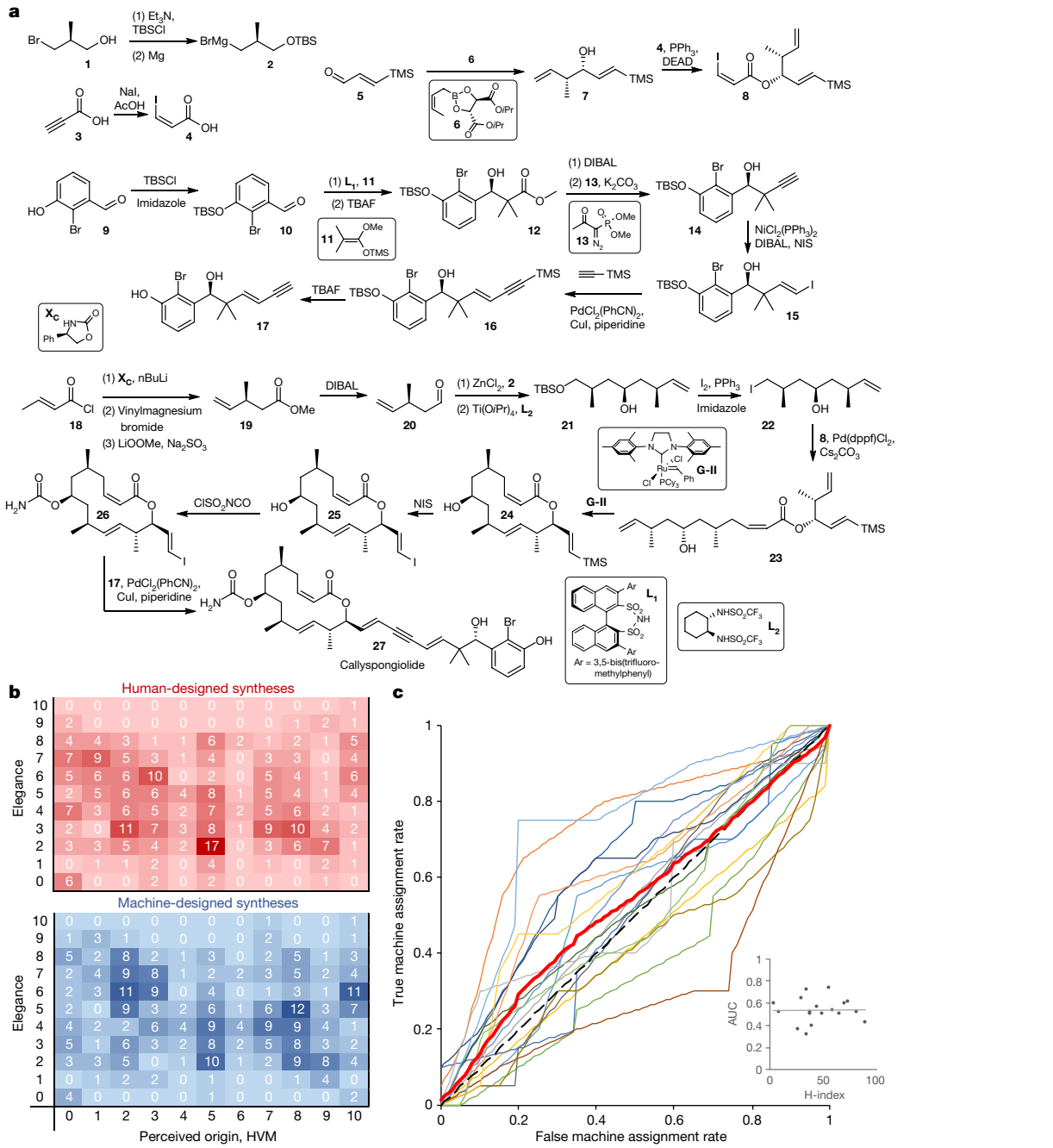


Fig. 2 | Synthesis Turing test. a, One of the total syntheses evaluated by the experts. This particular pathway leading to callyspongiolide was designed by Chematica, but was judged by experts as slightly more likely to be human-designed (and of average elegance, although without any chemical issues). In the test, syntheses were redrawn in ChemDraw in one unified format, such as that shown here, to preclude detection of their origin—that is, literature-derived pathways were drawn with experimental conditions but no yields, and Chematica-derived pathways were drawn with conditions suggested by the program (from the most relevant publication linked to Chematica's particular reaction rule). The answer key and raw test statistics for all 40 pathways are provided in Supplementary Information section 1. **b**, Distribution of HVM scores, quantifying the perceived origin of the pathways (see main text), and elegance *E* scores. The red heatmap is for literature pathways; the blue heatmap is for pathways designed by Chematica. Each cell corresponds to a particular combination of HVM and *E* scores. Numbers in white give the number of judges who voted for the given (HVM, *E*) combination; a darker colour means more votes. **c**, ROC curves representing the answers of individual experts (thin lines) and the average ROC curve for all experts (thick red line). ROC curves are constructed by plotting the true assignment rate against the false assignment rate for different thresholds, and are used to evaluate the accuracy of the classifier. The (0, 0)–(1, 1) diagonal (dashed black line; AUC = 0.5) corresponds to an uninformative, random-guessing scenario. In our Turing test, the mean ROC (red curve) and mean AUC (0.53; inset) of all responders are close to the random-guessing scenario. As shown in the inset, more experienced responders (higher H-index) did not achieve better results than less experienced responders (lower H-index). The correlation coefficient between AUC and H-index is only $R^2 = 0.000267$.

of the program) and the top-scoring pathways were retrieved. All 40 syntheses, arranged in no particular order, were posted online using a quiz service (https://www.quiz-maker.com); they are also included, along with the answer key, in Supplementary Information section 1.

On such a set, we queried 18 synthesis experts (see Acknowledgements). We asked these experts to assign a 'human versus machine' (HVM) score, on a 0-to-10 scale, corresponding to the perceived likelihood that a given pathway was designed by a human or by the machine

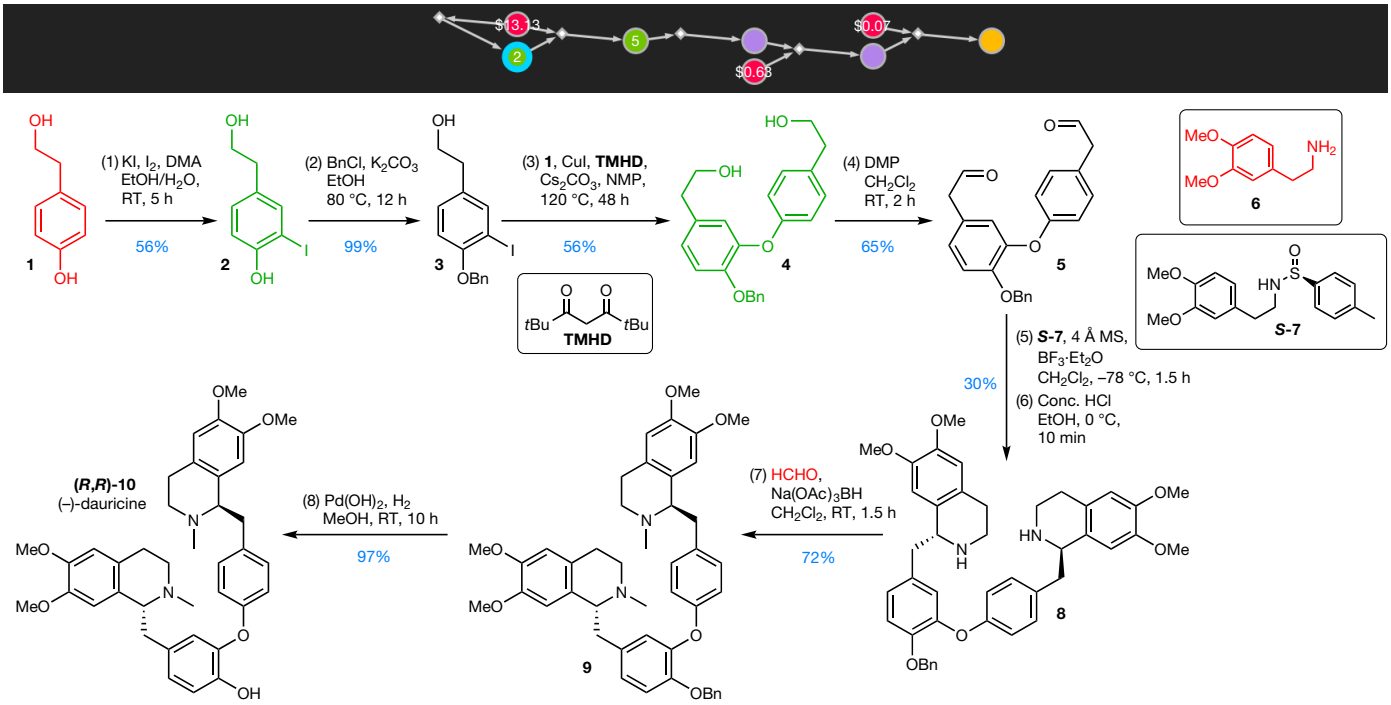


Fig. 3 | Synthesis of dauricine. Here we show the stereocontrolled synthesis of (–)-dauricine, planned by Chematica (top) and executed in the laboratory (bottom). Nodes correspond to specific molecules: orange, target; violet, unknown substances; green, known substances (with numerals denoting the number of literature-reported syntheses in which a particular molecule was used as a substrate[8,19,38,39]); red, terminal, commercially available chemicals (prices in USD per gram); blue halo, protection needed. Substrates drawn or listed in red (green) in the bottom panel correspond to the red (green) nodes in the top panel. Experimental yields are given in blue. RT, room temperature; MS, molecular sieves. See Methods and Supplementary Information section 2.2 for experimental details.

(HVM = 0 means definitely human-designed; HVM = 10 means definitely machine-designed). We also asked the experts to judge the synthetic elegance of the pathways, from uninspired ($E = 0$) to remarkable ($E = 10$).

The key question was whether the machine would pass the synthetic-chemical version of the so-called Turing test[32]—that is, whether a considerable portion of the experts would believe that the machine-generated synthetic plans were created by humans. The HVM and E scores assigned by the experts are summarized in Fig. 2b, c; detailed responses for all participants are summarized in Supplementary Figs. 2–4. For the machine-generated pathways, there were 144 (42%) incorrect votes (HVM < 5), 44 (12.8%) votes for "I do not know" (HVM = 5), and 155 (45.2%) correct votes (HVM > 5). Using the average expert scores for each pathway, 10 were (incorrectly) judged to be of human origin and 10 were (correctly) judged to be designed by Chematica (in the 20 literature pathways, 12 were judged as human-designed and 8 as machine-designed.) The average HVM scores over all pathways were only 0.6 points higher for Chematica-designed routes than for literature routes ($\langle HVM_{human\text{-}designed}\rangle = 4.58$, $\langle HVM_{machine\text{-}designed}\rangle = 5.17$). In terms of elegance, machine-designed pathways were voted to be slightly more elegant ($\langle E_{human\text{-}designed}\rangle = 4.55$, $\langle E_{machine\text{-}designed}\rangle = 4.75$; E did not correlate with HVM.) When $\langle HVM_{machine\text{-}designed}\rangle$ was transformed onto the 0–100 scale of the perception of human-likeness used to evaluate chatbots in Turing tests ($HL = 10 \times (10 - \langle HVM_{machine\text{-}designed}\rangle) = 48.23$), Chematica performed better than or comparably to chatbots described in ref. [33]. We also analysed the responses of the experts by using the methods used to evaluate binary classifiers. For each expert, we constructed the receiver operating characteristic (ROC) curve (Fig. 2c; the average of the individual ROC curves is shown in red). The area-under-the-curve (AUC) metric for this curve is 0.53, with a standard error of 0.03. This means that guesses of the cohort of experts were close, to within error, to random guessing (AUC = 0.5). These guesses did not correlate with the experts' H-indices (Fig. 2c, inset). Taken together, these results indicate that Chematica passes the Turing test, because the experts were generally not able to discern the origin (machine or human) of the natural-product syntheses provided in the quiz.

Experimental validation

We chose three natural-product targets of different complexities. The simplest target was (–)-dauricine (Fig. 3), a potent autophagy blocker and anticancer agent[34], which has been synthesized only in racemic form, via a Bischler–Napieralski reaction[35]. The intermediate-complexity target was a recently isolated[36] but not-yet-synthesized iboga alkaloid, called tacamonidine (Fig. 4). The most complex target was lamellodysidine A (Fig. 5), a bridged polycyclic sesquiterpene isolated[37] in 2017, yet lacking total synthesis. Lamellodysidine A comprises a tetracyclic carbon framework, with seven contiguous (including three quaternary) stereocentres and an acid-labile hemiacetal, which make its synthesis challenging and of contemporary synthetic interest.

For each of these targets, Chematica suggested multiple routes (see examples in Supplementary Information section 3.12), of which the top ones were chosen. Because our main objective was to verify the predictions of the program, no alterations to the proposed disconnections were allowed. When needed, organic chemists performing the syntheses were allowed to adjust reaction conditions (such as temperature, solvent, specific base or catalyst) for the sake of optimization.

