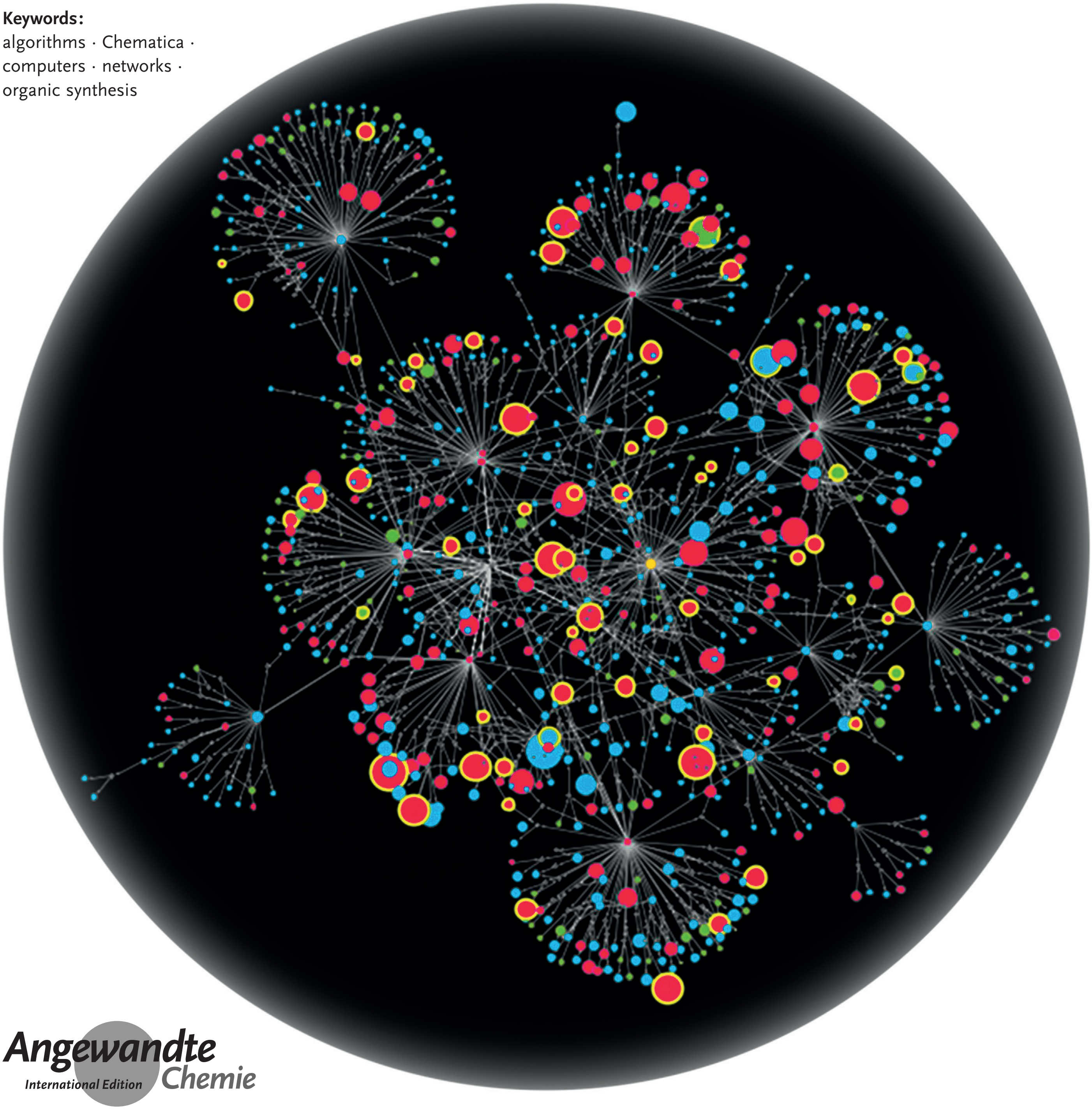
Computer-Aided Synthetic Planning

International Edition: DOI: 10.1002/anie.201506101
German Edition: DOI: 10.1002/ange.201506101

