Chemical Science

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2019, **10**, 4640

All publication charges for this article have been paid for by the Royal Society of Chemistry

Selection of cost-effective yet chemically diverse pathways from the networks of computer-generated retrosynthetic plans†

Tomasz Badowski,‡[a] Karol Molga ‡[a] and Bartosz A. Grzybowski *[ab]

Received 16th December 2018
Accepted 24th February 2019

DOI: 10.1039/c8sc05611k

rsc.li/chemical-science

As the programs for computer-aided retrosynthetic design come of age, they are no longer identifying just one or few synthetic routes but a multitude of chemically plausible syntheses, together forming large, directed graphs of solutions. An important problem then emerges: how to select from these graphs and present to the user manageable numbers of top-scoring pathways that are cost-effective, promote convergent *vs.* linear solutions, and are chemically diverse so that they do not repeat only minor variations in the same chemical theme. This paper describes a family of reaction network algorithms that address this problem by (i) using recursive formulae to assign realistic prices to individual pathways and (ii) applying penalties to chemically similar strategies so that they are not dominating the top-scoring routes. Synthetic examples are provided to illustrate how these algorithms can be implemented – on the timescales of ~1 s even for large graphs – to rapidly query the space of synthetic solutions under the scenarios of different reaction yields and/or costs associated with performing reaction operations on different scales.

Introduction

Recent years have witnessed the revival of interest in computer-assisted retrosynthetic planning, which has been an elusive goal since the late 1960s.[1–11] With foundational work on the representation of large collections of chemical reactions as networks[12–14] and the so-called bipartite graphs[15–18] and with modern hardware and algorithms allowing for rapid searches for synthetic pathways, synthetic planning by computers has finally become a tangible possibility. Indeed, several software platforms[7–11,19,20] have been developed differing in the details of search algorithms and also in the origin of synthetic rules (expert-coded[19,20] based on reaction mechanisms *vs.* automatically extracted from the literature[7–11]). The year 2018 also marked the first demonstration[20] – on our Chematica platform – of autonomous computer design and subsequent experimental validation of multiple efficient syntheses leading to medicinally important targets. Despite this undeniable progress, however, several challenges remain and need to be considered, especially if the programs are to be adopted by practicing organic chemists. One of the challenges we consider here is how to present to the program's user synthetic solutions that are not only viable but also economical and chemically diverse.

In the early stages of its development, Chematica was able to identify relatively small numbers of viable syntheses which were often variations of a similar synthetic theme. With the increasing knowledge base of reactions and with improved algorithms for the exploration of synthetic options,[19,20] however, the searches started to identify increasingly large numbers of chemically correct solutions which themselves formed large synthetic networks (*cf.* Fig. 1). The question then arose how to estimate and rank the realistic costs of these possible pathways, taking into account not only the absolute number of steps and the costs of starting materials but also the path structure – that is, its linearity *vs.* convergence, the placement of the convergence points within the pathway, or the optimal "timing" to use the most expensive reagents (see examples in Fig. 2). In addition, because organic molecules can be made in different ways and the ultimate choice of a pathway often reflects practical considerations (ranging from the availability of certain reagents or equipment to the familiarity of a given chemist with particular types of reactions/procedures), it is important to present to the user multiple choices differing in the key reactions they entail. We note that although the problems of (i) finding a desired number of the best/lowest-cost solutions within the so-called directed graphs with weighted nodes (*e.g.*, in random time-dependent networks,[21–23] transit networks,[21,24,25] or reaction networks[19,20,26]) and also (ii) identifying qualitatively different pathways (*e.g.*, within transportation networks[27]) have been individually studied in graph theory, the specific approaches are not easily extendable to realistic synthetic–organic planning (*cf.* Discussion in Section S3†). Curiously,

[a]*Institute of Organic Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, Warsaw 01-224, Poland. E-mail: nanogrzybowski@gmail.com*

[b]*IBS Center for Soft and Living Matter, Department of Chemistry, UNIST, 50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan, 689-798, South Korea*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c8sc05611k

‡ Authors contributed equally.

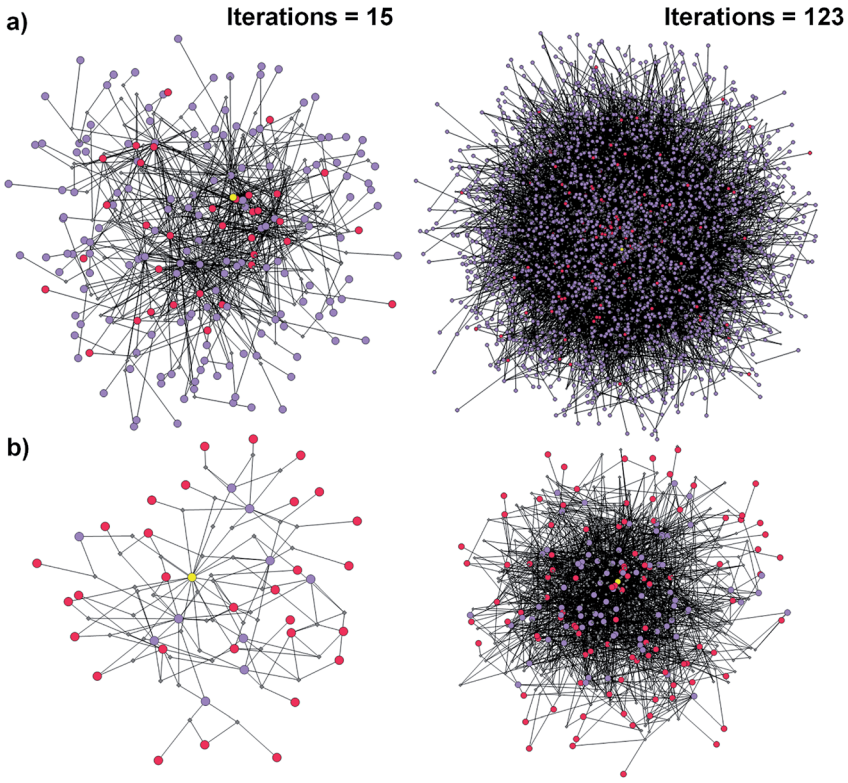

Fig. 1 Reaction networks/graphs inspected during synthetic searches and the subgraphs of viable solutions. (a) The graphs of all molecule nodes visited during Chematica's retrosynthetic searches for the syntheses of the triarylamine target (same as in Fig. 6). The network on the left is for the early stage of the search (15 iterations of expanding retrons into progeny synthons; overall within ~20 s computing time on a 64-core machine) and contains 456 nodes (*i.e.*, all molecules and reactions considered during the search). The graph on the right is later in the search (123 iterations within <2 min of computing time on a 64-core machine) and contains 5300 nodes. (b) The corresponding subgraphs of the networks from (a) contain only the viable syntheses. Note that as the search progresses, the subgraphs of solutions are themselves becoming quite complex, here increasing from 90 nodes after 15 iterations to 779 nodes after 123 iterations. Color coding of the nodes: yellow = target; violet = intermediates; red = commercially available starting materials; small gray diamonds = reaction nodes.

neither cost effectiveness nor diversity has been addressed in the existing retrosynthetic platforms, which might explain why the relevant publications usually describe just one top-scoring solution (diversity issue) and why the published pathways are often linear rather than convergent sequences (lack of realistic treatment of costs and yields). In our earlier versions of Chematica,[16,19,20] the selection algorithms were also rudimentary and the cost-calculation schemes were not only extremely slow (running for thousands of seconds for large graphs of solutions) but also did not properly capture the efficiencies of individual steps and the overall path structure (linearity *vs.* convergence), translating into unrealistic costs of starting materials consumed and/or reaction operations performed (*cf.* Section S3.1†). In light of these considerations, we see the improved approaches – reflecting true synthetic costs and operating within just seconds – described in the current paper as an important advance not only for Chematica but also for other efforts in this exciting area of research.