The satisfactory experimental yields (given next to reaction arrows in Figs. 3–5) demonstrate that all three syntheses worked as planned (see Methods and Supplementary Information sections 2–4 for synthetic details). The synthesis of dauricine (Fig. 3) was straightforward, with the key step being the Pictet–Spengler cyclization, which was performed using Davis auxiliary. The synthesis of tacamonidine (Fig. 4) was interesting because the multiple syntheses of its close analogue, tacamonine, are not readily adaptable to this target; they do not allow

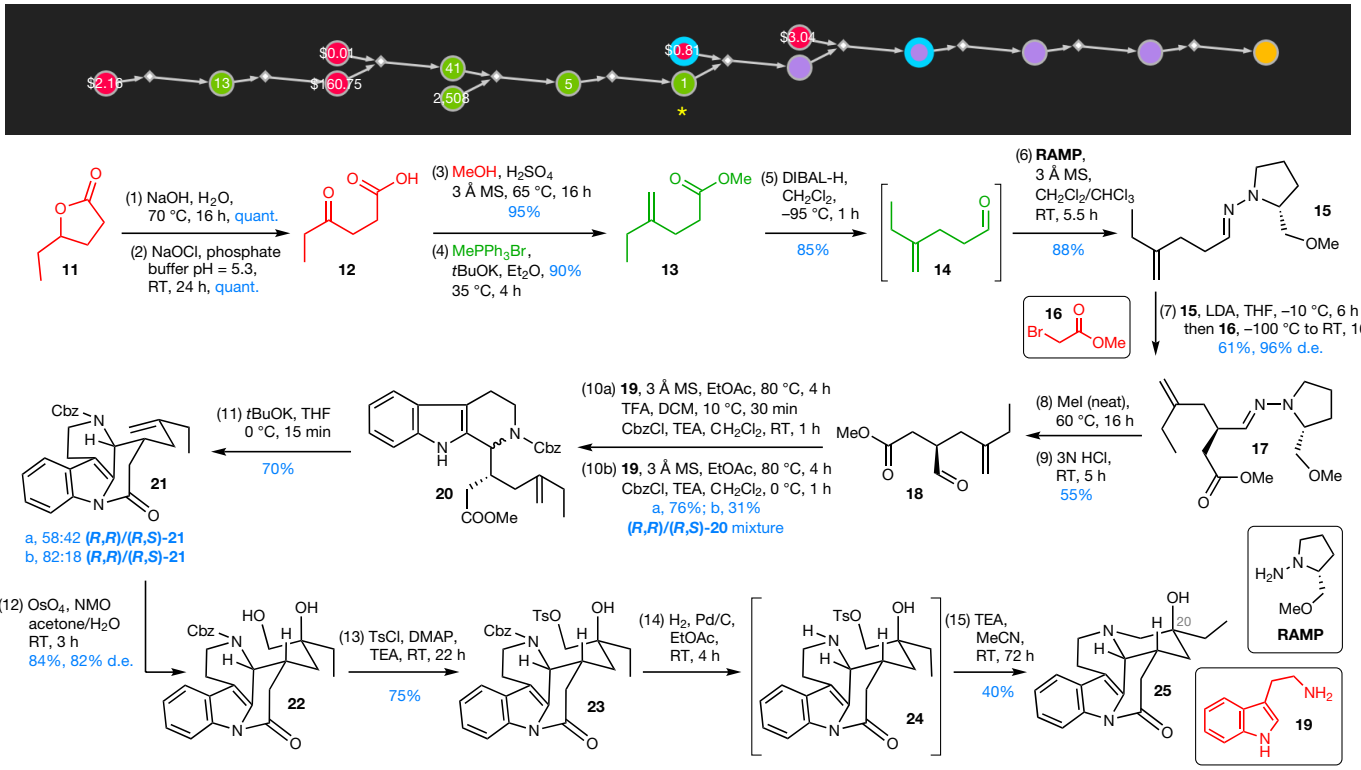


Fig. 4 | Total enantioselective synthesis of (*R*,*R*,*S*)-tacamonidine. The synthesis was planned by Chematica (top) and executed in the laboratory (bottom). The colours of the nodes (top) are as in Fig. 3. The yellow asterisk denotes intermediate **14**, which is known; its synthesis could be found from the literature-based NOC (network of chemistry) module of Chematica[19,38,39]. However, on reaching this starting material, we allowed the program to continue the de novo search and navigate the synthesis all the way to very simple, commercially available building blocks (terminal red nodes). Substrates drawn or listed in red (green) in the bottom panel correspond to the red (green) nodes in the top panel. Text in blue corresponds to experimental yield (quant., quantitative yield) or diastereomeric excess (d.e.). For compounds **22**–**24**, the minor diastereoisomers are not shown; for **25**, the formation of 20-epi-tacamonidine was not detected. See Methods and Supplementary Information section 3.5 for experimental details.

for the necessary stereochemical control, and the intermediates generated in these pathways are not amenable to the ultimate construction of the quaternary hydroxylated stereocentre of tacamonidine (see Supplementary Information section 3.2 for details). In its route, Chematica initially constructed the tricyclic tryptoline system by Pictet–Spengler cyclisation—a common, step-efficient method of preparing chiral tetrahydroisoquinoline and tetrahydrocarboline alkaloids—but then followed this with an original and logical sequence of

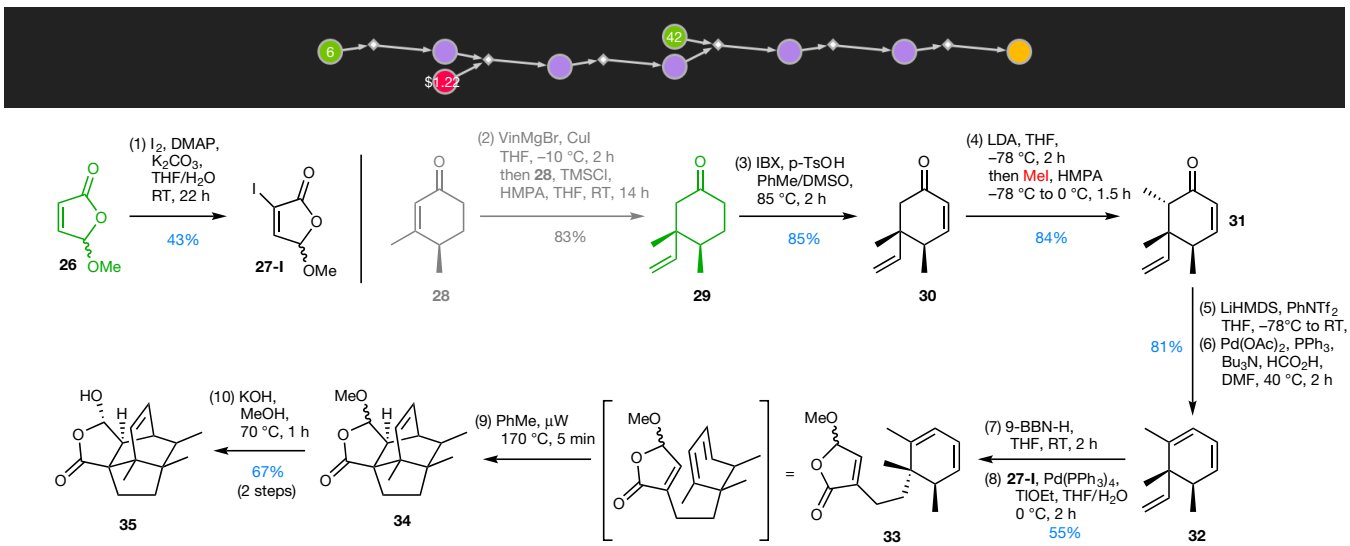


Fig. 5 | Total enantioselective synthesis of lamellodysidine A. The synthesis was planned by Chematica (top) and executed in the laboratory (bottom). The colours of the nodes (top) are as in Fig. 3. Substrates drawn or listed in red (green) in the bottom panel correspond to the red (green) nodes in the top panel. Experimental yields are given in blue. The synthesis of the starting material **29** is known; here, it was obtained via a modified procedure from enone **28**, which is available in three steps from dihydrocarvone. See Methods and Supplementary Information section 4.2 for experimental details.

intramolecular indole *N*-acylation, asymmetric dihydroxylation and final ring-closing alkylation. Finally, in the synthesis of lamellodysidine A (Fig. 5), the key elements were the conversion of **29** into triene **32**, its subsequent hydroboration and Suzuki coupling with vinyl halide **27-I** and, ultimately, intramolecular Diels–Alder cycloaddition, which provided access to the desired tetracyclic [2.2.2]-bicyclooctene framework of lamellodysidine. The stereodefined hemiacetal was then obtained via hydrolysis of methoxyfuranone. Chematica correctly predicted the stereochemistry of the Diels–Alder cycloadduct and the selective formation of the thermodynamically more stable and less hindered stereoisomer in the last step. With these experimental demonstrations, the number of Chematica-predicted pathways validated by experiment is now 16 (8 in ref. [9], 4 in ref. [20], 1 in ref. [26] and 3 here), which in total comprise more than 70 individual reaction steps.

In summary, our results indicate that computers are finally becoming capable of creating reliable synthetic plans, comparable to those designed by highly trained synthetic chemists, and with the upper bound on the complexity of the targets as in Extended Data Figs. 2–7. Reaching this level took decades of work[1–7] (nearly two for our team alone[8,9,20–27,38,39]) because automated synthetic planning at the expert level is so multifaceted. It requires large numbers of accurate rules that describe individual reactions, careful structural evaluation of the generated synthons, efficient algorithms for graph searching, scoring functions and, as we highlighted here, routines to mimic human-like strategizing over multiple steps. Recognizing that mastering the art of synthesis for extremely complex natural products will require additional advanced-chemistry rules, improvements in hardware and further acceleration of code, we believe that Chematica can now be a useful companion to practising synthetic chemists (see Methods section 'Extended discussion' for additional examples of syntheses of very complex targets, discussion of current limitations, performance under different constraints or risk-taking scenarios and pending improvements). Looking forward, the next challenge will be to teach the machine to discover completely new reaction types, which could then be validated by experiment and used in retrosynthetic planning.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2855-y.

1. Corey, E. J. & Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science* **166**, 178–192 (1969).
2. Gelernter, H. L. et al. Empirical explorations of SYNCHEM. *Science* **197**, 1041–1049 (1977).
3. Hanessian, S., Franco, J. & Larouche, B. The psychobiological basis of heuristic synthesis planning - man, machine and the Chiron approach. *Pure Appl. Chem.* **62**, 1887–1910 (1990).
4. Hendrickson, J. B. Systematic synthesis design. 6. Yield analysis and convergency. *J. Am. Chem. Soc.* **99**, 5439–5450 (1977).
5. Ugi, I. et al. Computer-assisted solution of chemical problems - the historical development and the present state of the art of a new discipline of chemistry. *Angew. Chem. Int. Edn Engl.* **32**, 201–227 (1993).
6. Todd, M. H. Computer-aided organic synthesis. *Chem. Soc. Rev.* **34**, 247–266 (2005).
7. Ravitz, O. Data-driven computer aided synthesis design. *Drug Discov. Today. Technol.* **10**, e443–e449 (2013).
8. Szymkuć, S. et al. Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016).
9. Klucznik, T. et al. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* **4**, 522–532 (2018).
10. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
11. Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019).
12. SciFinder[n], https://scifinder-n.cas.org (accessed 20 July 2020).
13. Lee, A. A. et al. Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chem. Commun.* **55**, 12152–12155 (2019).
14. Schwaller, P. et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).
15. Nicolaou, K. C. *Classics in Total Synthesis II: More Targets, Strategies, Methods* (Wiley-VCH, 2003).
16. Huang, P. *Efficiency in Natural Product Total Synthesis* (Wiley, 2018).
17. Yi, K. et al. CLEVERER: collision events for video representation and reasoning. Preprint at https://arxiv.org/abs/1910.01442 (2020).
18. Bergstein, B. *What AI still can't do.* MIT Technical Review https://www.technologyreview.com/s/615189/what-ai-still-cant-do/ (2020).
19. Kowalik, M. et al. Parallel optimization of synthetic pathways within the network of organic chemistry. *Angew. Chem. Int. Ed.* **51**, 7928–7932 (2012).
20. Lin, Y. et al. Reinforcing the supply chain of COVID-19 therapeutics with expert-coded retrosynthetic software. Preprint at https://doi.org/10.26434/chemrxiv.12765410.v1 (2020).
21. Beker, W., Gajewska, E. P., Badowski, T. & Grzybowski, B. A. Prediction of major regio-, site-, and diastereoisomers in Diels-Alder reactions by using machine-learning: the importance of physically meaningful descriptors. *Angew. Chem. Int. Ed.* **58**, 4515–4519 (2019).
22. Badowski, T., Gajewska, E. P., Molga, K. & Grzybowski, B. A. Synergy between expert and machine-learning approaches allows for improved retrosynthetic planning. *Angew. Chem. Int. Ed.* **59**, 725–730 (2020).
23. Badowski, T., Molga, K. & Grzybowski, B. A. Selection of cost-effective yet chemically diverse pathways from the networks of computer-generated retrosynthetic plans. *Chem. Sci.* **10**, 4640–4651 (2019).
24. Molga, K., Dittwald, P. & Grzybowski, B. A. Computational design of syntheses leading to compound libraries or isotopically labelled targets. *Chem. Sci.* **10**, 9219–9232 (2019).
25. Molga, K., Dittwald, P. & Grzybowski, B. A. Navigating around patented routes by preserving specific motifs along computer-planned retrosynthetic pathways. *Chem* **5**, 460–473 (2019).
26. Gajewska, E. P. et al. Algorithmic discovery of tactical combinations for advanced organic syntheses. *Chem* **6**, 280–293 (2020).
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. Emami, F. E. et al. A priori estimation of organic reaction yields. *Angew. Chem. Int. Ed.* **54**, 10797–10801 (2015).
29. Skoraczyński, G. et al. Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient? *Sci. Rep.* **7**, 3582 (2017).
30. Corey, E. J. & Cheng, X.-M. *The Logic of Chemical Synthesis* (Wiley, 1995).
31. Serratosa, F. *Organic Chemistry in Action: The Design of Organic Synthesis* (Elsevier, 1996).
32. Copeland, B. J. (ed.) *The Essential Turing: The Ideas That Gave Birth to the Computer Age* (Oxford Univ. Press, 2004).
33. Shah, H., Warwick, K., Vallverdú, J. & Wu, D. Can machines talk? Comparison of Eliza with modern dialogue systems. *Comput. Human Behav.* **58**, 278–295 (2016).
34. Yang, Z. et al. Dauricine induces apoptosis, inhibits proliferation and invasion through inhibiting NF-κB signaling pathway in colon cancer cells. *J. Cell. Physiol.* **225**, 266–275 (2010).
35. Kametani, T. & Fukumoto, K. Total synthesis of (±)-dauricine. *Tetrahedr. Lett.* **5**, 2771–2775 (1964).
36. Lim, K.-H. et al. Ibogan, tacaman, and cytotoxic bisindole alkaloids from *Tabernaemontana*. Cononusine, an iboga alkaloid with unusual incorporation of a pyrrolidone moiety. *J. Nat. Prod.* **78**, 1129–1138 (2015).
37. Torii, M. et al. Lamellodysidines A and B, sesquiterpenes isolated from the marine sponge *Lamellodysidea herbacea*. *J. Nat. Prod.* **80**, 2536–2541 (2017).
38. Fialkowski, M., Bishop, K. J. M., Chubukov, V. A., Campbell, C. J. & Grzybowski, B. A. Architecture and evolution of organic chemistry. *Angew. Chem. Int. Ed.* **44**, 7263–7269 (2005).
39. Grzybowski, B. A., Bishop, K. J. M., Kowalczyk, B. & Wilmer, C. E. The 'wired' universe of organic chemistry. *Nat. Chem.* **1**, 31–36 (2009).