Computer-Assisted Synthetic Planning: The End of the Beginning

*Sara Szymkuć, Ewa P. Gajewska, Tomasz Klucznik, Karol Molga, Piotr Dittwald, Michał Startek, Michał Bajczyk, and Bartosz A. Grzybowski**

Keywords:
algorithms · Chematica · computers · networks · organic synthesis



© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

Exactly half a century has passed since the launch of the first documented research project (1965 Dendral) on computer-assisted organic synthesis. Many more programs were created in the 1970s and 1980s but the enthusiasm of these pioneering days had largely dissipated by the 2000s, and the challenge of teaching the computer how to plan organic syntheses earned itself the reputation of a "mission impossible". This is quite curious given that, in the meantime, computers have "learned" many other skills that had been considered exclusive domains of human intellect and creativity—for example, machines can nowadays play chess better than human world champions and they can compose classical music pleasant to the human ear. Although there have been no similar feats in organic synthesis, this Review argues that to concede defeat would be premature. Indeed, bringing together the combination of modern computational power and algorithms from graph/network theory, chemical rules (with full stereo- and regiochemistry) coded in appropriate formats, and the elements of quantum mechanics, the machine can finally be "taught" how to plan syntheses of non-trivial organic molecules in a matter of seconds to minutes. The Review begins with an overview of some basic theoretical concepts essential for the big-data analysis of chemical syntheses. It progresses to the problem of optimizing pathways involving known reactions. It culminates with discussion of algorithms that allow for a completely de novo and fully automated design of syntheses leading to relatively complex targets, including those that have not been made before. Of course, there are still things to be improved, but computers are finally becoming relevant and helpful to the practice of organic-synthetic planning. Paraphrasing Churchill's famous words after the Allies' first major victory over the Axis forces in Africa, it is not the end, it is not even the beginning of the end, but it is the end of the beginning for the computer-assisted synthesis planning. The machine is here to stay.

From the Contents

1. Introduction

Only few areas of chemical research have experienced such a dramatic reversal of fortunes—from early excitement to current pessimism—as computer-assisted synthesis planning. This state of affairs is somewhat surprising given that it was the organic chemists who were among the first to recognize, already in the 1960s, the promise of modern computing in natural sciences. Unfortunately, the problem these pioneers tackled turned out to be too complex for the machines and algorithms of their day and the approaches they developed could be applied only to relatively simple targets for which human experts did not really need machine's assistance. Also, as eloquently narrated by Prof. P. Judson in his book on experts systems in chemistry,[1] the programs were cumbersome to use, and so they did not gain widespread acceptance in the community and slowly, one by one, faded from the scene of history. These early synthetic-chemical programs were neither right nor wrong—they were simply somewhat irrelevant to the everyday practice of organic synthesis. Whatever the reasons might have been, computer-assisted synthetic planning seems to have missed the great computer revolution of the 1990s and 2000s: While Deep Blue

[*] S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, Dr. P. Dittwald, M. Bajczyk, Prof. Dr. B. A. Grzybowski
Institute of Organic Chemistry, Polish Academy of Sciences
Kasprzaka 44/52, Warsaw 02-224 (Poland)
E-mail: nanogrzybowski@gmail.com

Prof. Dr. B. A. Grzybowski
Center for Soft and Living Matter of Korea's Institute for Basic Science (IBS) and Department of Chemistry, Ulsan National Institute of Science and Technology
50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan (South Korea)
E-mail: grzybor72@unist.ac.kr

Dr. M. Startek
Faculty of Mathematics, Informatics, and Mechanics
University of Warsaw
Banacha 2, 02-097 Warszawa (Poland)

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201506101.

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

was beating Garry Kasparov in chess, and when computers were revolutionizing fields ranging from economics to biology, only few chemical groups maintained active research on computational synthetic planning. One notable exception where vibrant and fruitful development continued has been chemical databases and automated management of synthetic knowledge (e.g., electronic notebooks). Nowadays, repositories such as Reaxys,[2a] SciFinder,[2b] ChemSpider,[2c] and SPRESI[2d] offer invaluable help to chemists searching for literature sources and/or examples of "analogous" reactions. Even in this area, however, the power of modern computers is heavily underused, and the searches are largely manual, step-by-step procedures with only rudimentary capabilities to evaluate reaction sequences.