Computer-assisted retrosynthetic searches rely on iterative expansion of the parent/retron nodes into daughter/synthon nodes and on navigating the thus-created synthetic space (with the help of various scoring functions) to ultimately reach simple and commercially available substrates. Since the search procedures are not the subject of our current work (for details, see ref. 9, 19 and 20), the starting point for our analyses is an already existing large graph of molecules considered/"visited" during synthetic planning. In a more technical parlance, we consider a large directed bipartite graph (Fig. 1a) composed of two types of nodes: molecules represented in all figures as circular nodes and reactions represented as smaller diamond-shaped nodes. The molecule nodes are of three types: the target (marked *yellow* in the figures), its progeny nodes (in specific chemical examples in Fig. 6–10 colored *green* if a molecule is known in the literature[19,20] and *violet* otherwise) corresponding to synthetic intermediates, and commercially available starting materials (*red*). To enable meaningful cost estimates of the synthetic pathways, the starting materials must have realistic prices standardized to a certain common quantity – in Chematica, there are over 200 000 such nodes from the Sigma-Aldrich catalog and their prices are all standardized to "per gram," which is easily convertible to "per mmol" we use here. The reaction nodes carry with them some "fixed cost" of performing a reaction operation *r* to obtain some unit quantity of the product – this cost can be loosely construed as a cost of labor plus equipment/solvent/purification and does not yet account for

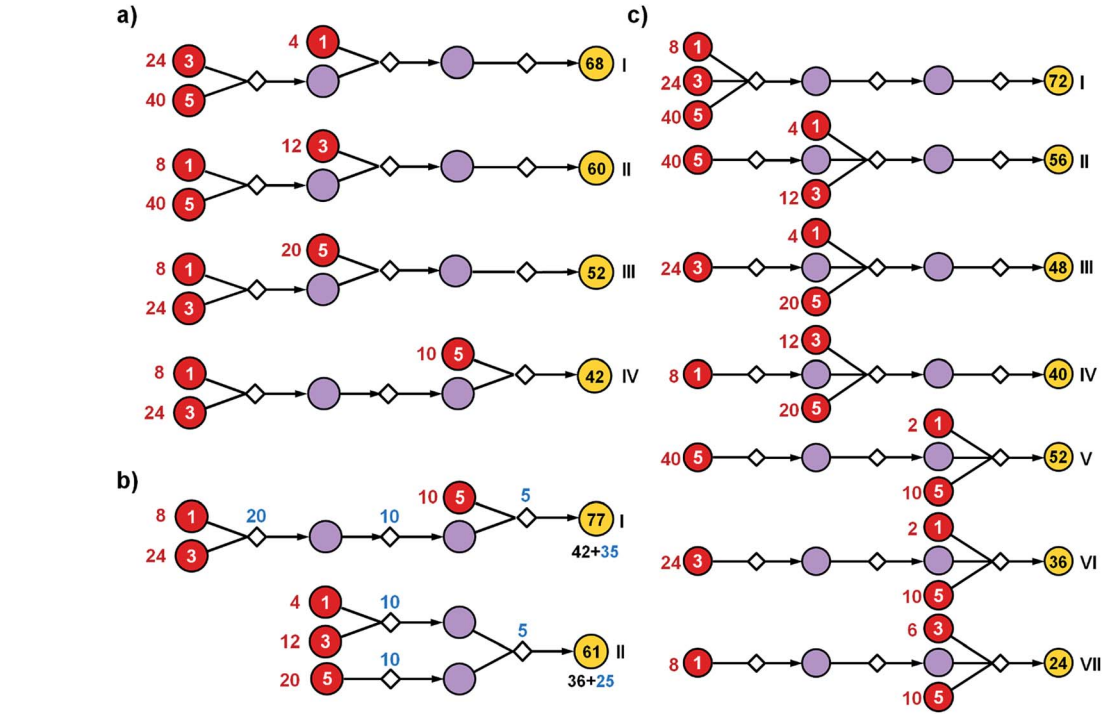

Fig. 2 Effects of path structure on the cost of syntheses. The schemes illustrate hypothetical synthetic plans in which 1 mmol of the final product (yellow nodes on the right) is made from the same substrates (red nodes with prices per mmol given by the white numbers, *i.e.*, 1, 3 or 5 of some currency). All syntheses use three reactions. The reaction operations in this bipartite representation are indicated by small diamond nodes and are assumed to proceed in 50% yield each. Intermediates are denoted by violet circles. Black numbers within yellow circles give the calculated cost of making 1 mmol of the target while red numbers next to the substrate nodes are costs of the starting materials that need to be used for this purpose. (a) More expensive substrates should be introduced later in the synthesis. Here, placing the substrate with cost "3" at the second step (II) lowers the cost of materials needed for the synthesis of 1 mmol of the target from 68 to 60. Likewise, as the substrate with cost "5" is introduced in the first (I), the second (III), or the third step (IV), the overall cost is progressively lowered, from 60 to 52 to 42. (b) The realistic overall cost of synthesis should also incorporate the cost of reaction operations per some unit scale (here "5" for making 1 mmol of a given reaction product). Blue numbers next to reaction nodes denote the scale required for each 50%-yield reaction to ultimately produce 1 mmol of the target. Given this 50% yield of each step and with reference to the optimal solution (IV) from panel (a), 4 mmol of the product of the first step has to be prepared (cost $5 \times 4 = 20$), 2 mmol of the product of the second step (cost $5 \times 4 = 10$), and ultimately 1 mmol of the target (cost $5 \times 1 = 5$). Overall, the cost of the pathway is 77 and is the sum of materials' costs (42) and reaction operations' cost (35). In contrast, with the same yields, the more convergent approach marked as (II) reduces both the labor cost (25 *vs.* 35) and the cost of starting materials (36 *vs.* 42). (c) A multicomponent reaction (MCR) used *en route* to a given target offers most significant cost savings if this "convergence point" is placed later within the pathway and when it uses the most expensive substrates. Here, these conditions are met for (VII).

the prices of specific substrates and/or reaction yield that are considered only when evaluating specific pathways. The algorithms we describe below do not change if the "fixed costs" are the same or different for different reaction types (as in the example in Fig. 3). Finally, the reactions are assumed to proceed with a certain yield – although yields of each reaction in the network can be estimated individually (by machine-learning[28] or by thermodynamic models[29]), we assume here, without losing generality of the algorithms, that the yields of all reactions in the graph are the same. In Chematica, the specific value of such a "global"/average yield can be set by the user allowing him/her to query the graph of synthetic solutions under different yield scenarios.

Results and discussion

Scoring and selecting cost-optimal pathways

With algorithmic details described in the ESI, Section S1,† the general procedure for pathway selection is illustrated in Fig. 3. Within the initial network in Fig. 3a, we define (i) chemical nodes as "synthesizable" if they are targets of at least one synthetic pathway tracing back to commercially available substrates and (ii) reaction nodes as "viable" if all their substrates are synthesizable. In the first step, the algorithm finds all synthesizable nodes in the network in a depth-first-search-like manner and using the fact that a chemical is synthesizable only if it is commercially available or is a product of some viable reaction. If the target is not among the synthesizable nodes, then the selection algorithm stops without returning any pathways. Otherwise, it proceeds as follows. A subgraph of the network induced by synthesizable nodes is computed and retained (Fig. 3b). This step removes all substance nodes that are not synthesizable and reactions that are not viable. Then, the remaining subgraph is further restricted to one induced by ancestors of the target and the target itself (Fig. 3c). This step removes nodes which do not belong to any pathways leading to the target. Over the remaining subgraph, called the solution graph, the cost of each chemical node is taken as the smallest

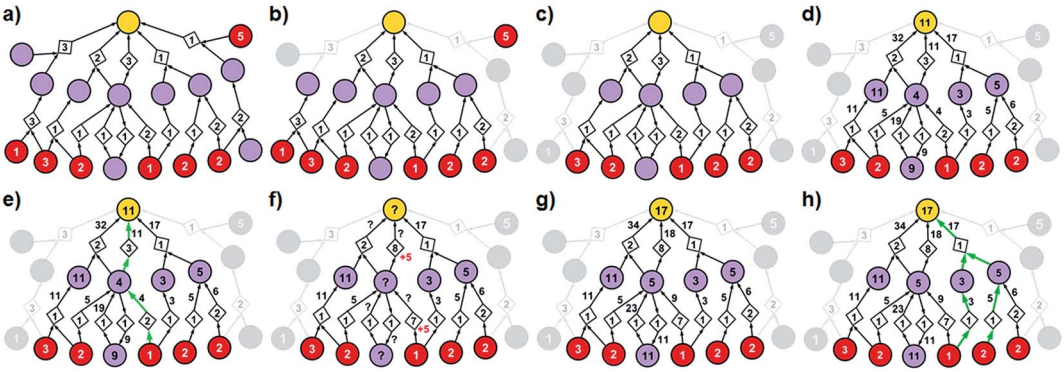

Fig. 3 Stages of selecting cost-effective yet diverse pathways from a synthetic graph. Parameters used are 50% yield for all reactions and penalty $P = 5$. (a) A hypothetical chemical reaction network created during a retrosynthetic search. Hypothetical costs of substrates per mmol are given over red nodes and fixed costs of reaction operations are indicated inside the diamond-shaped reaction nodes. Note that not all pathways terminate in commercially available starting materials (red nodes) as the search algorithm visited/probed some intermediates that did not lead to complete synthetic solutions. Such nodes and the pathways they are involved in are removed from consideration in (b) and (c). (d) The costs of all nodes in the remaining subgraph are computed by propagating from starting materials to the target as described in detail in the main text (see also Fig. 4). (e) The lowest-cost synthesis of the target is selected and here indicated in green. (f) Penalty P is added to the reactions from the selected pathway (here, $P = +5$, red numbers). Nodes whose costs increase due to such penalization are marked with question marks and are recalculated as in (g). The new "best" synthetic pathway is selected in (h) and the penalization-selection cycle can be again repeated as needed.