Methods

Additional algorithmic considerations

Computer-aided synthesis of simple versus complex molecules. Simple molecules are made in few synthetic steps, usually via sequences of disconnections into increasingly simpler synthons. Stereocentres, if present, are often sourced from starting materials, and the use of protection chemistries is limited. There are often many different ways to make a product, and the space of synthetic options to explore is relatively small, of the order of billions (for example, $\mathcal{O}(100^5)$ for a five-step synthesis[8]). By contrast, synthesis of complex natural products is more like chess, in which sequences of preparatory moves—which do not simplify the structure per se—build the position that leads to a key disconnection. The stereocentres have to be carefully set up and the protecting-group strategies must be thoughtfully planned. There are often very few viable ways of reaching the target, yet the space of potential syntheses can be extremely large[8]—$\mathcal{O}(100^n)$, where the number of steps n is typically tens. Navigation of this vast space must be guided by very accurate (yet often chemically advanced) reaction rules, using methods to prioritize which reaction steps to take and which to avoid.

Limitations of data-driven AI in advanced synthesis planning. There are multiple reasons why purely data-oriented AI methods trained on even the largest reaction repositories (such as Reaxys or SciFinder), but without mechanistic insight, are not adequate for natural-product syntheses. First, the reaction rules automatically derived from such repositories do not capture all the relevant stereochemical information (not to mention that a sizeable fraction of reactions suffer from manual-entry errors; see ref. [27] for comprehensive discussion of these and other aspects). Second, because failed and side-reactions are usually not published, the machine cannot deduce the scope of potentially conflicting groups; this is especially problematic when a given reaction type has only a few literature precedents for complex scaffolds and it is not warranted to assume that not-reported reactions are impossible (see also next paragraph). Third, these repositories are largely dominated by simple, non-stereoselective reactions, so scoring functions derived from them to select the most promising reaction moves are geared towards these simpler and more popular chemistries (see ref. [22] for a detailed discussion). Because chess-like, total syntheses of natural products are a very small fraction of these repositories, the machine is not incentivized to learn and purposefully apply sequences of position-building, but not structure-simplifying, steps; in addition, it does not learn for which rare scaffolds it is worth performing FGIs or structure-complexifying steps, when to perform two or more different reactions on the same molecule, which reactions should be performed in tandem, or how to navigate around intermittent reactivity conflicts.

Reaction rules, reactivity conflicts and selectivity issues. The main set of Chematica's >100,000 expert-coded reaction transforms are mechanism-based. These rules generalize and are broader than the underlying literature precedents. From the beginning of our effort on Chematica, the rules have been added to the software following an iterative procedure, in which we test the performance of the program in finding routes to increasingly complex types of scaffolds. In doing so, we focus on adding methodologies that enable synthesis of entire target classes (for example, cationic cyclizations of polyenes to make steroid targets) rather than specific literature precedents. After adding a given methodology and testing the performance of the algorithm (that is, its ability to synthesize a given class of targets it was previously not able to synthesize), a new class of targets is selected and the process is repeated. However, the software also supports a comparable number of rules that were machine-extracted from reaction repositories. But these transforms are used only optionally (mostly for rare types of ring), as their quality is considerably lower than the expert-coded set; they were not used in this study. The issues of rule quality are described in detail elsewhere[27]. In the context of natural-product synthesis, machine extraction is ill-suited for stereoselective reactions, especially those in which stereochemistry is dictated by distant stereocentres already present in the molecule; in complex scaffolds, these directing motifs may lie far from the core of atoms that change their immediate environments, and their automated extraction from the underlying reaction precedent is problematic. Another aspect of the rules is that they must be accompanied by comprehensive lists of groups that are incompatible with the reaction under its specified reaction conditions. In Chematica, reactivity conflicts are treated very stringently. Nearly 400 potentially conflicting groups are considered when coding each transform; selection of specific conflicting groups is especially meaningful because it is based on the underlying reaction mechanism, in conjunction with appropriate reaction conditions. In this way, only chemically sound options are allowed and the number of conflict-free reaction candidates that match a given retron is typically a few tens to roughly 100. With the lower-quality machine-extracted rules, the numbers of reactions considered per retron are typically much higher (for example, roughly 46,000 during expansion and more than 300 during rollouts in 3M Monte Carlo tree searches[10]), exacerbating the combinatorial explosion of the search space. Specification of reaction conditions for each transform is also essential for assessing chemoselectivity and preventing potential side-reactions: "machine-extraction methods can learn about chemoselectivity problems by considering hypothetical, computer-generated reactions not reported in the literature. They assume that such a hypothetical reaction is not feasible if, starting from the same substrates, it generates a product that is different than a product already reported in literature—which is a dangerous chemical oversimplification, since different products can often be obtained by simply changing reaction conditions or a catalyst. In effect, such methods disqualify large numbers of potentially feasible though not yet reported reactions."[22]

The need to fine tune the rules. A reaction rule might be perfectly applicable to a given retron, but the synthons it generates may be problematic. In some cases, the synthons contain highly strained motifs (such as small rings with triple bonds or unsaturations at bridgehead atoms), which are readily eliminated by globally imposed filters that eliminate such motifs for all reactions[8,9,27]. However, sometimes the synthons are not obviously problematic, but contain motifs that, under the conditions of a specific reaction, are known to be unstable (for example, prone to rearrangement). Accordingly, Chematica's reaction rules are accompanied by lists of motifs or scaffolds that are problematic in a particular reaction class (see example in Extended Data Fig. 1c). Another problem arises when synthon or retron molecules are not themselves strained, but excessive strain develops in the transition state, for example, during intramolecular cyclization within a polycyclic system. Such problematic motifs, at least the most obvious ones, are identified and prohibited on the basis of molecular-mechanics calculations (these simulations benefit from the knowledge of the reaction mechanism and narrow the conformational space to the known angles of approach; see refs. [9,27] for discussion). Yet another class of problems pertains to reaction types for which the number of possible substituents is not only extremely large—too large to enumerate exhaustively—but also their mutual placement and steric and/or electronic properties are essential. One example is aromatic substitutions, which in Chematica are guided by electron-density calculations (see supplementary information for ref. [9]). In reaction types for which the reaction core is uniquely defined and the number of literature examples is large (thousands or more), machine-learning models enable very accurate predictions of regio-, site- or diastereoselectivity[21]. However, the training set must be selected very carefully, excluding reactions that involve the same types of retron and synthon, but proceed through different reaction mechanisms (for example, nucleophilic aromatic substitution versus transition-metal-catalysed coupling between an aryl chloride and an amine). Machine-learning

models give the best results—not only accurate for a given test or training set but also transferrable to not-yet-seen types of substrate—when they are trained using physically meaningful descriptors, which capture steric and electronic features of the molecules involved (see ref. [21] for discussion).

Scoring functions, search algorithms and retrieval of top-scoring pathways. Scoring functions used within Chematica are described and compared in ref. [22]. Regarding the search algorithms that these functions guide, most recent works on computer-aided retrosynthesis use Monte Carlo tree searches. However, in our experience, applying Monte Carlo tree searches to complex molecules did not produce viable pathways, and in many cases did not identify any pathways, probably because the roll-outs could not solve the complex retrons (which was not a serious problem for simple retrons during planning of short syntheses[10]). On the basis of these experiences, the newest version of Chematica relies on a combination of algorithms, some that search wide and others that search deep. Our earlier algorithm (see supplementary information for ref. [9]) always popped from the priority queue search graph nodes (representing found beginnings of synthetic pathways) with the lowest scores (computed as sums of costs of encountered terminal substrates and of scoring functions of reactions and nonterminal substrates). To increase the search width, the improved algorithm uses a beam-search-inspired[40] priority queue, which keeps a given number (beam width) of encountered nodes with the lowest scores at each search depth. The queue ensures that before a node with the lowest overall score is popped the nodes with beam-width record scores at all lower search depths are popped. The new algorithm also allows the use of several such queues with different chemical scoring functions simultaneously, in which case new nodes are pushed to all queues, while pops are done from each queue in turn. In a typical scenario, two such queues are used, one with a scoring function that prefers wide searches, and another with a scoring function that prefers to go deeply, trying to reach commercially available and/or literature-known starting materials as rapidly as possible. When used together, the first queue allows us to discover promising beginnings of routes (that is, breaking some of the beam-width score records at lower depths), which can then be quickly finished with the help of the second queue.

Although details of search performance vary from target to target, typical execution times for complex natural products are hours, during which time the search may expand 10,000–100,000 synthon nodes. Because each synthon has, on average, $\mathcal{O}(100)$ progenies, the total chemical space being evaluated in a search is therefore up to $\mathcal{O}(10^7)$ molecules. We stress that most of the CPU time is not spent on performing the reaction and search operations (which are very rapid), but on enforcing proper stereochemistry of the reactions (via the Stereofix module[8,9]) and evaluating the molecules using the auxiliary routines discussed above.

Finally, even for complex targets Chematica is capable of identifying multiple plausible syntheses (compare some tacamonidine examples in Supplementary Information section 3.11). These syntheses are scored and ranked (and clustered, to avoid similar pathways dominating the rankings) by previously described algorithms[23].

Synthetic details

Syntheses of (−)-dauricine. The stereoselective route designed by Chematica (Fig. 3; details in Supplementary Information section 2) applies the 4-hydroxyethylphenol (**1**) substrate in two unique bond-forming steps and installs stereocentres present in the structure of dauricine in a single step via a stereoselective Pictet–Spengler cyclization. Starting from **1**, a straightforward sequence of iodination, benzylation, coupling and bisoxidation proceeded smoothly (compare yields in blue in Fig. 3) to provide the dialdehyde **5** on a gram scale. The stereoselectivity of the subsequent Pictet–Spengler cyclization was controlled by a chiral auxiliary attached to the homobenzylamine **6**. Specifically, we followed Koomen's approach[41], and used a *N*-sulfinyl (Davis auxiliary) decorated amine ***S*-7** as the coupling partner. This two-step condensation–auxiliary-removal procedure afforded **8** in 30% yield (corresponding to about 55% per ring formed), isolated as a single diastereoisomer (apparent diastereomeric ratio > 20:1). To complete the synthesis, **8** was subjected to robust reductive amination conditions to provide the benzyl-protected dauricine **9** in 72% yield, which was further debenzylated (97%) to achieve (−)-dauricine **10**. Spectroscopic and physical data matched the values we measured for the commercially purchased standard of (−)-dauricine (AdooQ Bioscience).

Synthesis of (*R*,*R*,*S*)-tacamonidine. The graph of the top-scoring pathway for (*R*,*R*,*S*)-tacamonidine is shown at the top of Fig. 4. Within five steps (plus protection and deprotection of carboxylic acid and secondary amine; protections indicated by blue halos) from the target, the program reached 4-methylenehexanal as the starting substrate. The green colour of this node indicates that a synthesis of this compound has been reported in the literature and is available within the Network of Organic Chemistry[19,38,39], with which Chematica can communicate. However, because the established five-step sequence[42] involves low-yielding steps, and starts from volatile and irritating 2,3-dibromopropene, we permitted the de novo search to continue to identify an alternative five-step sequence. This alternative sequence involves straightforward, scalable reactions (hydrolysis of a lactone, oxidation of a secondary alcohol, esterification, Wittig olefination and ester reduction to an aldehyde) and terminates in commercially available γ-hexalactone and methanol (red, terminal nodes, with prices in USD per gram).

This route was validated in the laboratory and is detailed at the bottom of Fig. 4, with conditions mirroring those proposed by the program (see detailed screenshots in Supplementary Information section 3.4). The synthesis commenced with quantitative hydrolysis of γ-hexalactone **11**, followed by oxidation of the secondary alcohol. The resulting carboxylic acid (**12**) was then converted to its methyl ester in high yield (86%). Further Wittig exomethylenation produced the unsaturated ester (**13**), which was immediately subjected to DIBAL-H reduction to yield **14** (78% over two steps); this was used in the subsequent step without isolation. The first stereocentre was introduced using a standard Enders procedure: condensation of **14** with RAMP (90% yield) afforded **15**, which was deprotonated with LDA, and alkylated with methyl bromoacetate to give **17** in 61% yield. The resulting hydrazone **17** (96% diastereomeric excess) was subjected to reaction with MeI and subsequent acid hydrolysis to remove the chiral auxiliary, and to provide chiral aldehyde **18** in 55% yield. This aldehyde was allowed to react with tryptamine **19** to give the desired imine, which was subsequently protonated[43] with TFA to induce the Pictet–Spengler cyclization. The isolated product was immediately protected with CbzCl to provide **20** in 76% yield as an inseparable (at this stage) mixture of (*R*,*R*) and (*R*,*S*) diastereoisomers (method 'A'; see also Supplementary Fig. 24a). The obtained mixture was subjected to reaction with *t*BuOK in THF to give a mixture (*R*,*R*:*R*,*S* = 58:42) of tetracyclic amides **21**, from which the desired (*R*,*R*) isomer was isolated via flash column chromatography in 31% yield over four steps. Alternatively, we sought to obtain **21** via a *N*-acyliminium Pictet–Spengler cyclization[44] (method 'B'; see also Supplementary Fig. 24b). To do so, aldehyde **18** was reacted with tryptamine **19** to form an imine, which was directly treated with CbzCl to induce cyclization. Subsequent lactamization afforded a 82:18 mixture of *R*,*R*:*R*,*S* **21** from which **(*R*,*R*)-21** was isolated (relative configuration confirmed by COSY, HSQCAD, HMBCAD and 2D-NOESY techniques; see Supplementary Information section 3.10), albeit in lower yield (17% over three steps). The remaining steps towards (*R*,*R*,*S*)-tacamonidine (**25**) required formation of the quaternary hydroxylated stereocentre. For the Sharpless asymmetric dihydroxylation, Chematica suggested either AD-mix-α or OsO_4; both variants were performed. AD-mix-α gave 23% yield (65% yield based on recovered starting material) and 81% diastereomeric excess. The reaction with only OsO_4/NMO provided the

diol mixture in higher yield (84%) and similar diastereomeric excess (82%), pointing to the stereodirecting effect of the scaffold (quantum mechanical calculations suggest that OsO_4 coordination from the *si* face is energetically favoured over the *re* face by 1.6 kcal mol^{-1}; such subtle effects are not considered by Chematica during synthesis planning). The primary alcohols within the diols were then tosylated (75%), the amine groups were deprotected and the intermediate was subjected to intramolecular amination to afford the *R*,*R*,*S*-tacamonidine target **25** in 40% yield over the last two steps. Spectroscopic (^{1}H NMR, ^{13}C NMR, COSY, HSQCAD, HMBCAD, 2D-NOESY, IR, UV–Vis and HRMS) and physical (optical rotation) data confirmed the structure of the target and matched literature[36]. This data and all synthetic details are provided in Supplementary Information section 3.5–3.10.

Synthesis of lamellodysidine A. The graph of the top-scoring pathway for lamellodysidine A is shown at the top of Fig. 5. Within six steps from the target, Chematica reached cyclohexanone **29** and vinyl bromide **27-Br**, the syntheses of which have already been reported in the literature. The key elements of Chematica's plan are the conversion of **29** into triene **32**, its subsequent hydroboration and Suzuki coupling with vinyl halide **27**, and finally intramolecular Diels–Alder cycloaddition to provide access to the desired tetracyclic [2.2.2]-bicyclooctene framework of lamellodysidine. The stereodefined hemiacetal is obtained via hydrolysis of methoxyfuranone, which leads to the thermodynamically more stable and less hindered stereoisomer. This plan was executed in the laboratory and commenced with the addition[45] of the vinyl cuprate to enone[46] **28** to give **29** in 83% yield after acidic workup. Subsequent oxidation[47] with IBX-DMSO gave enone **30** in 85% yield, which was then methylated under standard conditions (LDA, MeI, HMPA in THF) to give enone **31** in 84% yield. Subsequent enolization with LiHMDS and triflation with $PhNTf_2$ gave the vinyl triflate, which was then reduced with $Pd(OAc)_2$/HCOOH and gave access to the desired triene **32** in 81% yield. The remaining steps to lamellodysidine A required attachment of the methoxyfuranone fragment, which was realized by the one-pot hydroboration of **32** with 9-BBN-H and subsequent coupling with **27-I**, which is available in a single step from 5-methoxy-2(5*H*)-furanone **26** (a more reactive iodoacetal **27-I** was used instead of the proposed bromoacetal **27-Br**). The obtained intermediate **33**, which had all fragments of lamellodysidine A, was subjected to short (roughly 5 min) microwave heating in toluene to trigger the intramolecular Diels–Alder reaction and afford the expected cycloadduct **34**. Subsequent hydrolysis of crude **34** yielded the more stable diastereoisomer exclusively and gave lamellodysidine A in 67% yield and with 17.7% overall yield from **28**. Spectroscopic (^{1}H NMR, ^{13}C NMR, COSY, HSQCAD, HMBCAD, 2D-NOESY, IR and HRMS) and physical (optical rotation) data confirmed the structure of the target and matched literature[37]. This data and all synthetic details are provided in Supplementary Information section 4.2–4.3.

Extended discussion

Limits of the complexity of the target. Beyond the examples provided in the main text, it is important to estimate the upper level of the complexity of the targets for which Chematica can plan plausible syntheses. The examples in Extended Data Figs. 2–5 demonstrate that the program designs plausible routes to targets such as cephanolide B[48], conidiogenone B[49], scabrolide A[50] and taxuyunnanine D[51], the syntheses of which were recently published in leading chemical journals (pathways to the less complex targets aplykurodinone-1[52] and dendrobine[53] are shown in Extended Data Figs. 6, 7).

On the other hand, Chematica was not able to find routes to targets such as CJ-16,264[54], ryanodol[55] or taxol[56]. In some cases (for example, CJ-16,264), failure could be attributed to the program not yet being taught a certain class of reaction (for example, stereoselective substrate-controlled condensation of enolate with imide[57] for the synthesis of CJ-16,264). However, when the program fails to find a plausible route despite having the requisite chemical knowledge, it is likely that an insufficiently small fraction of the synthetic space has been explored. This seems to be the case for taxol. For such complex targets, the numbers of viable options per step may be very high (around 200 or more); searches over a 100^n–200^n-sized search space usually time out, exceeding the available RAM (roughly 500 GB on our machines). However, Chematica was able to design synthesis to a less oxidized taxane (taxuyunnanine D[51]; Extended Data Fig. 5), which lies roughly half-way in Baran's pyramid of taxanes[56] and was synthesized by Baran's group in 2014[51].

The synthesis of taxol[56] (building on a modified taxane carbon framework) is interesting for another reason—the allowable level of risk taking. In several steps of the synthesis route, multiple equivalent sites were present but reactions were still performed on one of them selectively, owing to a skilful choice of advanced reagents (for example, initial, selective oxidation at C13 using a chromium(V)–hydroxyacid complex), isotopic exchange (to prevent oxidation of a secondary alcohol during C1 oxidation with DMDO) or solvent effects (only a unique protic and non-nucleophilic mixture of HFIP/TMSOH (2:1) led to a high selectivity of C13 oxidation)[56]. In these steps, the humans were taking calculated and creative risks that capitalized on the structure of the particular scaffold. By contrast, in its typical configuration, Chematica is told to act more conservatively, using generally applicable reaction types and assigning a high penalty for any potential non-selectivities encountered (if a competing position, non-equivalent by symmetry, is found, the reaction is deemed non-selective and is heavily penalized). However, the non-selectivity parameter in the scoring function in Chematica can be lowered by the user. For complex targets, this could permit riskier yet more elegant disconnections (with the hope that non-selectivity nuances can be tweaked by the structure of the scaffold). However, for simpler targets (or intermediates), lowering the penalty usually leads to serious non-selectivity issues that cannot be remedied. One potential solution to this problem might be adjusting the degree of the penalty proportionally to the complexities of the targets or intermediates. In the absence of such a self-adjusting scoring scheme, the remedy is to code additional reaction rules for specific or complex scaffolds (which encompass much wider reaction cores). Still, for targets of industrial or medicinal interest and mid-complexity natural products, we recommend the use of a high non-selectivity penalty, as used here.

Performance with different search parameters. In Chematica, it is possible to exclude from the searches specific structures or reactions. Because each reaction used by Chematica comes with a common name (for example, Michael addition or Sharpless dihydroxylation) and typical conditions and reagents, the user can exclude entire reaction classes (and/or reagents) by using appropriate keywords. This could be useful when looking for routes that have no precedent in the literature or illustrating the usefulness of certain methodologies if using Chematica as a teaching aid. The use of such exclusion is illustrated in the synthesis of scabrolide A (Extended Data Fig. 4), where we excluded SAMP and RAMP hydrazones to minimize the use of chiral auxiliaries. Another example is provided in Extended Data Fig. 8. The top panel illustrates the results of an unconstrained search for the synthesis of mevastatin. The suggested route relies on an intramolecular Diels–Alder reaction to construct the 6–6 ring system of mevastatin. By contrast, in the bottom panel, the program was forbidden from using any Diels–Alder reactions. Predictably, the route with the additional constraint is longer, although it is still chemically plausible (the rings are formed via ring-closing metathesis).

The multistep strategizing routines described in the main text (tactical combinations, FGIs, bypasses and tandem reactions) may also be used optionally, although this is not recommended. For very simple targets, numerous synthetic solutions may be found even without these enhancements, but for more complex ones, problems arise. First,

for very complex natural products, searches without any multistep strategizing typically time out with no pathways found; second, for less advanced (but not trivially simple) targets, the routes—even if found in realistic search times—are often less elegant (longer or more roundabout). One example of the latter is provided in Extended Data Fig. 9; we used Chematica to search for syntheses of ramelteon (a sleep medication). Without strategizing algorithms, only an unremarkable 13-step pathway was found, even after several hours of searching. By contrast, with strategizing algorithms, the program found a much more concise (8 steps) and elegant route after only about 10 min. In this route, Chematica used a tactical combination of Robinson annulation followed by the aromatization of cyclohexenone (in the retrosynthetic direction, dearomatization of phenol offered no immediate structural simplification but set the scene for the structure-simplifying Robinson annulation). In another example (Extended Data Fig. 10), the program was searching for syntheses of tybost, a drug used for the treatment of HIV. The route designed without any strategizing routines involved an additional protection–deprotection sequence mid-way; no better pathway was identified after hours of searching. By contrast, with the strategizing routines, a more elegant pathway (which relied on two bypasses and one FGI) was found after only about 15 min. In this route, the program sourced this synthesis from substrates that were already appropriately protected (so that no mid-way protection–deprotection sequence was needed) and easily available from appropriate amino acids.

Pending and future improvements. As discussed above, searches for very complex targets can time out, exceeding the available RAM. Preventing such outcomes will require hardware improvements, which is one of the main focuses of the ongoing commercial deployment of Chematica as Synthia. Improving memory size and management is also a prerequisite for using advanced options such as library-wide searches, in which Chematica designs syntheses to several targets simultaneously, often benefiting from the use of common intermediates (provided the targets are not completely unrelated). The effectiveness of such searches has been demonstrated[24] for small libraries of medicinally relevant targets, although the large search graphs were already straining memory limits; for more complex natural products, these problems are compounded.

Several of our algorithmic improvements have yet to be incorporated into Chematica/Synthia. One is the neural-network tool to estimate the pK_a of CH acids (https://pka.allchemy.net/)[58] and thus choose the loci of reactions such as alkylations more accurately (when many similar sites are present). Another class of improvements regards an algorithm that increases the diversity of the search results: after initial pathways are found, but are similar and use the same key disconnection (algorithmically detected as the reaction that offers the highest degree of structural simplification), the program prevents the use of this reaction type in subsequent steps. In this way, the searches are forced to abandon the already-visited regions of the search space and seek materially different solutions. A similar objective can, to some extent, be achieved by starting a new search with the exclusion of a given reaction (compare above and Extended Data Fig. 8), but this loses the search graph already explored. Another improvement focuses on optimizing the endings of the already-found pathways. The point here is that, as the search graphs spread out from the target outwards, the algorithm might spend less and less time on smaller and smaller intermediates. This means that the endings of the pathways might not be optimized (compare the synthesis of scabrolide A in Extended Data Fig. 4). To remedy this, we have been implementing an algorithm that, using an already-found path, moves to an intermediate, say, five steps away from the starting materials and initializes a local search from this molecule. If it finds a shorter and/or lower-scoring route to this intermediate, then the five-step ending is replaced by the shorter, say, three-step, one. An intermediate five steps from the new end of the updated pathway is then chosen and another search is performed. This cycle is repeated until the endings cannot be shortened or optimized any more.

Finally, it is often useful to scrutinize Chematica-designed pathways by using additional quantum-mechanics, molecular-mechanics or machine-learning post-filters. During the searches for the pathways, the algorithm explores thousands to millions of intermediates and reactions, the evaluation of which must be very rapid (a fraction of a second each). When a pathway or several top-scoring pathways are found and selected as interesting, it is no longer a problem to spend an additional few seconds on each intermediate or step and use more accurate evaluation methods. For instance, in the current version of Chematica, the user can examine strain within each intermediate; such molecular-mechanics calculations take only about 1–10 s. Similarly, we have developed, but not yet connected to Chematica, hybrid machine-learning–knowledge-based models to evaluate scaffold-directing (non-covalent) effects for certain reaction types for which the trajectory of approach can be reasonably approximated. These methods rest on a vectorized representation of distances between the atoms of the approaching substrates.

Data availability

All data that support the findings of this study are available within the paper and its Supplementary Information, or from the corresponding authors on reasonable request.

Code availability

In Supplementary Data, we provide the pseudocode for the multistep retrosynthetic design, pathway generation and retrieval (PSEUDOCODE_Aug2.pdf), an example of one of the reaction rules as coded in Chematica (RULE.pdf), and additional details of the availability and execution of the software (README_Aug2.pdf).

40. Sammut, C. in *Encyclopedia of Machine Learning and Data Mining* (eds Sammut, C. & Webb, G. I.) 120 (Springer, 2017).
41. Gremmen, C., Willemse, B., Wanner, M. J. & Koomen, G.-J. Enantiopure tetrahydro-β-carbolines via Pictet–Spengler reactions with *N*-sulfinyl tryptamines. *Org. Lett.* **2**, 1955–1958 (2000).
42. Gansäuer, A., Worgull, D., Knebel, K., Huth, I. & Schnakenburg, G. 4-exo cyclizations by template catalysis. *Angew. Chem. Int. Ed.* **48**, 8882–8885 (2009).
43. Hadjaz, F., Yous, S., Lebegue, N., Berthelot, P. & Carato, P. A mild and efficient route to 2-benzyl tryptamine derivatives via ring-opening of β-carbolines. *Tetrahedron* **64**, 10004–10008 (2008).
44. Taylor, M. S. & Jacobsen, E. N. Highly enantioselective catalytic acyl-Pictet–Spengler reactions. *J. Am. Chem. Soc.* **126**, 10558–10559 (2004).
45. Goetz, A. E., Silberstein, A. L., Corsello, M. A. & Garg, N. K. Concise enantiospecific total synthesis of tubingensin A. *J. Am. Chem. Soc.* **136**, 3036–3039 (2014).
46. White, J. D., Grether, U. M. & Lee, Ch.-S. (R)-(+)-3,4-dimethylcyclohex-2-en-1-one. *Org. Synth.* **82**, 108 (2005).
47. Nicolaou, K. C., Zhong, Y.-L. & Baran, P. S. A new method for the one-step synthesis of α,β-unsaturated carbonyl systems from saturated alcohols and carbonyl compounds. *J. Am. Chem. Soc.* **122**, 7596–7597 (2000).
48. Xu, L., Wang, C., Gao, Z. & Zhao, Y.-M. Total synthesis of (±)-cephanolides B and C via a palladium-catalyzed cascade cyclization and late-stage sp^3 C–H bond oxidation. *J. Am. Chem. Soc.* **140**, 5653–5658 (2018).
49. Xu, B., Xun, W., Su, S. & Zhai, H. Total syntheses of (−)-conidiogenone B, (−)-conidiogenone, and (−)-conidiogenol. *Angew. Chem. Int. Ed.* **59**, 16475 (2020).
50. Hafeman, N. J. et al. The total synthesis of (−)-scabrolide A. *J. Am. Chem. Soc.* **142**, 8585–8590 (2020).
51. Wilde, N. C., Isomura, M., Mendoza, A. & Baran, P. S. Two-phase synthesis of (−)-taxuyunnanine D. *J. Am. Chem. Soc.* **136**, 4909–4912 (2014).
52. Zhang, Y. & Danishefsky, S. J. Total synthesis of (±)-aplykurodinone-1: traceless stereochemical guidance. *J. Am. Chem. Soc.* **132**, 9567–9569 (2010).
53. Guo, L., Frey, W. & Plietker, B. Catalytic enantioselective total synthesis of the picrotoxane alkaloids (−)-dendrobine, (−)-mubironine B, and (−)-dendroxine. *Org. Lett.* **20**, 4328–4331 (2018).
54. Nicolaou, K. C. et al. Total synthesis and structural revision of antibiotic CJ-16,264. *Angew. Chem. Int. Ed.* **54**, 9203–9208 (2015).
55. Chuang, K. V., Xu, C. & Reisman, S. E. A 15-step synthesis of (+)-ryanodol. *Science* **353**, 912–915 (2016).
56. Kanda, Y. et al. Two-phase synthesis of taxol. *J. Am. Chem. Soc.* **142**, 10526–10533 (2020).
57. Lambert, T. H. & Danishefsky, S. J. Total synthesis of UCS1025A. *J. Am. Chem. Soc.* **128**, 426–427 (2006).