cost of all syntheses that can produce this chemical. The cost of any reaction node in the network is the smallest cost of all synthetic pathways containing this particular reaction and giving this reaction product. Accordingly, for each non-starting-material chemical, c, in the network, its cost is prescribed recursively by $\mathrm{cost}(c) = \min_{r \in \mathrm{pred}(c)}(\mathrm{cost}(r))$, while for each reaction node we have $\mathrm{cost}(r) = \mathrm{fixed\_cost}(r) + \sum_{c \in \mathrm{pred}(r)} \frac{\mathrm{cost}(c)}{\mathrm{yield}(r)}$, where fixed_cost($r$) was discussed in the preceding paragraph (cost of performing synthetic operations on some unit scale) and pred denotes the set of predecessors of a given node in the network. In the subsequent step, the costs of all nodes in the network are calculated bottom-up (*i.e.*, from the starting materials to the target) using a Dijkstra-like algorithm similar to the one for finding minimum-weight B-paths in weighted hypergraphs[24] (see also ref. 26).

To illustrate how these operations work, let us first consider a simple tree in Fig. 4 in which each of the intermediates can be made in only one way and all reactions have, say, 50% yield. For the left branch, the substrate with price "3" enters in the reaction with a fixed cost of "1". Because of the 50% yield, making 1 mmol of this reaction product requires 2 mmol of the substrate, and the total reaction cost is $1 + 3/50\% = 7$. For the reaction in the right branch, the unit cost is different ("2"; this reaction may be just harder to perform experimentally) but the cost calculation is analogous, $2 + (1 + 2)/50\% = 8$. These costs are assigned to the intermediates and propagated to the target in another 50%-yield reaction – the overall cost of making 1 mmol of the target will be $1 + (7 + 8)/50\% = 31$. The result of this recursive procedure agrees with the overall chemical balance – indeed, to make 1 mmol of the target, we used 4 mmols of each substrate (cost $4 \times 3 + 4 \times 2 + 4 \times 1 = 24$) and performed the initial two reactions (from the substrates) on twice the scale of the final reaction making the target – hence,

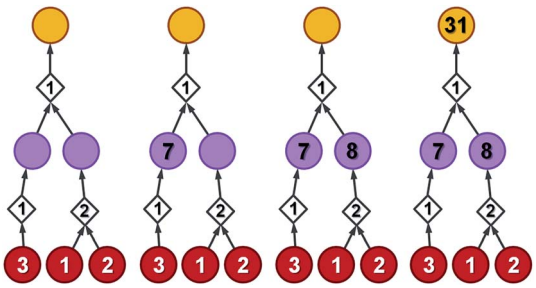

Fig. 4 A simple example illustrating bottom-up propagation of costs. All reactions have the same yield of 50%. Costs of starting materials (per mmol) are given by the numbers in the red nodes. Fixed costs of synthetic operations are indicated inside the diamond-shaped reaction nodes. For details of the calculations, see the main text.

the cost of reaction operations is $(2 \times 1 + 2 \times 2) + 1 \times 1 = 7$ and the total cost of making 1 mmol of the target is $24 + 7 = 31$. We note that such calculations can be performed rapidly for arbitrary graphs including those that contain cycles (see the small cycle involving the violet node in the bottom row of networks in Fig. 3) – the cycles, however, are chemically unproductive and the costs they entail are always higher than for acyclic pathways (compare the costs of paths $1 \rightarrow 4$ *vs.* $1 \rightarrow 4 \rightarrow 9 \rightarrow 4$ in Fig. 3d).

Coming back to Fig. 3d, we observe that in realistic networks, there is generally more than one pathway to make a given chemical – for instance, the second-from-the-left intermediate can be made in three ways, *via* reactions with costs of 5, 19, and 4. Of these, we chose the least expensive option and assign to the intermediate the cost of 4, as prescribed by the formula $\mathrm{cost}(c) = \min_{r \in \mathrm{pred}(c)}(\mathrm{cost}(r))$. Having scored all nodes within the graph, we then easily identify the most cost-effective pathway by subsequent choices (from target "down") of the lowest-scoring reactions at each synthetic generation (in our example, "11"

followed by "4"; Fig. 3e). The information about other pathways (*i.e.*, their fragments and estimated costs) is kept in a priority queue, like in an A* algorithm, and the graph is re-searched *via* a greedy-descent-type algorithm to find the second, third, *etc.* best pathways (see algorithmic details in the ESI, Section S1.4†).

Note that if one wishes to find pathways composed of minimal numbers of steps – which is a common situation in small-scale pharmaceutical synthesis whereby time is of essence and one might not even care about the prices of substrates or yields but just focus on synthesizing the target as rapidly as possible in amounts adequate for the upcoming assays – then the algorithm's parameters should be set to 100% yields, zero cost for all starting materials, and all fixed costs set to some common value. Under such assumption, the overall pathway score is simply the sum of the fixed_cost(r) over all r's (with the exception of some pathways which are not trees; see ESI, Section S1.1†). In another limiting case, when the fixed costs (labor costs) are negligible (fixed_cost(r) = 0), the cost is equal to the total cost of starting materials needed to synthesize 1 mmol of the target (taking into account the loss of mass for realistic yields <100%). The full scoring scheme we consider takes into account not just the number of steps (through the fixed_cost cost term) or costs of starting materials but also both of these factors simultaneously along with yields and most optimal placement of convergence points within a pathway.

Assigning penalties and ensuring synthetic diversity

The selection algorithm described so far can return n best-scoring pathways but does not guarantee in any way that these pathways are structurally diverse. For instance, two top-scoring solutions for the synthesis of triarylamine in Fig. 6d rely on the key Buchwald–Hartwig-type amination of the bromopyrimidine and differ only in the method of preparation of the diarylamine. In the same spirit, negligible modifications such as changing an aryl bromide to an iodide are formally different pathways to the computer but are pretty much equivalent to a user chemist. To avoid these and other unproductive repetitions and to select cost-effective yet chemically diverse pathways, we proceed as follows. After finding the best pathway (*cf.* above), the algorithm repeats the following sequence of steps until it finds the requested number of pathways or discovers that there are no more pathways left in the network:

(i) A penalty P is added to the fixed costs of each reaction from the most-recently-found pathway (Fig. 3f) and, to avoid reusing similar synthetic solutions in other pathways, also to other reactions in the network that have the same product and non-trivial (*i.e.*, having at least four carbon atoms) substrates;

(ii) A depth-first-search-like algorithm is used to identify the nodes (both reaction and molecule nodes) whose cost is affected due to the newly imposed penalization (nodes marked with question marks in Fig. 3f);

(iii) The costs of all affected nodes are recalculated by a modified Dijkstra algorithm (Fig. 3g);

(iv) Finally, a new lowest-cost pathway is identified and cycles (i)–(iv) are repeated. For all other algorithmic details, see the ESI, Sections S1.5 and S1.6.†

Algorithms' performance

One of our key motivations for developing the selection and diversity routines has been to allow queries of the solution space on timescales much shorter than those involved in the initial retrosynthetic planning creating this space. During retrosynthetic planning, Chematica has to perform multiple operations ranging from relatively rapid matching of the reaction-rule templates (such matching is common to all retrosynthesis platforms) to much slower and Chematica-peculiar assignments of proper stereo- and regiochemistry, calculations of electronic populations for some reaction types, and several more (for details, see ref. 19, 20 and 30). In effect, searches for the solutions take from minutes for medicinal-chemistry targets to hours for complex natural products, in the end presenting to the user a given number (on the order of 100) of top-scoring solutions and, at this point, discarding the remaining ones. Retaining (*e.g.*, saving on disk) the entire space of solutions allows the user to query it multiple times under different scenarios (costs of reactions, average yields, and magnitudes of imposed diversity penalties). Importantly, querying a solution graph does not involve all the slow routines of retrosynthetic planning and should thus be possible on much shorter time scales – indeed, typical times for assigning costs and selecting pathways are on the order of 1 s, even for large solution graphs and for different target molecules. Specifically, Fig. 5a shows the times t_{100} to select (on a machine with 2.5 GHz AMD Opteron 6380 processors) 100 lowest-cost pathways from solution graphs ranging in size from 90 to *ca.* 12 000 nodes – these solution graphs are for the actual synthetic examples we discuss later in the text (triarylamine, Fig. 6; Bayer's Clofedanol, Fig. 7; Amgen's AMG641 modulator of the calcium sensing receptor, Fig. 8). As seen, these t_{100} times are on the order of 0.25 s without any diversity penalties and ~0.5 s when diversity penalties P are added and costs of nodes need to be recalculated as new pathways are being selected. We note that the times to select n lowest cost pathways, t_n, scale approximately linearly with n and are also below 0.5 s for the largest solution graphs (Fig. 5b).