58. Roszak, R., Beker, W., Molga, K. & Grzybowski, B. A. Rapid and accurate prediction of pKa values of C–H acids using graph convolutional neural networks. *J. Am. Chem. Soc.* **141**, 17142–17149 (2019).
59. Crosby, S. R., Harding, J. R., King, C. D., Parker, G. D. & Willis, C. L. Oxonia-Cope rearrangement and side-chain exchange in the Prins cyclization. *Org. Lett.* **4**, 577–580 (2002).
60. Kormann, C., Heinemann, F. W. & Gmeiner, P. A consecutive Diels–Alder approach toward a Tet repressor directed combinatorial library. *Tetrahedron* **62**, 6899–6908 (2006).
61. Owens, K. R. et al. Total synthesis of the diterpenoid alkaloid Arcutinidine using a strategy inspired by chemical network analysis. *J. Am. Chem. Soc.* **141**, 13713–13717 (2019).
62. Jung, M. E. & Davidov, P. Efficient synthesis of a tricyclic BCD analogue of ouabain: Lewis acid catalyzed Diels–Alder reactions of sterically hindered systems. *Angew. Chem. Int. Ed.* **41**, 4125–4128 (2002).
63. Sheu, J.-H., Ahmed, A. F., Shiue, R.-T., Dai, C.-F. & Kuo, Y.-H. Scabrolides A–D, four new norditerpenoids isolated from the soft coral *Sinularia scabra*. *J. Nat. Prod.* **65**, 1904–1908 (2002).
64. Cui, W.-X. et al. Polycyclic furanobutenolide-derived norditerpenoids from the South China Sea soft corals *Sinularia scabra* and *Sinularia polydactyla* with immunosuppressive activity. *Bioorg. Chem.* **94**, 103350 (2020)
65. Mendoza, A., Ishihara, Y. & Baran, P. S. Scalable enantioselective total synthesis of taxanes. *Nat. Chem.* **4**, 21–25 (2012).
66. Liao, W. & Yu, Z.-X. DFT study of the mechanism and stereochemistry of the Rh(I)-catalyzed Diels–Alder reactions between electronically neutral dienes and dienophiles. *J. Org. Chem.* **79**, 11949–11960 (2014).
67. Xu, B., Xun, W., Wang, T. & Qiu, F. G. Total synthesis of (+)-aplykurodinone-1. *Org. Lett.* **19**, 4861–4863 (2017).
68. Wang, Y.-M., Bruno, N. C., Placeres, Á. L., Zhu, S. & Buchwald, S. L. Enantioselective synthesis of carbo- and heterocycles through a CuH-catalyzed hydroalkylation approach. *J. Am. Chem. Soc.* **137**, 10524–10527 (2015).

Acknowledgements Development of Chematica was partly supported by US DARPA under the Make-It Award, 69461-CH-DRP #W911NF1610384 (K.M., S.S., E.P.G., P.D., T.B., B.A.G.); the same award also supported the synthesis of dauricine (A.A.B., M.M.). Synthesis of tacamonidine was supported in part (B.M.-K., T.K., B.A.G.) by the National Science Center, NCN, Poland under the Symfonia Award (#2014/12/W/ST5/00592). Synthesis of lamellodysidine A was supported in part (P.G., B.A.G.) by the National Science Center, NCN, Poland under the Maestro Award (#2018/30/A/ST5/00529). J.M. and O.P. thank the Foundation for Polish Science for financial support under award TEAM/2017-4/38. B.A.G. acknowledges support from the Institute for Basic Science Korea, project code IBS-R020-D1. We thank B. Sieredzińska for help in the synthesis of tacamonidine and S. Trice (Merck, KGaA) for help in organizing the Turing test. We thank the following experts for their participation in the Turing test (in alphabetical order): P. Baran (Scripps), J. Bode (ETH Zurich), M. Burke (University of Illinois), M. Christmann (Freie Universität Berlin), H. Davies (Emory University), M. Giedyk (ICHO PAN), D. Huryn (University of Pittsburgh), M. Krische (University of Texas), S. Matsubara (Kyoto University), N. Maulide (Universität Wien), G. Molander (University of Pennsylvania), R. Sarpong (Berkeley), P. Schreiner (Justus Liebig University Giessen) and J. Siitonen (Rice University), as well as four others, who prefer to remain anonymous.

Author contributions B.M.-K. and T.K. synthesized tacamonidine. P.G. and O.P. synthesized lamellodysidine A. A.A.B. synthesized dauricine. K.M., S.S., E.P.G., P.D. and T.B. were key developers of Chematica; K.M., S.S., E.P.G. and P.D. implemented the Turing test. O.S.-K. determined the structures and relative configuration of compounds ***(R,R)*-21**, ***(R,S)*-21** and final (*R,R,S*)-tacamonidine. K.M. and W.B. performed analyses of the Turing-test results. K.A.S. and M.M. supervised synthesis of dauricine. J.M. supervised synthesis of lamellodysidine A. B.A.G. supervised the synthesis of tacamonidine, designed the Turing test and supervised, along with K.M., its administration, and conceived and directed the development of Chematica from its inception 20 years ago. K.M. and B.A.G. wrote the paper, with contributions from other authors.

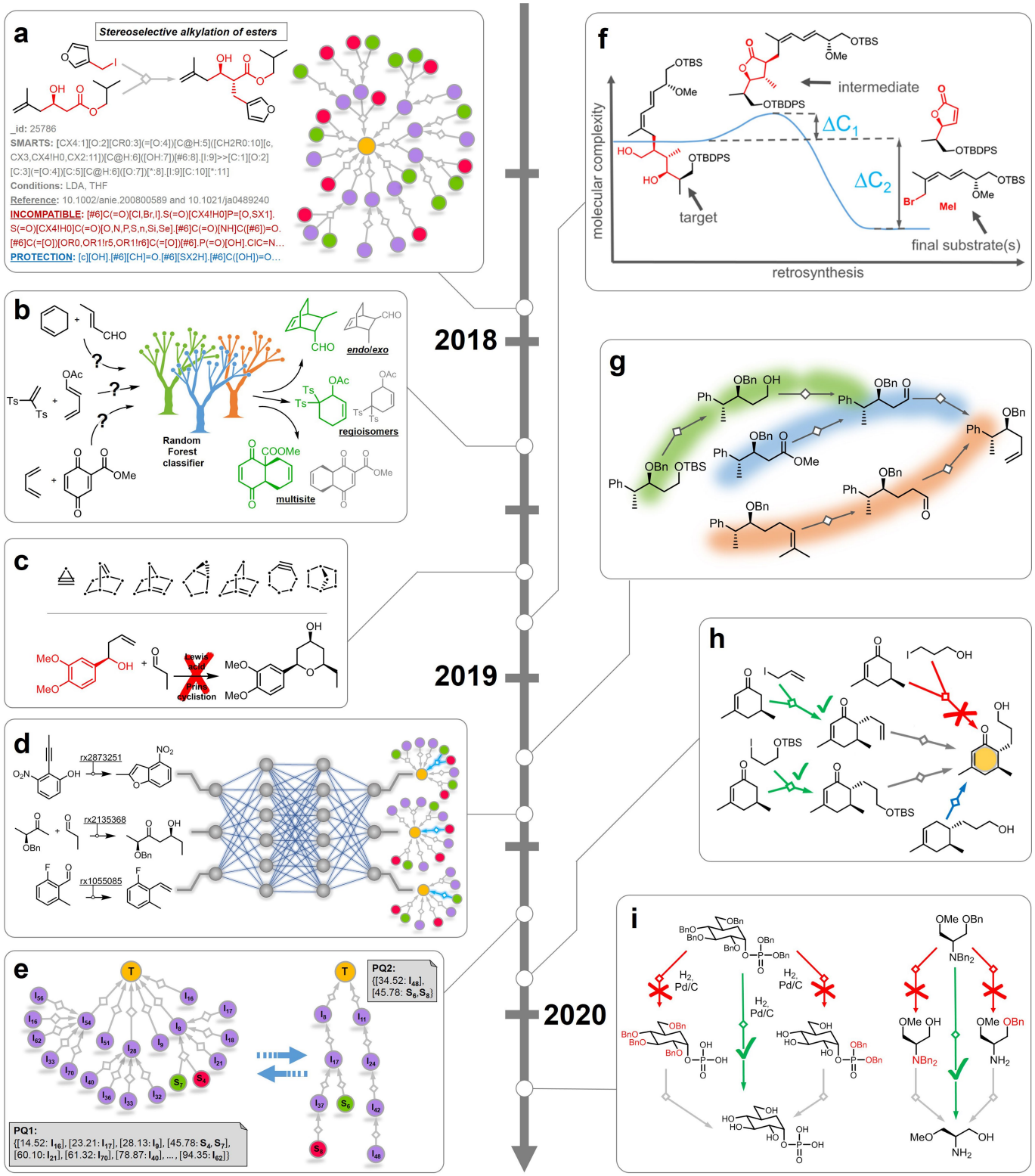
Competing interests Although Chematica was originally developed and owned by B.A.G.'s Grzybowski Scientific Inventions, LLC, neither he nor the co-authors currently hold any stock in this company, which is now property of Merck KGaA, Darmstadt, Germany. S.S., E.P.G. P.D., T.B., K.M. and B.A.G. continue to collaborate with Merck KGaA, Darmstadt. The algorithms described in this paper are currently being transitioned into Chematica's commercial version, called Synthia™. All queries about access options to Chematica/Synthia™, including academic collaborations, should be directed to S. Trice (sarah.trice@sial.com).

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2855-y.

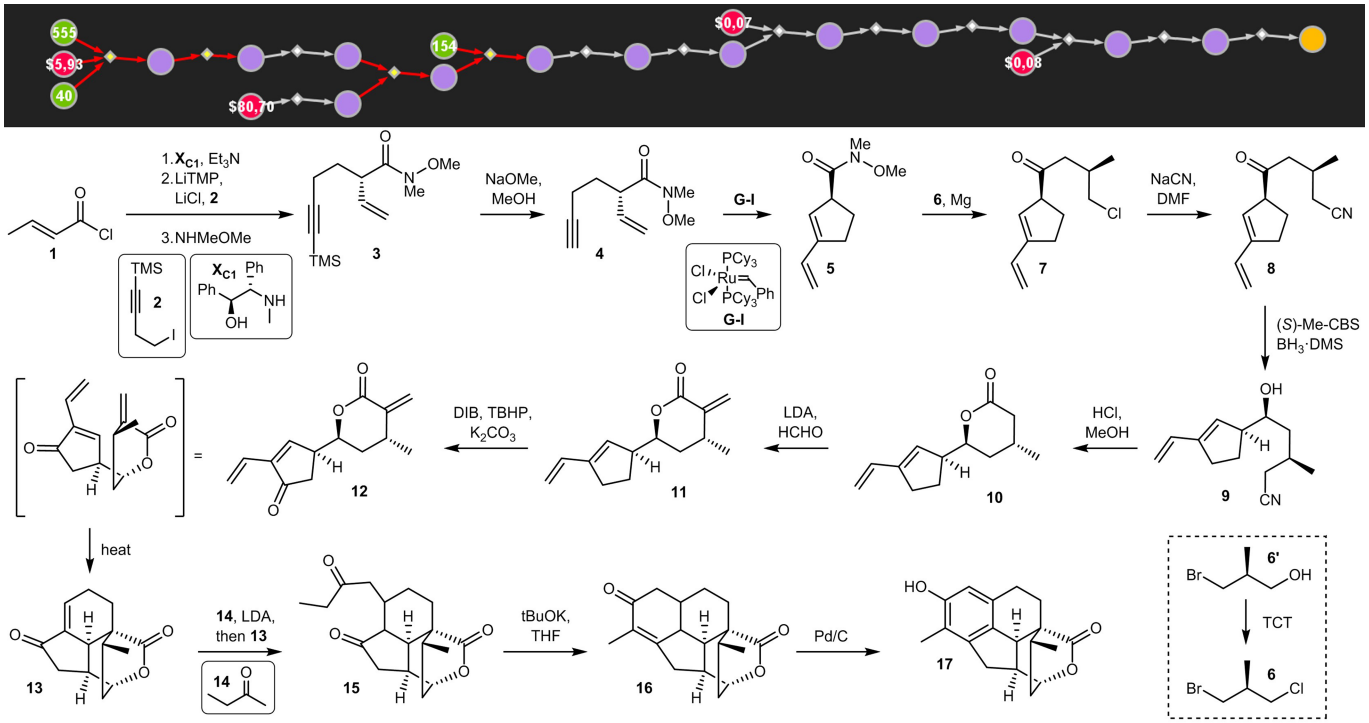
Correspondence and requests for materials should be addressed to K.M., J.M., M.M. or B.A.G.

Peer review information *Nature* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

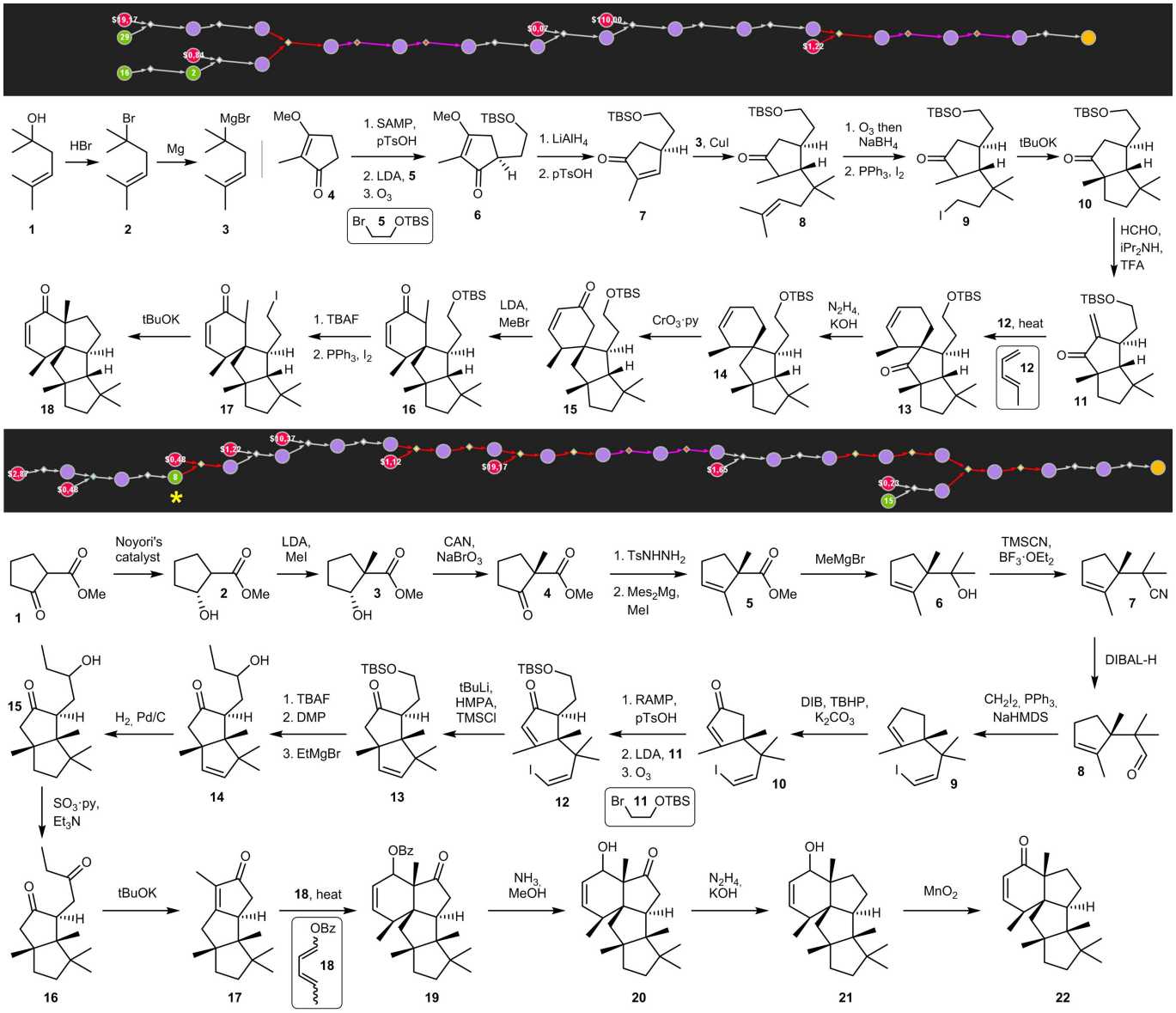


Extended Data Fig. 1 | See next page for caption.

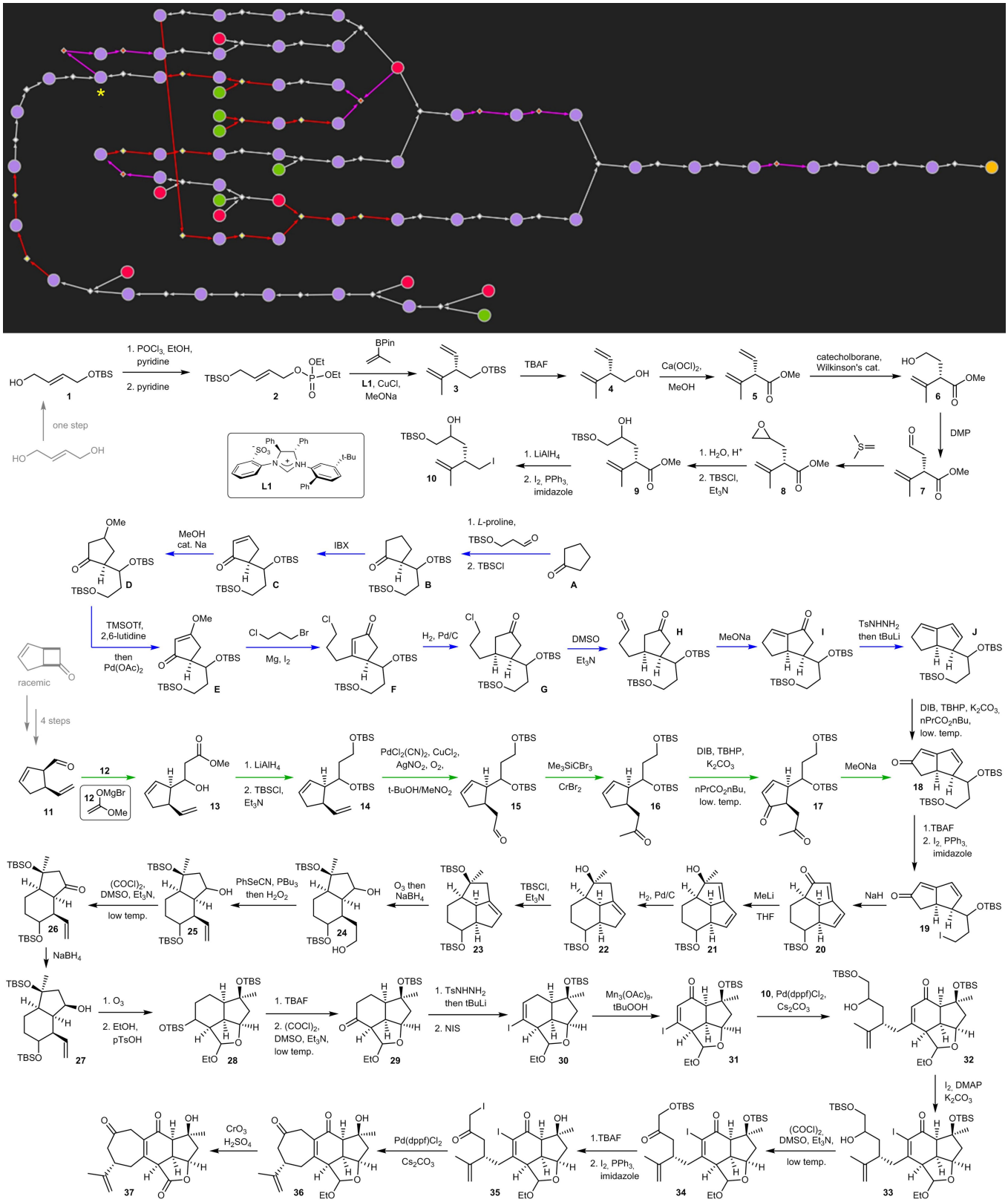
Extended Data Fig. 1 | New components of Chematica essential to its ability to plan the syntheses of complex, natural-product targets. Only key algorithmic improvements (since the publication of ref. [9]) are highlighted. **a**, Increase in the knowledge base of reactions rules to more than 100,000, including a large fraction of advanced stereoselective transforms. **b**, Implementation of various machine-learning molecular mechanics and quantum-mechanics routines to further evaluate the correctness of the reaction prediction. Illustrated here is the machine-learning method (random forest classifier) that evaluates the applicability of Diels–Alder cyclizations[21]. **c**, Information about specific motifs in the synthons that are not only too strained (top)[8] but also prone to side reactions. An electron-rich allylic alcohol substrate in the Prins cyclization may undergo a competitive oxonia-Cope rearrangement[59] (bottom). **d**, Scoring functions, either improved heuristics-based or best-in-class neural networks[22]. **e**, Search algorithms that combine two strategies: searching broadly to explore wide spectrum of options and deeply to reach stop-point substrates as soon as possible. Each search strategy maintains its own priority queue (PQ), with different queues sharing results. **f**, Large numbers of previously unrecognized two-step reaction sequences that allow the program to overcome local maxima of structural complexity. Image reproduced with permission from ref. [26] (https://doi.org/10.1016/j.chempr.2019.11.016; Elsevier), which is published under a Creative Commons license (CC BY-NC-ND 4.0; http://creativecommons.org/licenses/by-nc-nd/4.0/). **g**, Hard-coded sequences of some 100 FGIs to rapidly reach less reactive synthons. **h**, Bypasses—that is, routines that navigate around intermittent reactivity conflicts (red reaction arrow), by first converting the conflicting group into a non-conflicting one (here, a primary alcohol into an alkene or a silyl ether) and only then performing a high-gain, structure-simplifying step (here, stereoselective alkylation of cyclohexenone). Without the bypass algorithm, the search would explore other, less-structure-simplifying options such as the allylic oxidation indicated by blue arrow. **i**, The ability to perform two different reactions on the retron simultaneously, if multiple reaction loci are reactive under the reaction conditions. Here, treatment with hydrogen and Pd catalyst should remove both phosphonate esters and benzyl ethers (left). Under these conditions, only esters or only ethers cannot be selectively removed. Attempting such selective removal, Chematica would see the unremoved groups (marked in red) as incompatible; in effect, it would not be able to perform the desired global deprotection. Similarly, global debenzylation of an aminoalcohol should be performed in a single step (right).



Extended Data Fig. 2 | Enantioselective synthesis of a pentacyclic diterpenoid, cephanolide B, designed by Chematica. This target was recently prepared in racemic form in 12 steps[48], with Pd-catalysed carbonylative cyclization as the key step. In its design, Chematica used 13 steps to reach commercially available crotonyl chloride and a known iodoalkyne **2** (available in two steps from the commercially available oxirane and TMS-acetylene). The synthesis commences with the formation of enantioenriched diene **5** via stereoselective alkylation of the enolate (with stereochemistry controlled by a chiral auxiliary) and subsequent metathesis of the enyne **4**. Subsequently, addition of a Grignard reagent derived from bromide **6**, cyanation, reduction of ketone, lactonization, methylenation and oxidation of the less hindered allylic position derives triene **12**. This is then used in an elegant, intramolecular Diels–Alder cycloaddition (the feasibility of which was confirmed separately by molecular-mechanics calculations) to form the tetracyclic skeleton of cephanolide B. The synthesis of the target is then accomplished via the (non-intuitive) construction of the aromatic part via Robinson annulation of **13** with butanone **14** and oxidation of the thus-obtained enone.