Illustrative synthetic examples

To illustrate how the above procedures work in practice, we considered several realistic synthetic-design examples in which the solution graphs were created by Chematica within 2–10 minutes using its standard scoring functions (see ref. 19 and 20) and comprised pathways terminating in commercially available starting materials (with prices in USD per gram, converted by the program to per mmol). We queried the solution graphs varying the average yields, the fixed costs of individual reactions on a 1 mmol scale, and the diversity penalties (henceforth denoted, respectively, Y, RxC, and P).

(i) Pathway ordering under various yield scenarios. In the first example, Chematica designed routes to an unsymmetrical triarylamine used previously in the context of photochemical synthesis of complex carbazoles in continuous flow.[31] Within *ca.* 2 min the program searched the graph of 17 881 nodes (6826 intermediates, 293 starting materials, and 10 761 reactions; Fig. 6a), from which a solution graph composed of 3176 nodes

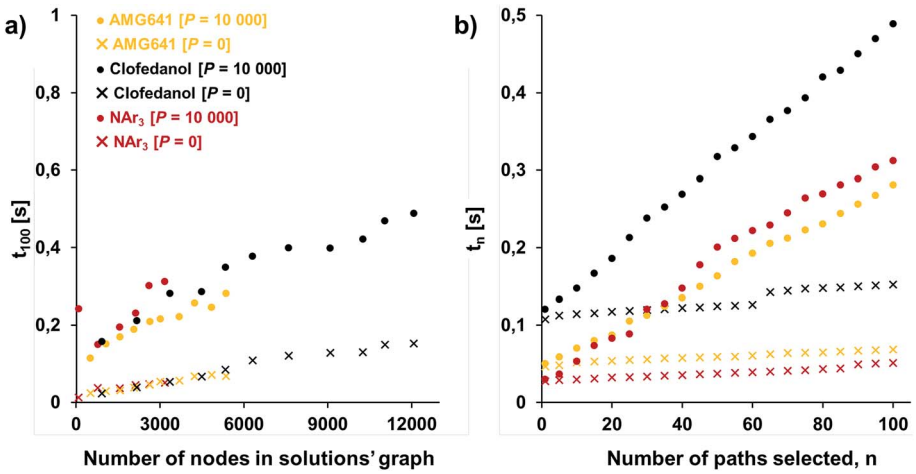

Fig. 5 Typical times to select pathways from the solution graphs. (a) Times t_{100} to select 100 lowest-cost pathways from graphs of different sizes (graph size increases as the search algorithm identifies new solutions); (b) times t_n to select n lowest-cost pathways from the maximum-size solution graphs considered here (~3000 nodes for triarylamine from Fig. 6; ~12 000 nodes for Bayer's Clofedanol from Fig. 7; and ~5400 nodes for Amgen's AMG641 modulator of the calcium sensing receptor from Fig. 8). Cross markers are for selection without diversity penalty P; solid circles are for selection with penalty $P = 10\,000$.

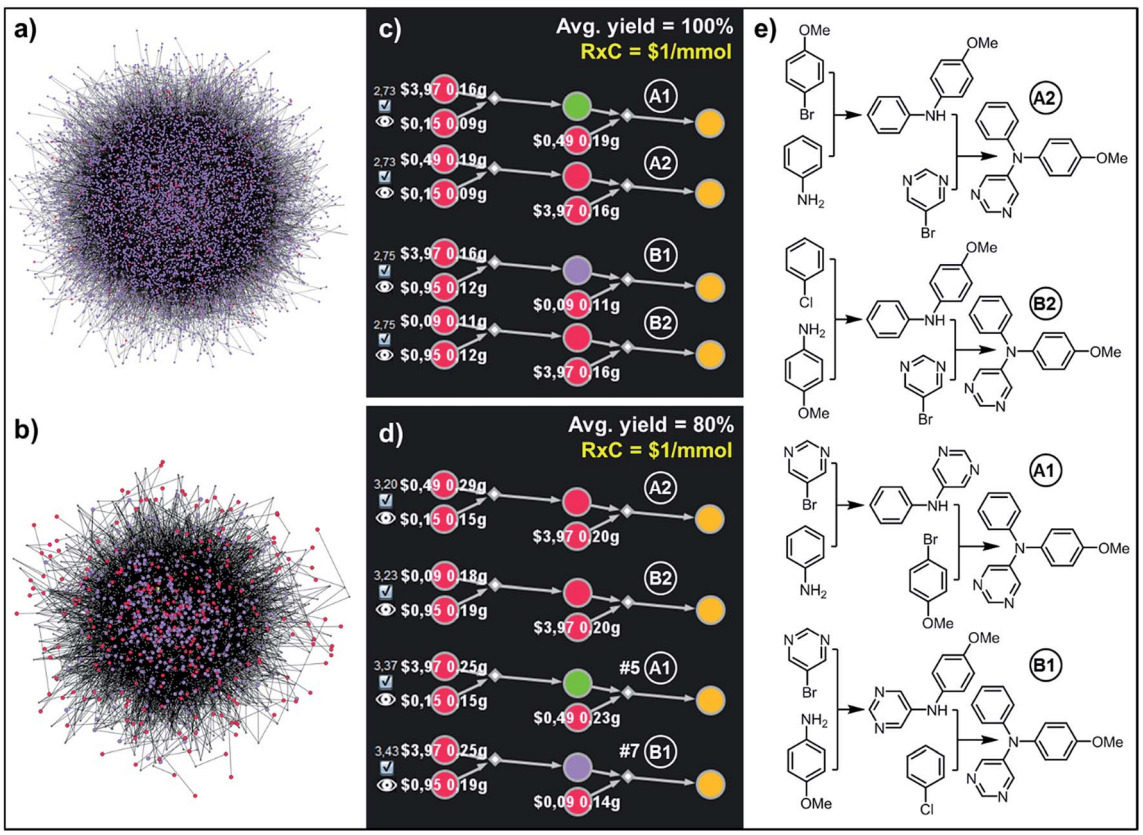

Fig. 6 Top-scoring syntheses of unsymmetrical triarylamine[31] proposed by Chematica under different yield scenarios. (a) Full graph searched during the retrosynthetic planning and (b) the solution graph. (c) Chematica's screenshots of the four top-scoring syntheses obtained with yields of all steps set to 100%. (d) Ordering of these top-scoring solutions changes when the yield is set to a more realistic 80%. (e) Chemical details of the pathways. The top scoring pathway A2 is identical to the one performed experimentally in ref. 31. For further details including reaction conditions proposed by Chematica, see the ESI, Section S4.† In (c and d), *RxC* specifies the fixed cost of performing each reaction on a 1 mmol scale ($1) while the color coding of the nodes in Chematica's pathway miniatures is as follows: yellow = target; green = intermediates whose syntheses have been already reported in the literature and are stored in the Network of Organic Chemistry, NOC;[12,14] violet = intermediates not known in the NOC; red = starting materials commercially available from Sigma-Aldrich. Pairs of white numbers over the starting materials specify the costs and amounts of these starting materials necessary to make 1 mmol of the target.

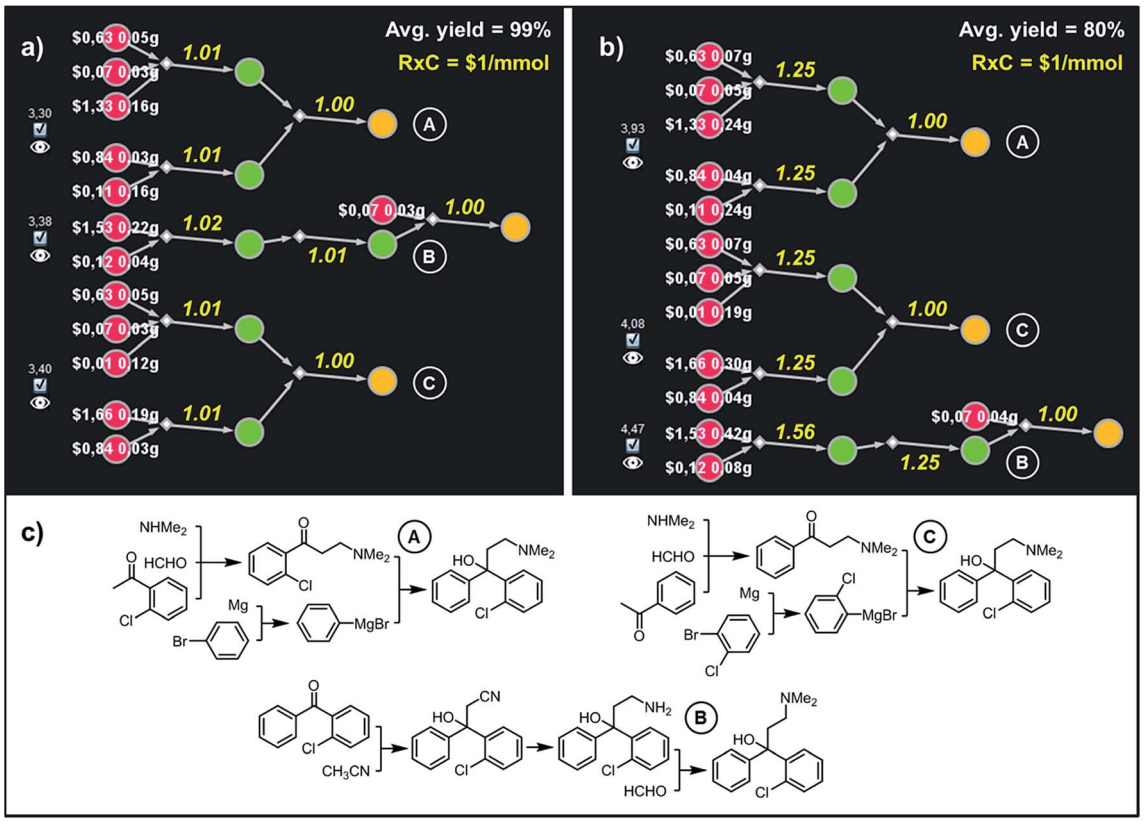

Fig. 7 Top-scoring syntheses of Clofedanol proposed by Chematica under different yield scenarios. (a) Chematica's screenshots of the three top-scoring syntheses obtained with yields of all steps set to 99%. (b) Ordering of these top-scoring solutions changes when the yield is set to a more realistic 80%. (c) Chemical details of the pathways. *RxC* specifies the fixed cost of performing each reaction on a 1 mmol scale ($1). Yellow numbers over reaction arrows are fixed costs rescaled to the scale required to ultimately make 1 mmol of the target. Colors of the nodes and all legends are explained in the caption to Fig. 6. For further details including reaction conditions suggested by Chematica, see the ESI, Section S5.†

(392 intermediates, 293 starting materials, and 2490 reactions) was selected (Fig. 6b). When this solution graph was queried with the fixed cost of each reaction operation per mmol, $RxC =$ \$1, and with an average yield of $Y = 100\%$ – that is, naively omitting mass losses at each step – the costs of the top-scoring pathways in Fig. 6c and e were simple sums of costs of performing reactions (here, \$1 per mmol step × 2 steps = \$2 per mmol) plus the costs of starting materials. While all these solutions, relying on Buchwald–Hartwig amination, were chemically correct, the algorithm was not able to capture the differences in the costs of various starting materials being used in the first *vs.* second steps. In particular, the costs of syntheses utilizing the most expensive reagent, bromopyrimidine, in the first (A1) *vs.* the second steps (A2) were exactly the same ("\$2.73," numbers to the left of the pathways in Fig. 6c), which is in contrast to the considerations from Fig. 2 showing that most expensive substrates should come in "closer to the target". This problem was avoided by specifying a more realistic average yield ($Y = 80\%$, close to the average yield of all known reactions, see ref. 28) such that the pathway costs now reflected mass loss at each step – under this condition, the top-scoring pathways A2 and B2 (Fig. 6d and e) used the expensive bromopyrimidine in the second step. We note that the top-scoring pathway A2 was actually validated experimentally in ref. 31 which inspired this example.

In the second example, more relevant to pharmaceutical chemistry, Chematica designed pathways leading to Clofedanol, a dry cough suppressant. Choosing from the solution graph created within 10 min search time and comprising 12 074 nodes in total, the cost-optimal pathways were sought with the same fixed per-mmol cost of each reaction ($RxC =$ \$1) but under two different average-yield scenarios, $Y = 99\%$ and $Y = 80\%$. Under the first scenario, the lowest-cost pathway – marked as (A) in Fig. 7 – commences with the three component Mannich reaction of 2-chloroacetophenone, followed by addition of phenylmagnesium bromide to the obtained ketone. This solution resembles the route patented in 2009 by Zhejiang Hisoar Pharma.[32] The second-scoring synthetic plan, (B), starts with the addition of acetonitrile to appropriate benzophenone, reduction of the nitrile to an amine, and reductive dimethylation with formaldehyde. This strategy is, in fact, the same as the method of preparation described in Bayer's initial (1962) patent covering Clofedanol.[33] Finally, the third-best solution, (C), also relies on the Mannich reaction of acetophenone, followed by the addition of Grignard reagent derived from *o*-bromochlorobenzene.[34] In contrast, when the average yield is $Y = 80\%$, Bayer's pathway is disfavored. In re-evaluating it, the algorithm recalculates the amounts and costs of necessary starting materials (*e.g.*, one now needs 0.42 g of benzophenone *vs.* 0.22 g under 99%-yield

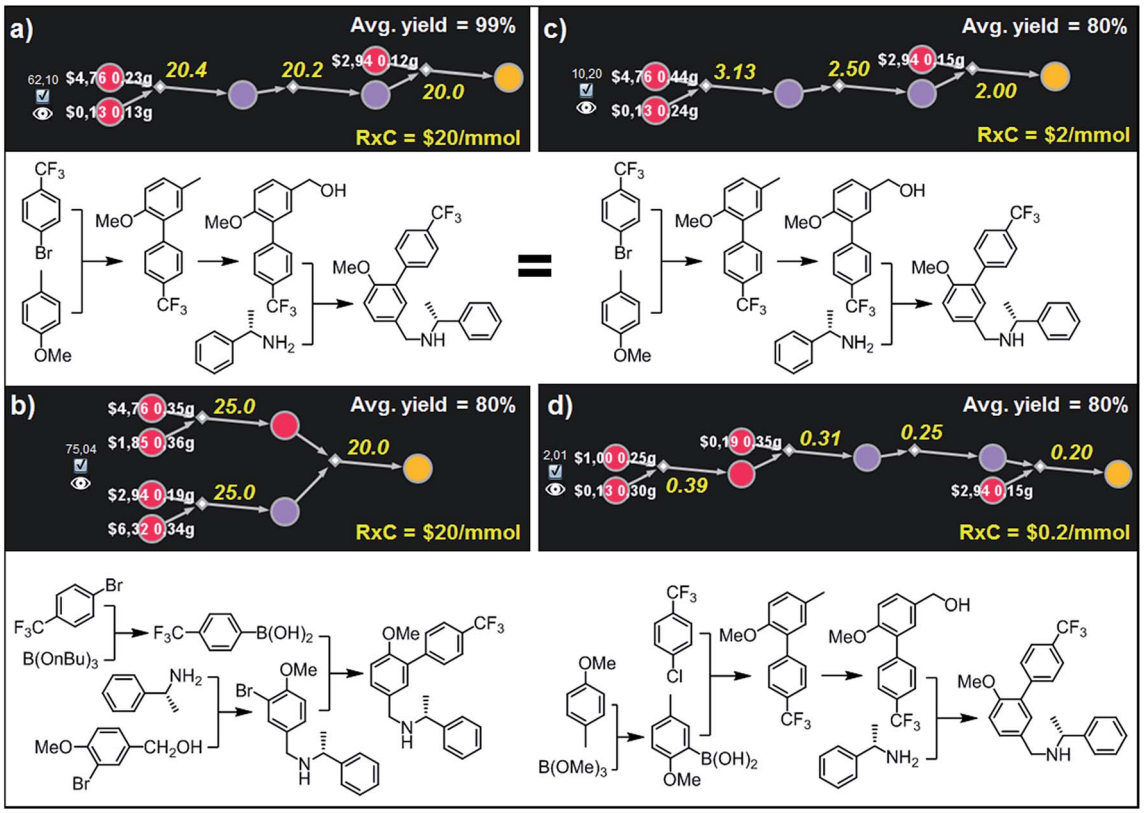

Fig. 8 Top-scoring syntheses of AMG641 proposed by Chematica under different yield and reaction-cost scenarios. Chematica's miniatures and the pathways shown below them were selected from the solution graph (created in 7 min of retrosynthetic planning; 5363 nodes in total) assuming the following values of parameters: (a) $Y = 99\%$, $RxC = \$20$; (b) $Y = 80\%$, $RxC = \$20$; (c) $Y = 80\%$, $RxC = \$2$; (d) $Y = 80\%$, $RxC = \$0.2$. Yellow numbers over reaction arrows are fixed reaction costs rescaled to the scale required to ultimately make 1 mmol of the target. Pairs of white numbers over the starting materials specify the costs and amounts of these starting materials necessary to make 1 mmol of the target. Colors of the nodes are explained in the caption to Fig. 6. For further details including reaction conditions suggested by Chematica, see the ESI, Section S6.†

assumption) and scales the costs of performing synthetic steps on larger scales (compare yellow numbers in Fig. 7a and b; *e.g.*, the addition of acetonitrile to benzophenone must now yield over 1.5 mmol of the adduct if 1 mmol of Clofedanol is expected at the end). Consequently, pathway (B) appears to be less economically feasible and is ranked lower than both approaches taking advantage of the Mannich reaction. Of course, when making such comparisons in industrial reality, it would be essential to use substrate catalogs with wholesale prices available to a specific organization, not catalog prices of Sigma-Aldrich focusing on the sales of small quantities of specialty chemicals. Fortunately, connecting a requisite catalog to Chematica or any other retrosynthetic program is a technically trivial task.