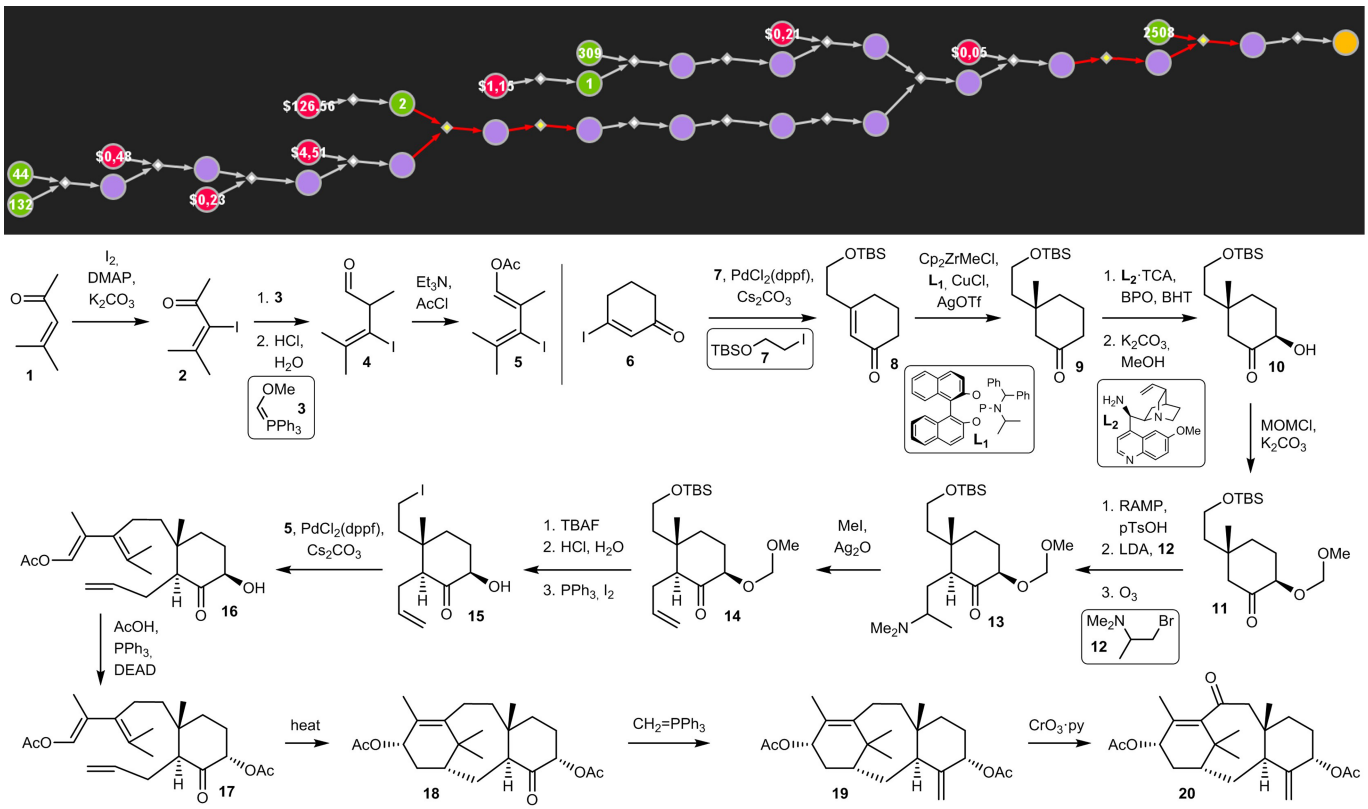


Extended Data Fig. 3 | Enantioselective synthesis of a cyclopiane diterpene, conidiogenone B, and its derivative designed by Chematica. Synthesis of conidiogenone B, which includes a challenging 6–5–5–5 ring system and six contiguous stereocentres (of which three are quaternary), was recently accomplished in 14 steps[49] (starting from trimethylcyclopentenone, itself one step from a buyable substrate) and relied on a substrate-controlled Nicholas/Pauson–Khand reaction and Danheiser annulation. Chematica's plan (top panel) also uses 14 steps and relies on intramolecular alkylations to construct five-membered rings and Diels–Alder cycloaddition to build the six-membered ring of conidiogenone B. The synthesis commences with the chiral-auxiliary-controlled alkylation of cyclopentenone **4** with protected bromoethanol **5** to install the first stereocentre. Subsequent Stork–Danheiser transposition is followed by a substrate-controlled addition of a tertiary organocuprate and intramolecular alkylation to yield the bicyclic ketone **10**, which is further methylenated to enone **11**. Formation of the six-membered ring of conidiogenone B is accomplished via the Diels–Alder reaction of **11** with diene **12** to give the tricyclic ketone **13**, which is further elaborated into iodoketone **17**. Formation of the last ring of conidiogenone B is accomplished via the intramolecular alkylation of the ketone. In the bottom panel, Chematica was asked to design a plan for a more complex derivative of conidiogenone B, which differs by an extra methyl group (at a new quaternary stereocentre). Within 18 steps from the target, Chematica reached a known enantioenriched ketoester **4** (marked with a yellow asterisk) which was then sourced, in a few minutes of additional searching, to the commercially available and inexpensive **1**. The synthesis commences with the reduction of the ketone (with stereochemistry controlled by Noyori's catalyst). Subsequent substrate-controlled alkylation and oxidation are followed by elaboration of ester **4** into iodoenone **10**. Stereoselective alkylation with protected bromoethanol **11** and subsequent cyclization yields the bicyclic ketone **13**, which is further elaborated to tricyclic enone **17**. We make two notes here. First, owing to the presence of a matched stereocentre, conversion of **10** to **12** could probably be performed as one step, without Enders' auxiliary to control the stereochemical outcome. Chematica did not recognize this possibility, probably because it has not yet been taught detailed rules that govern substrate-directed alkylations controlled by quaternary stereocentres. Second, desmethyl analogue of enone **17** was also used in the published synthesis of conidiogenone B, but, to form the six-membered ring, it was subjected to Danheiser annulation followed by ozonolysis-aldol condensation rather than to Diels–Alder cyclization. The formation of the last ring of conidiogenone B is accomplished via intermolecular Diels–Alder reaction with electron-rich diene **18** (available in a single step from pent-3-enal) approaching from the less hindered face of the enone (see refs. [60–62] for similar Diels–Alder cyclizations promoted by Lewis-acid catalysts). From this point, the target molecule is obtained in three straightforward steps.

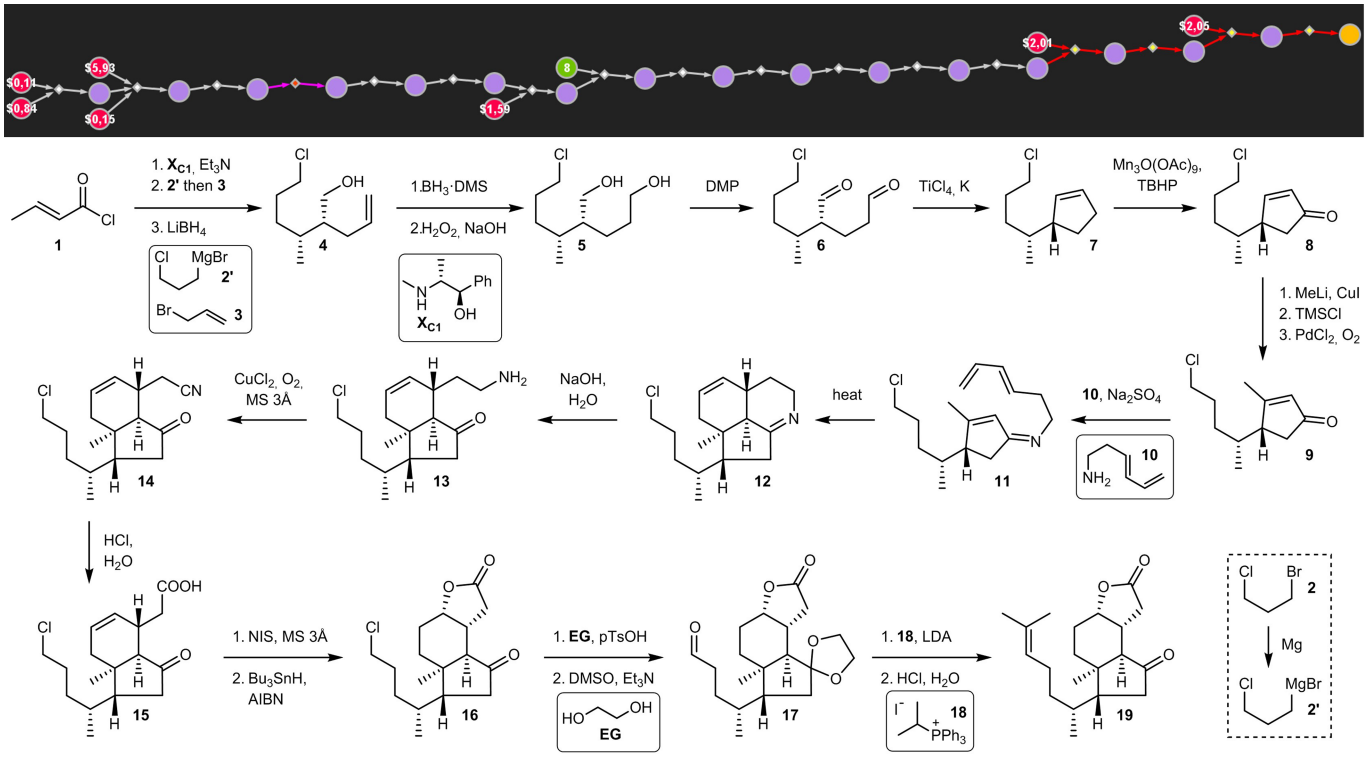


Extended Data Fig. 4 | See next page for caption.

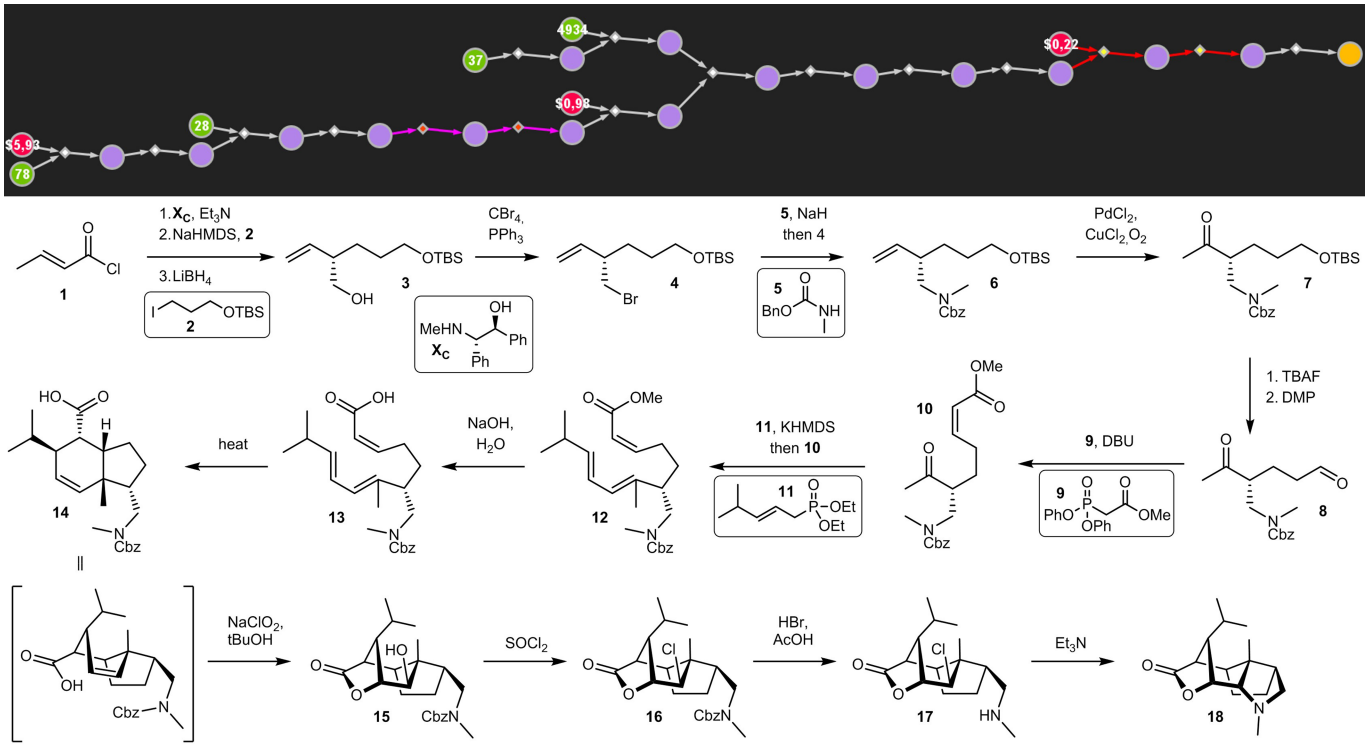
Extended Data Fig. 4 | Chematica's synthetic plan for scabrolide A. Scabrolide A is a polycyclic furanobutenolide-derived norcembranoid diterpenoid that belongs to a family of marine natural products isolated from *Sinularia* soft corals[63,64]. The molecule poses a synthetic challenge owing to its compact, densely functionalized core: a fused 5–6–7 carbocyclic scaffold decorated with five adjacent stereocentres and one additional remote stereocentre on the seven-membered ring. A recent literature pathway[50] (to the enantiomer from ref. [64]) comprises 21 synthetic steps and relies on the intramolecular Diels–Alder cycloaddition and late-stage [2+2] photocycloaddition/fragmentation sequence. During computer planning of the enantiomer from ref. [63], several constrains were imposed; for example, Chematica was asked to design an enantioselective strategy (using the REMOVE_DIAST variable to exclude reactions that lead to a single racemic diastereoisomer), and was not allowed to use SAMP or RAMP hydrazones (to minimize the use of chiral auxiliaries), or highly strained bridgehead intermediates. The route proposed by the software is longer (about 30 steps) and more conservative in the sense that it relies on only broadly applicable chemistries. When planning its route, Chematica did not know the highly scaffold-specific (though elegant) fragmentation–recombination–elimination sequence of steps used towards the end of the literature pathway. The synthesis proposed by the machine relies on an intramolecular aldol addition of **17** followed by FGI, which sets the scene for the closure of a six-membered ring via alkylation reaction to yield intermediate **20**. Subsequent substrate-controlled, stereoselective addition installs the tertiary alcohol. Reduction (with double-bond migration) of intermediate **21** followed by reductive ozonolysis sets the scene for the construction of the second five-membered ring of scabrolide's scaffold. The fourth and final, seven-membered ring is closed via Pd-mediated coupling. The starting material initially identified by the software (aldehyde **11**) is not commercially available, but can be sourced in four steps from (±)-*cis*-bicyclo[3.2.0]hept-2-en-6-one. Looking for alternative endings of the pathways, that terminate in commercially available, achiral and inexpensive starting materials, we restarted the search from a node marked in the graph view (top) by a yellow asterisk (bicyclic intermediate **18**). The alternative ending (blue reaction arrows in the bottom scheme) was found within about half an hour and commenced from readily available, protected hydroxyaldehyde and cyclopentanone. The initial ending, starting from the aldehyde **11**, is marked by green arrows.



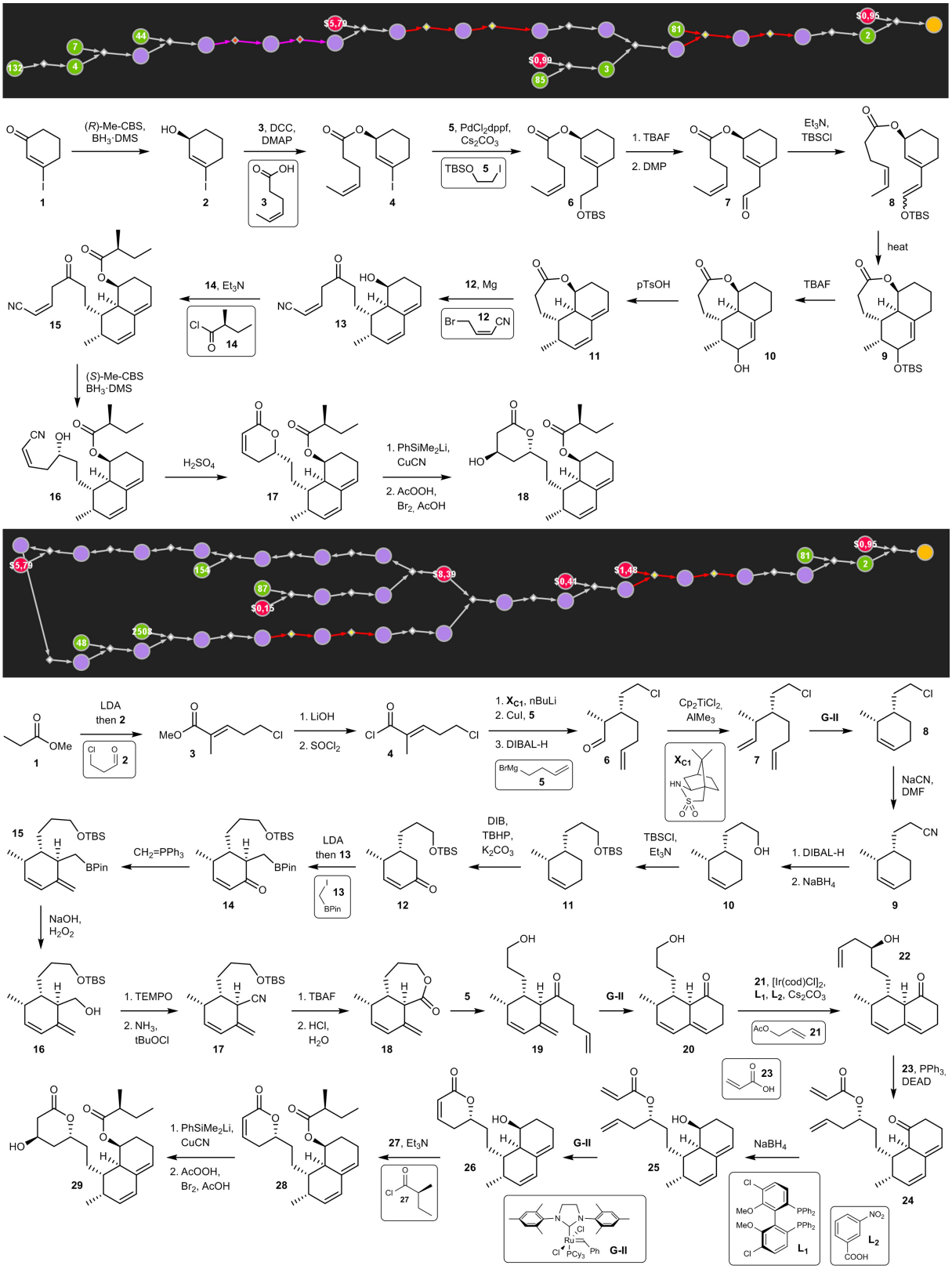
Extended Data Fig. 5 | Chematica-designed, enantioselective synthesis of taxuyunnanine D, a less oxidized taxane. The previous synthesis[51,65] of this target was accomplished in 12 steps via a two-phase cyclase-oxidase strategy, and required extensive exploration of conditions to achieve satisfactory selectivity during C–H oxidations. Here, within 14 steps from the target molecule, Chematica reached simple and known starting materials: iodocyclohexenone **6** and protected iodoethanol **7**. The synthesis commences with the Pd-mediated coupling of **6** and **7**. Subsequent catalyst-controlled methylation and oxidation introduce the all-carbon quaternary and C5 hydroxylated stereocentres of taxuyunnanine D. Subsequently, protection of alcohol, stereoselective alkylation of cyclohexanone (with proposed Enders' auxiliary controlling the stereochemical outcome, but probably also feasible when performed directly; see notes in the caption of Extended Data Fig. 3), Hofmann elimination, removal of protecting groups and Appel reaction yield iodide **15**, which is coupled with iododiene **5** (available in four steps from enone **1**) to give triene **16**, setting the scene for the key formation of the taxane skeleton via electron-neutral intramolecular Diels–Alder cycloaddition (such an electronically neutral system that lacks electron-withdrawing groups may require activation with high temperature or a transition-metal catalyst[66]). Formation of taxuyunnanine D from the [4+2] cycloadduct **18** is then accomplished in two steps and requires olefination of ketone and allylic oxidation. The latter step appears less risky compared to the known solution[51], because **19** lacks any competitive allylic CH_2 groups, which are prone to oxidation and could cause selectivity problems.