(ii) Pathway ordering under various yield and fixed-reaction-cost scenarios. The example in this section is intended to illustrate how the optimal pathways vary when both the average reaction yields and the fixed costs of performing individual reactions on a given scale change. Specifically, we query the graph of synthetic solutions leading to Amgen's AMG641 (ref. 35) – an orally efficacious, positive allosteric modulator of the calcium sensing receptor – varying Y from 99% down to 80% and RxC from \$20 (expensive, probably small-scale synthesis) to \$0.2 (relatively inexpensive, probably larger scale production). For $Y = 99\%$ and $RxC = \$20$, the best-scoring solution in Fig. 8a is a three step linear sequence initialized by an elegant one-pot *ortho*-lithiation/Pd-mediated coupling[36] with 4-trifluoromethylbromobenzene; subsequent oxidation of the benzylic position[37] and alkylation of commercially available chiral amines yield the target molecule. When reaction costs remain the same but one accounts for the mass loss at each step ($Y = 80\%$; Fig. 8b), the best solution is a convergent three-step sequence, mirroring the original Amgen's route and using benzylic alcohol to alkylate the amine in one-pot oxidation-reductive amination[38] and boronic acid (prepared in one step from appropriate bromoarene) to construct the biphenyl part of AMG641.

Although both of these routes are chemically correct, they might not be optimal if AMG641 goes into large-scale production characterized by lower reaction-operation costs (*e.g.*, achieved by solvent recycling, use of crystallization rather than chromatography, *etc.*). To emulate such hypothetical scale-up, we kept $Y = 80\%$ but decreased RxC to \$2 and then to \$0.2. In the first case, the best-scoring solution (Fig. 8c) is actually the same as in Fig. 8a for $Y = 99\%$. We note, however, that the overall cost of this plan recalculated with $Y = 80\%$, $RxC = \$2$ constrains is, as expected, very different (\$10.2 per mmol *vs.*

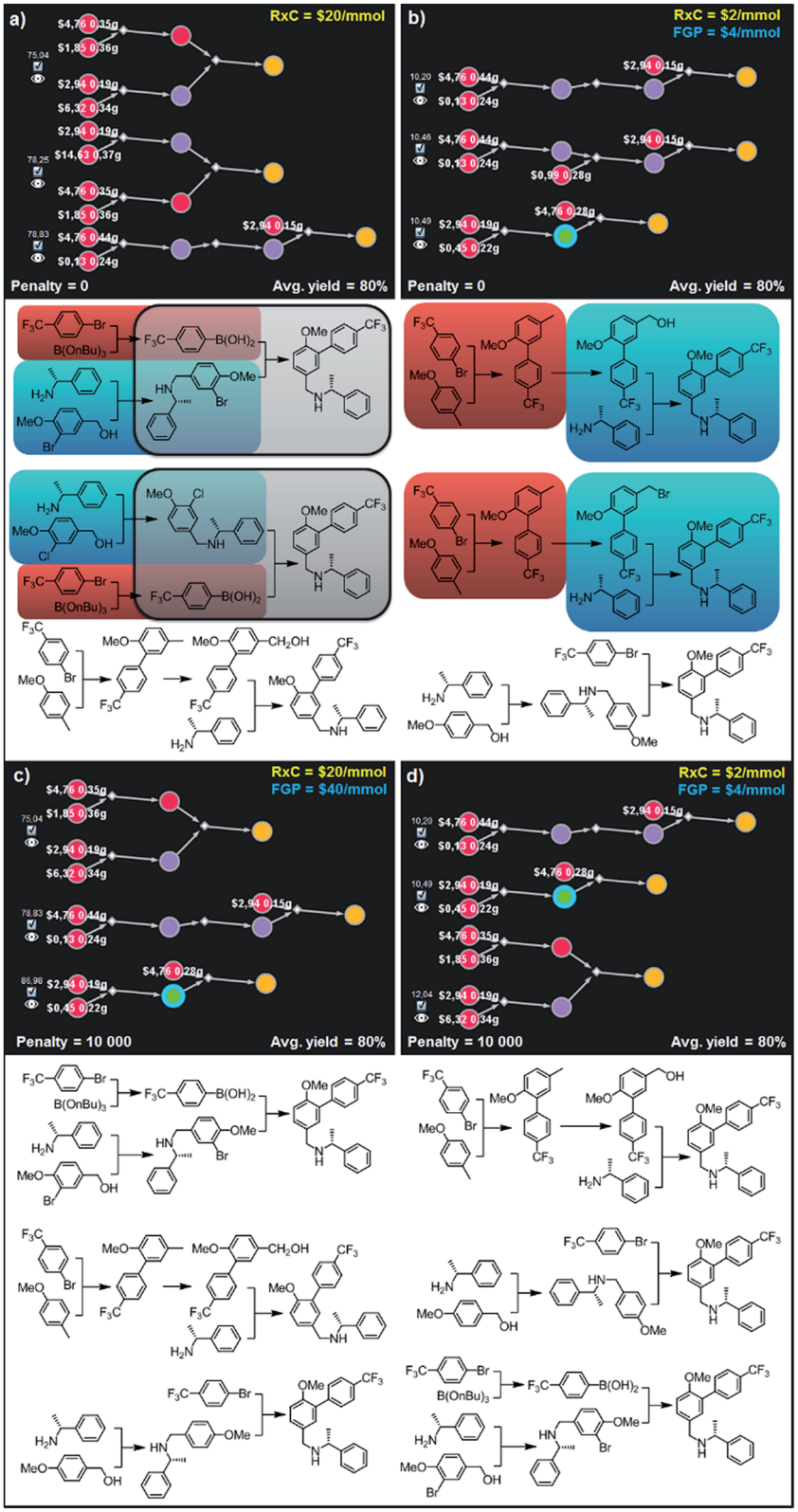

Fig. 9 Top-scoring syntheses of AMG641 proposed by Chematica without and with the application of diversity penalties. Chematica's sets of solutions shown in (a and b) have several identical (red) or very similar (blue and grey) steps when the pathways are selected from the solution graph with no penalty for reuse of already used reactions. When such a penalty is added (c and d) diversity of the synthetic plans returned is improved and each transformation is used only once. FGP is the cost of reactions requiring protection (of substances whose nodes are surrounded by blue halos). In Chematica, the cost of protection is set by the user, typically at twice the cost of reactions not requiring protection (here, FGP = 2 *RxC*). For further details including reaction conditions suggested by Chematica, see the ESI, Section S7.†

$62.1 per mmol in Fig. 8a), reflecting slightly higher quantities and costs of starting materials ($2.56 *vs.* $1.38) but much lower costs of reaction operations (7.63$ *vs.* 60.6$). Finally, further decrease of *RxC* to $0.2 adds an extra step (labor/operations are now cheap!) but sources the synthesis from very inexpensive starting materials (4-methylanisole and chloroarene). This four step linear sequence is shown in Fig. 8d and begins with the Suzuki coupling of aryl chloride and boronic acid prepared *via ortho*-lithiation and trapping of the obtained aryllithium with trimethyl borate. Subsequent oxidation of the benzylic position and junction with an appropriate amine leads to the target molecule. Taken together, the examples we discussed in this section illustrate that by varying the *Y* and *RxC* parameters, the machine makes pathway selections that reflect the economical differences between medicinal chemistry and manufacturing operations.