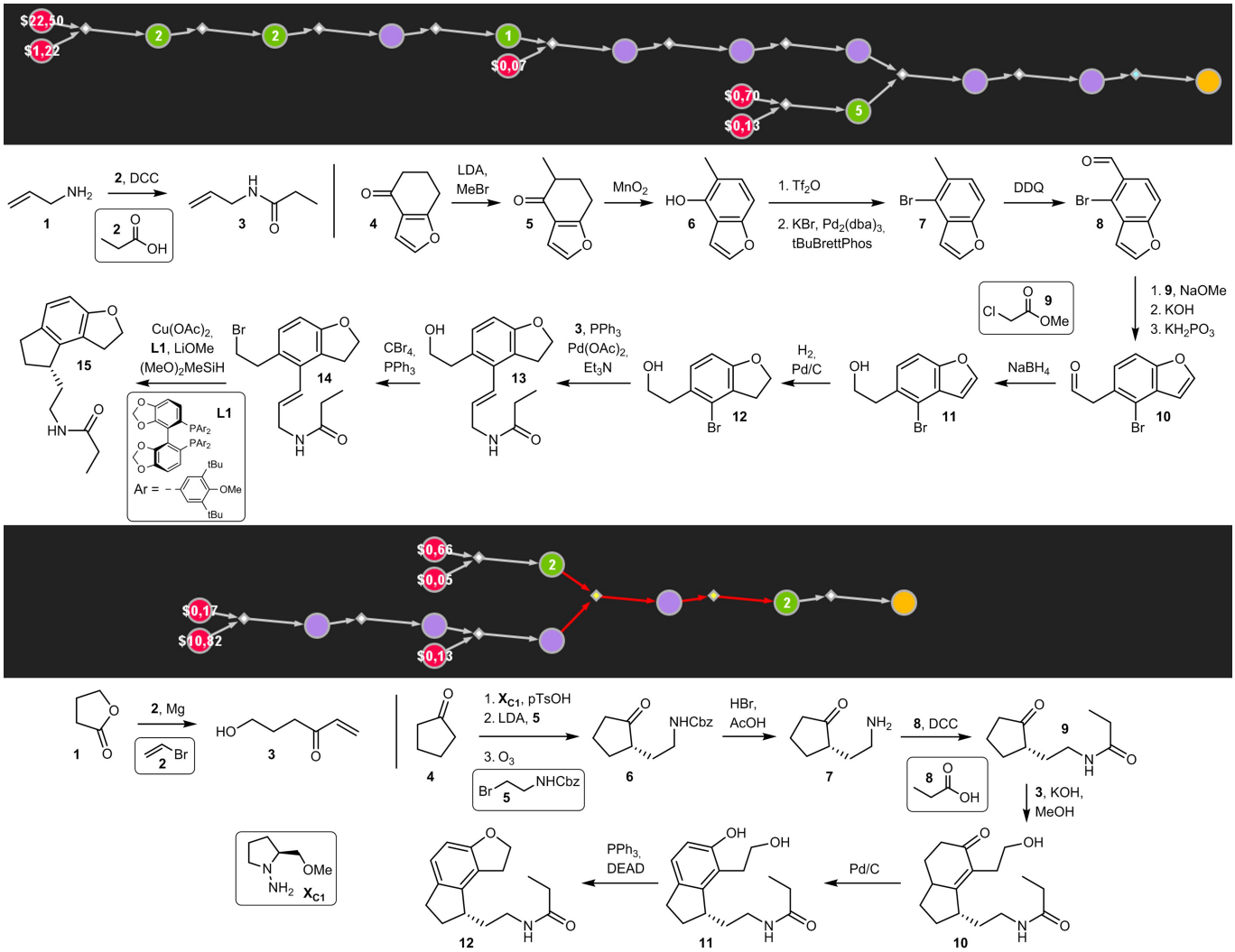
Extended Data Fig. 6 | Chematica-designed, enantioselective synthesis of a marine steroid, aplykurodinone-1. Prior syntheses[67] of this target, featuring six contiguous stereocentres, either relied on the late-stage introduction of the side chain via Michael addition to cyclopentenone (which suffers from low selectivity), or started[67] from chiral building blocks (in the latter case, in 11 steps but from much more advanced, chiral substrates). Chematica used 17 steps to reach achiral and commercially available substrates: crotonyl chloride, allyl bromide and bromochloropropane **2**. This synthesis commences with the installation of two contiguous stereocentres via stereoselective *vic*-difunctionalization of unsaturated amide and subsequent hydroboration and bisoxidation, followed by McMurry coupling to give cyclopentene **7**. From there on, oxidation of the less hindered allylic position, methylation of cyclopentenone, reoxidation and formation of imine (elegantly ensuring that a single regioisomer would form in the Diels–Alder reaction) with aminodiene **10** (available in four steps from ethyl sorbate) derives triene **11**, which is then used in an intramolecular Diels–Alder cycloaddition that forms the desired 6–5 ring system of aplykurodinone-1. Hydrolysis of the imine linker and conversion of the primary amine to the carboxylic acid via oxidation and hydrolysis yields **15**, which is then subjected to iodolactonization followed by dehalogenation to form the entire 5–6–5 ring system. The synthesis is completed by elaborating the remaining alkyl chloride to the desired alkene.



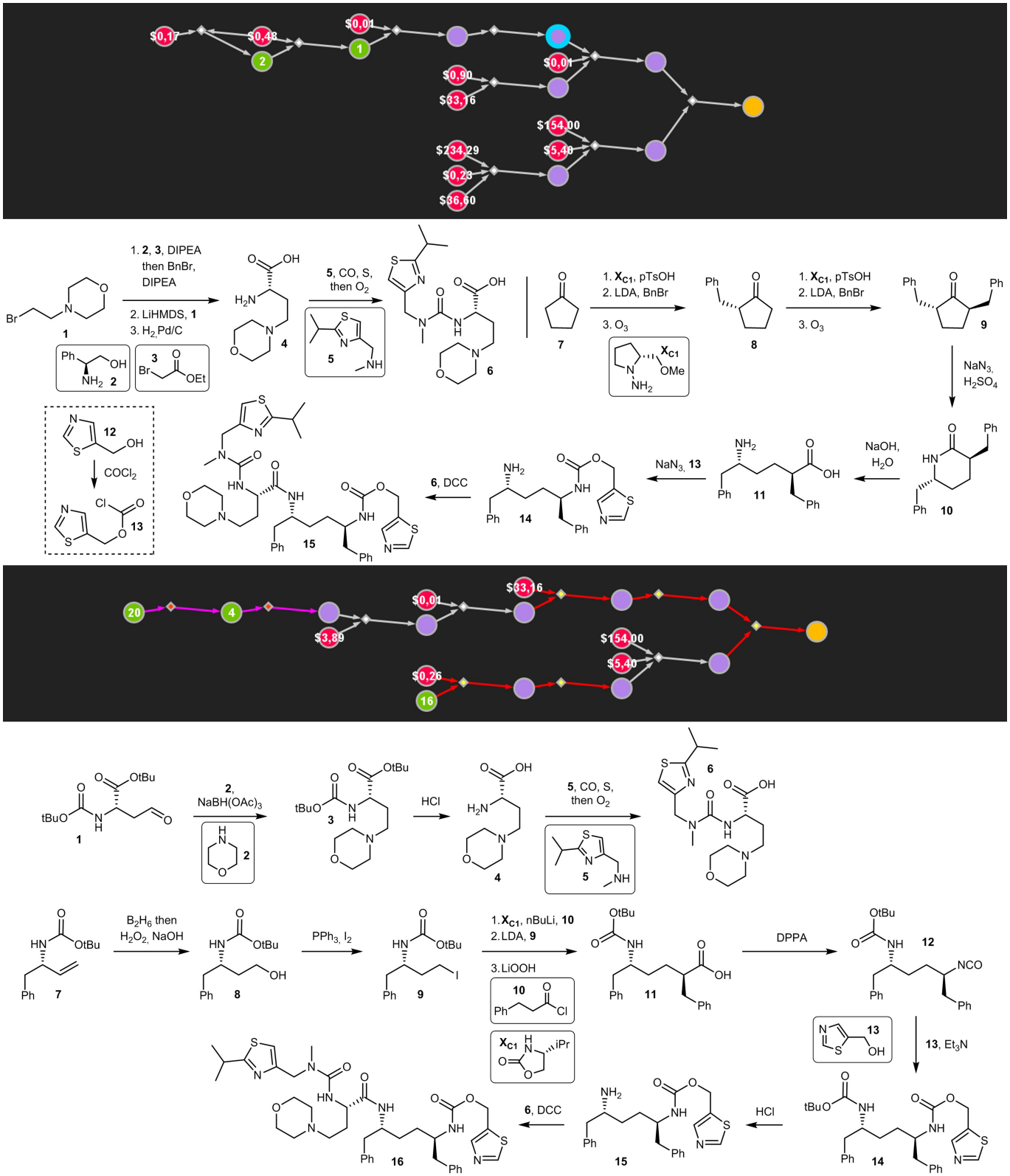
Extended Data Fig. 7 | Chematica-designed enantioselective synthesis of a tetracyclic alkaloid, dendrobine. Synthesis of this target, which features a challenging 5–6–5–5 ring system and seven contiguous stereocentres, was performed recently[53] in 11 steps, taking advantage of enantioselective Diels-Alder reaction, substrate-controlled hydroboration and reduction of imine as the key steps. In Chematica's synthetic plan, within 14 steps from the target, the software reached the commercially available crotonyl chloride and known 3-iodopropanol (that is, simpler starting materials than the Danishefsky's diene and unsaturated imide used in the literature synthesis). The synthesis commences with the chiral-auxiliary-controlled alkylation of the amide enolate. Ensuing steps allow for the preparation of enoate **10**. Further homologation with allylic phosphonate **11** (available in two steps from an appropriate alcohol) and hydrolysis yield the triene **13**, setting the scene for an intramolecular Diels–Alder reaction that forms the desired 6–5 ring system. Subsequent hydroxylactonization gives tricyclic alcohol **15**, which is then efficiently transformed into the target molecule via stereoretentive chlorination of the alcohol, Cbz removal and substitution of chloride.



Extended Data Fig. 8 | Enantioselective synthesis of mevastatin designed by Chematica with all its reaction knowledge and on exclusion of user-specified reaction types. Top, synthetic plan obtained when the program was allowed to use all of its reaction knowledge base. Under these circumstances, the planned route relies on an intramolecular Diels–Alder reaction to construct mevastatin's 6–6 ring system. The synthesis commences with stereoselective reduction of a ketone to give iodoalcohol **2**, which is transformed in five steps into triene **8**. Subsequent cycloaddition (note that such an electronically neutral system that lacks electron-withdrawing groups may require activation with high temperature or a transition-metal catalyst[66]) and elaboration of the side chain give the target molecule in the total of 14 steps. Bottom, synthetic plan designed by Chematica when it was forbidden from using the key Diels–Alder reaction and was thus forced to come up with a completely different approach; the synthesis is now much longer. The formation of each ring is accomplished via ring-closing metathesis.



Extended Data Fig. 9 | Pathways leading to ramelteon designed by the software with and without multistep strategizing routines. The top synthetic pathway was designed without the new, multistep heuristics. The scaffold of the target was constructed via Cu-catalysed hydroalkylation of alkenes[68]. Although the pathway does not contain chemically erroneous steps, it is long, relies heavily on reductions and oxidations, and involves many FGIs. The bottom route, designed with the new strategizing routines, is more concise and elegant. The key element in this path is a strategy that relies on Robinson annulation followed by dehydrogenation of enones (in the retrosynthetic direction, when planning the route, the program strategizes and first performs a seemingly unproductive dearomatization of a phenol, which then enables Robinson annulation).



Extended Data Fig. 10 | Pathways leading to tybost designed by the software with and without multistep strategizing routines. The top synthetic pathway was designed without the new multistep algorithms. This route is longer and requires additional protection and deprotection operations on intermediate **11** (node in blue halo). The program was not able to find better routes even after hours of searching. In the bottom route, when the program was allowed to strategize, it found a more elegant route that relies on two bypasses (two sets of red reaction arrows) and one FGI (pair of violet reaction arrows). The software navigated the pathways to starting materials that already had relevant groups protected (such that no protections were required mid-way into the pathway) and were easily available from appropriate amino acids.