(iii) Selection of diverse pathways. The pathways we described in previous examples were all chemically viable and the selection algorithm adapted to different scoring/pricing scenarios, but within each scenario, the variability among the *n* best-scoring pathways was far from satisfactory – in other words, the *n* top-scoring pathways selected for given values of *Y* and *RxC* could rely on the same or chemically equivalent transformations. This limits the menu of solutions the user is presented with. To illustrate the problem and how to remedy it by imposing penalties *P* on the reuse of equivalent transforms (see Fig. 3 and accompanying algorithm described earlier in the text), we required that Chematica returns three top scoring syntheses of the AMG641 target under the two *Y*–*RxC* scenarios from Fig. 8b,c but with *vs.* without diversity penalization. With no penalization ($P = 0$) the selection algorithm proposed sets of synthetic plans in which the same or similar transformations were used several times. For example, the first and second solutions shown in Fig. 9a ($Y = 80\%$, *RxC* = $20 per mmol) rely on the Suzuki coupling of *p*-trifluoromethylbenzeneboronic acid with either bromo- or chloroarene (*grey*). Additionally, the necessary haloarene is prepared *via* alkylation of the same benzylamine with either appropriate bromobenzyl bromide or chlorobenzyl alcohol (*blue*) while the preparation of the boronic acid in both plans starts from the same bromobenzene (*red*) undergoing Br/Li exchange and trapping with tributyl borate. Similar redundancy was observed in results obtained under different *Y*–*RxC* scenarios ($Y = 80\%$, *RxC* = $2 per mmol) and is illustrated in Fig. 9b. Here, the only minor difference between the top and the second-best solutions is, in fact, the leaving group of the benzylating agent used in the *N*-alkylation. In both

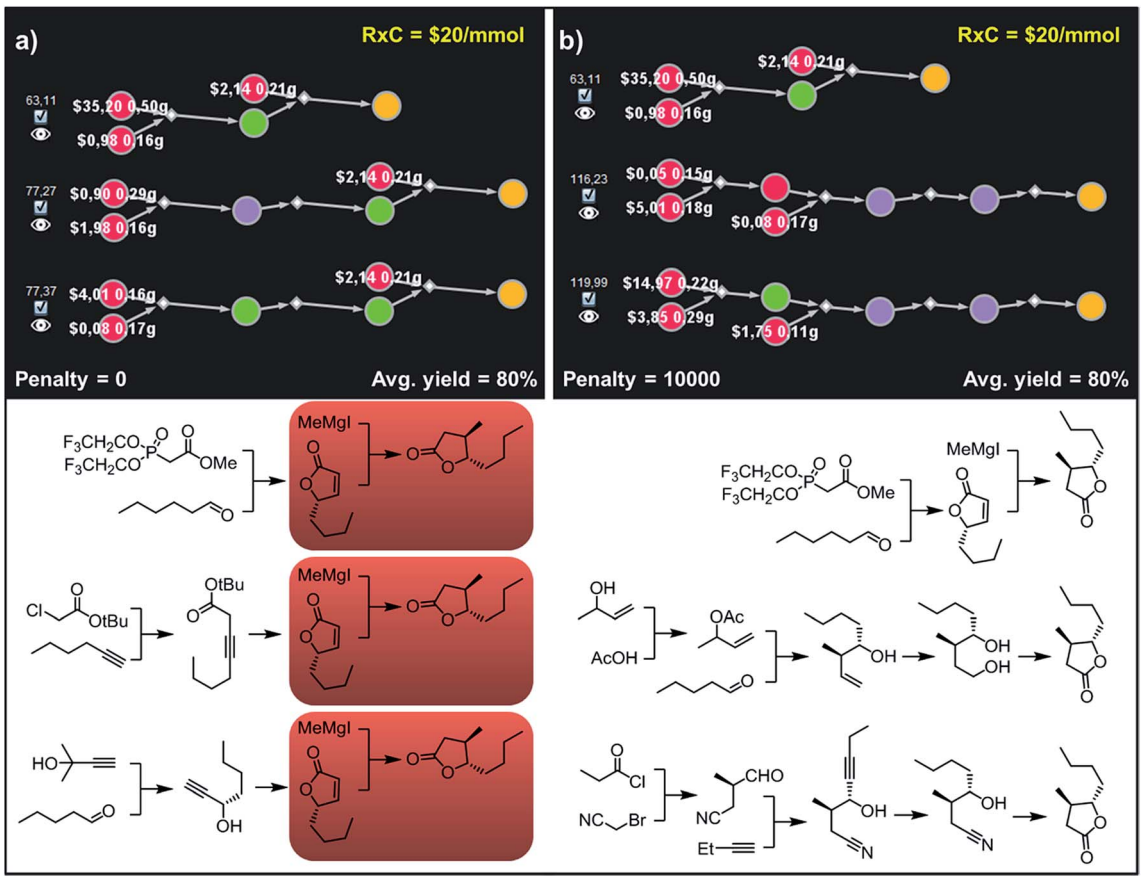

Fig. 10 Top-scoring syntheses of trans whisky lactone proposed by Chematica without and with the application of diversity penalties. (a) Without any diversity penalties, all pathways use the same method to install the C3 stereocenter (red frames). (b) When $P = 10\,000$ penalties are imposed, all three top-scoring plans are substantially different and rely on different methodologies. For further details including reaction conditions suggested by Chematica, see the ESI, Section S8.†

pathways, the last step (*blue*) requires the same amine undergoing alkylation while the construction of the biphenyl part of AMG641 takes advantage of the identical lithiation–arylation (*red*). In sharp contrast, results obtained after applying large diversity penalty ($P = 10\,000$) are chemically diverse. In particular, sets of synthetic plans shown in Fig. 9c and d rely on the (i) alkylation of the amine with the *m*-bromobenzyl alcohol and subsequent Suzuki coupling, (ii) *ortho*-lithiation/arylation, followed by hydroxylation and *N*-alkylation, or (iii) alkylation of the amine with *p*-methoxybenzyl alcohol and late-stage lithiation/arylation. All of the transformations used in these sets of plans are unique and used only once in the entire series of solutions – though, we observe, these relatively simple syntheses still bear some "thematic" similarity.

Accordingly, to allow for more synthetic latitude and diversity, our final example deals with more complex enantioselective syntheses of *trans*-whisky lactone (3-methyl-4-octanolide) isolated from oak wood and responsible for the taste of aged spirits.[39] With no penalization applied ($P = 0$) each of the three top-scoring synthetic plans relies on the formation of butenolides and subsequent trans-selective 1,4-addition of organocuprate (derived from methylmagnesium iodide; Fig. 10a, red frames) to set the C3 stereocenter, mimicking previous literature approaches.[40,41] The necessary enantioenriched butenolide can be obtained from hexanal *via* proline-mediated aminoxylation-olefination[42] (Fig. 10a, top path). We note that this approach was demonstrated experimentally during preparation of the structurally similar *trans*-cognac lactone.[42] Alternatively, the butenolide can be prepared *via* enantioselective isomerization-cyclisation[43] of β,γ-alkynoic ester which is available in one step from hexyne and chloroacetate (Fig. 10a, middle), or *via* enantioselective addition[44] of protected acetylene to pentanal, followed by carbonylative cyclisation[45,46] (Fig. 10a, bottom). In contrast, after applying diversity penalty ($P = 10\,000$), the alternative pathways no longer hinge on the 1,4-addition and both contiguous stereocenters are forged prior to the formation of the lactone. In particular, the second-best solution (Fig. 10b, middle path) now takes advantage of the Krische's crotylation[47] of pentanal setting both stereocenters. Hydroboration of the homoallylic alcohol thus obtained yields a 1,4-diol undergoing oxidative cyclisation[48] to the target molecule. Finally, the third-plan (Fig. 10b, bottom) commences with a chiral-auxiliary-controlled cyanomethylation of the enolate with bromoacetonitrile.[49] Subsequent addition of butynal controlled by a chiral catalyst[50,51] yields hydroxynitrile, which then undergoes reduction of alkyne and intramolecular alcoholysis to the whisky lactone target.

Conclusions

In summary, we described a family of algorithms that select and score the most economical and diverse synthetic pathways from large graphs of synthetically viable solutions. This problem has not been addressed in detail in previous literature on computer-assisted retrosynthesis likely because – until now – few solutions were produced during retrosynthetic searches and *any* chemically viable outcome has been deemed a success. Now, with much improved algorithms and modern computing power, the situation has changed and one faces the *embarras de richesse* problem, with very large numbers of potential solutions, all chemically plausible. With the algorithms like the ones we described, one can save the entire solution space and then query it rapidly, within seconds, for pathways meeting desired cost scenarios (instead of re-running the slow retrosynthetic search with different parameters). As we mentioned in the text, to truly reflect the realistic costs of specific organizations, the algorithm should be interfaced with catalogs of starting materials with prices peculiar to these organizations. In the future, one could also think of augmenting the penalization schemes – here, used to ensure chemical diversity – to downplay the use of reagents that are undesirable (toxic, volatile, *etc.*) or reaction types known to be particularly difficult or finicky.

Author contributions

T. B. designed and implemented the selection algorithms. K. M. validated the algorithms for synthetic correctness and provided examples of syntheses described in the text. B. A. G. conceived Chematica in graduate school and has directed the development of its various aspects – including the current work – ever since. All authors contributed to the writing of the manuscript.

Conflicts of interest

While 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 a property of Merck KGaA, Darmstadt, Germany. The authors continue to collaborate with Merck KGaA, Darmstadt, within the DARPA "Make-It" award. All queries about access options to Chematica (now rebranded as Synthia™), including academic collaborations, should be directed to Dr Sarah Trice at sarah.trice@sial.com.

Acknowledgements

This work was supported by the U.S. DARPA under the "Make-It" Award, 69461-CH-DRP #W911NF1610384. B. A. G. also gratefully acknowledges personal support from the National Science Center, NCN, Poland (Symfonia Award #2014/12/W/ST5/00592) and from the Institute for Basic Science Korea, Project Code IBS-R020-D1. We would like to thank Dr Piotr Dittwald for generating images of reaction networks.

References

1 E. J. Corey and W. T. Wipke, *Science*, 1969, **166**, 178–192.
2 E. J. Corey, W. T. Wipke, R. D. Cramer and W. J. Howe, *J. Am. Chem. Soc.*, 1972, **94**, 421–430.
3 E. J. Corey, W. T. Wipke, R. D. Cramer and W. J. Howe, *J. Am. Chem. Soc.*, 1972, **94**, 431–439.
4 H. L. Gelernter, A. F. Sanders, D. L. Larsen, K. K. Agarwal, R. H. Boivie, G. A. Spritzer and J. E. Searleman, *Science*, 1977, **197**, 1041–1049.

5 S. Hanessian, J. Franco and B. Larouche, *Pure Appl. Chem.*, 1990, **62**, 1887–1910.
6 J. B. Hendrickson, *J. Am. Chem. Soc.*, 1977, **99**, 5439–5450.
7 J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade and H. Y. Ando, *J. Chem. Inf. Model.*, 2009, **49**, 593–602.
8 A. Bøgevig, H.-J. Federsel, F. Huerta, M. G. Hutchings, H. Kraut, T. Langer, P. Löw, C. Oppawsky, T. Rein and H. Saller, *Org. Process Res. Dev.*, 2015, **19**, 357–368.
9 M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604–610.
10 C. W. Coley, L. Rogers, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 1237–1245.
11 C. W. Coley, W. H. Green and K. F. Jensen, *Acc. Chem. Res.*, 2018, **51**, 1281–1289.
12 M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2005, **44**, 7263–7269.
13 K. J. M. Bishop, R. Klajn and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2006, **45**, 5348–5354.
14 B. A. Grzybowski, K. J. M. Bishop, B. Kowalczyk and C. E. Wilmer, *Nat. Chem.*, 2009, **1**, 31–36.
15 C. M. Gothard, S. Soh, N. A. Gothard, B. Kowalczyk, Y. Wei, B. Baytekin and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2012, **51**, 7922–7927.
16 M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Weckiewicz, P. E. Fuller, B. A. Grzybowski and K. J. M. Bishop, *Angew. Chem., Int. Ed.*, 2012, **51**, 7928–7932.
17 P. E. Fuller, C. M. Gothard, N. A. Gothard, A. Weckiewicz and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2012, **51**, 7933–7937.
18 C. Chaouiya, *Briefings Bioinf.*, 2007, **8**, 210–219.
19 S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2016, **55**, 5904–5937.
20 T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. J. Trice and B. A. Grzybowski, *Chem*, 2018, **4**, 522–532.
21 L. R. Nielsen, K. A. Andersen and D. Pretolani, *Comput. Oper. Res.*, 2005, **32**, 1477–1497.
22 E. Miller-Hooks, *Networks*, 2001, **37**, 35–52.
23 D. Pretolani, *Eur. J. Oper. Res.*, 2000, **123**, 315–324.
24 G. Gallo, G. Longo, S. Pallottino and S. Nguyen, *Discrete Appl. Math.*, 1993, **42**, 177–201.
25 S. Nguyen and S. Pallottino, *Eur. J. Oper. Res.*, 1988, **37**, 176–186.
26 R. Fagerberg, C. Flamm, R. Kianian, D. Merkle and P. F. Stadler, *J. Cheminf.*, 2018, **10**, 19.
27 V. Akgün, E. Erkut and R. Batta, *Eur. J. Oper. Res.*, 2000, **121**, 232–246.
28 G. Skoraczyński, P. Dittwald, B. Miasojedow, S. Szymkuć, E. P. Gajewska, B. A. Grzybowski and A. Gambin, *Sci. Rep.*, 2017, **7**, 3582.
29 F. S. Emami, A. Vahid, E. K. Wylie, S. Szymkuć, P. Dittwald, K. Molga and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2015, **54**, 10797–10801.
30 W. Beker, E. P. Gajewska, T. Badowski and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2019, **58**, 4515–4519.
31 A. C. Hernandez-Perez, A. Caron and S. K. Collins, *Chem.–Eur. J.*, 2015, **21**, 16673–16678.
32 T. He and Y. Wenqiu, CN101844989, 2009.
33 R. Lorenz, R. Gosswald and H. Henecka, US3031377A, 1962.
34 R. S. Sulake, C. Chen, H.-R. Lin and A.-C. Lua, *Bioorg. Med. Chem. Lett.*, 2011, **21**, 5719–5721.
35 P. E. Harrington, D. J. St. Jean, J. Clarine, T. S. Coulter, M. Croghan, A. Davenport, J. Davis, C. Ghiron, J. Hutchinson, M. G. Kelly, F. Lott, J. Y.-L. Lu, D. Martin, S. Morony, S. F. Poon, E. Portero-Larragueta, J. D. Reagan, K. A. Regal, A. Tasker, M. Wang, Y. Yang, G. Yao, Q. Zeng, C. Henley and C. Fotsch, *Bioorg. Med. Chem. Lett.*, 2010, **20**, 5544–5547.
36 M. Giannerini, V. Hornillos, C. Vila, M. Fañanás-Mastral and B. L. Feringa, *Angew. Chem., Int. Ed.*, 2013, **52**, 13329–13333.
37 A. D. Cort, L. Mandolini and S. Panaioli, *Synth. Commun.*, 1988, **18**, 613–616.
38 C. Guérin, V. Bellosta, G. Guillamot and J. Cossy, *Org. Lett.*, 2011, **13**, 3534–3537.
39 K. Otsuka, Y. Zenibayashi, M. Itoh and A. Totsuka, *Agric. Biol. Chem.*, 1974, **38**, 485–490.
40 P. Koschker, M. Kähny and B. Breit, *J. Am. Chem. Soc.*, 2015, **137**, 3131–3137.
41 B. Mao, K. Geurts, M. Fañanás-Mastral, A. W. van Zijl, S. P. Fletcher, A. J. Minnaard and B. L. Feringa, *Org. Lett.*, 2011, **13**, 948–951.
42 D. A. Devalankar, P. V. Chouthaiwale and A. Sudalai, *Tetrahedron: Asymmetry*, 2012, **23**, 240–244.
43 H. Liu, D. Leow, K.-W. Huang and C.-H. Tan, *J. Am. Chem. Soc.*, 2009, **131**, 7212–7213.
44 D. Boyall, F. López, H. Sasaki, D. Frantz and E. M. Carreira, *Org. Lett.*, 2000, **2**, 4233–4236.
45 W.-Y. Yu and H. Alper, *J. Org. Chem.*, 1997, **62**, 5684–5687.
46 W. P. Gallagher and R. E. Maleczka, *J. Org. Chem.*, 2003, **68**, 6775–6779.
47 X. Gao, I. A. Townsend and M. J. Krische, *J. Org. Chem.*, 2011, **76**, 2350–2354.
48 M. Ito, A. Osaku, A. Shiibashi and T. Ikariya, *Org. Lett.*, 2007, **9**, 1821–1824.
49 M. T. Crimmins, M. Shamszad and A. E. Mattson, *Org. Lett.*, 2010, **12**, 2614–2617.
50 J. A. Marshall and M. P. Bourbeau, *Org. Lett.*, 2003, **5**, 3197–3199.
51 R. Takita, K. Yakura, T. Ohshima and M. Shibasaki, *J. Am. Chem. Soc.*, 2005, **127**, 13760–13761.