The main thesis of this Review is that these limitations can finally be overcome and modern computers 1) can enable dramatically improved searches of chemical databases that span the entire chemical space of known molecules; and 2) can be taught to design optimized synthetic pathways towards previously unexplored targets including relatively complex ones. As we will see, the common denominator of these two types of problems is that both reaction databases and the options generated by computers during retrosynthetic planning can be described as networks. This mathematical insight is quite critical since it enables the algorithms commonly used in the telecom industry or in chess-playing programs to be adapted and extended to search and analyze the gigantic networks of synthetic possibilities.

The Review is divided into three parts. In the first (Section 2), we focus on the applications of network algorithms to search the universe of known (i.e., literature reported) reactions. We discuss the key differences in searching databases versus searching networks, and the advantages of using the so-called bipartite representation of the Network of Organic Chemistry (NOC).[3a,b] We then introduce the concepts of various scoring functions which—with appropriately crafted search algorithms—can scrutinize and evaluate hundreds of millions of synthetic possibilities per second and can identify synthetic routes that are optimized for various user-specified criteria (e.g., the overall monetary cost, popularity of substrates involved, avoidance of toxic/hazardous intermediates, etc.)

While searches of all known chemical reactions can be computationally demanding, the underlying algorithms are well developed in other fields of science, and hence can be extended to the problem of chemical synthesis in a relatively straightforward manner. But for the completely de novo synthetic planning which we discuss in the second part (Section 3), new method development has been required. We begin by narrating the historical milestones (Section 3.1). Although only few of the early synthetic-planning programs are still available today, their importance for the development of the field should not be underestimated, as they galvanized the creation of various types of machine-readable molecular notations (e.g., SMILES)[4] and also structure editors (e.g., Stewart Rubenstein's ChemDraw emerging from Corey's LHASA effort[5,6]). Programs such as P. Y. Johnson's SYNLMA[7] contributed by introducing the elements of formal logic into synthetic planning, Gelernter's SYNCHEM[8] was a pioneering effort of searching expanding trees of synthetic possibilities, and Hanessian's CHIRON[9] emphasized the recognition of similar structural patterns in the targets and in the substrates. These are all essential aspects of synthetic planning—the question remains, though,

Sara A. Szymkuć graduated summa cum laude from the Faculty of Chemistry of the Warsaw University of Technology with thesis on multicomponent reactions applied to peptide mimics. She is currently a graduate student at the Institute of Organic Chemistry of the Polish Academy of Sciences in Warsaw. Her scientific interests focus on computer assisted organic chemistry, chemical networks, and multicomponent reactions.

Tomasz Klucznik graduated in chemistry and biotechnology from the Gdańsk University of Technology. He is currently a graduate student at the Institute of Organic Chemistry of the Polish Academy of Sciences in Warsaw. He received several best-presentation prizes at various chemical conferences and has been actively involved in the popularization of chemistry among high-school and college students. He is interested in chemical networks, theory of organic chemistry, total synthesis, methodology of organic synthesis, chemistry of sulfur, and philosophy.

Ewa P. Gajewska graduated with distinction in chemistry and biotechnology from the Gdańsk University of Technology. In 2013, she was a recipient of the Outstanding Achievements Award from Poland's Ministry of Science and Higher Education. She is currently a graduate student at the Institute of Organic Chemistry of the Polish Academy of Sciences in Warsaw. Her scientific interests include chemical networks and the fundamentals of organic reaction mechanisms.

Karol Molga graduated summa cum laude from the Faculty of Chemistry of the Warsaw University of Technology, completing his thesis on organonickel chemistry. He is currently a graduate student at the Institute of Organic Chemistry of the Polish Academy of Sciences in Warsaw where he works on chemical networks and the logic of computer-assisted retrosynthetic analysis.

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

why all these innovations have not given rise to widely accepted computational methods/tools such as those used in quantum chemistry or molecular dynamics?

Reflecting on this question, we will revisit (in Section 3.2) the fundamentals of the problem, and we will discuss the intellectual connection of synthetic planning to other fields, among them chess, the Rubik's cube (Table 1 and Refs. [10–15]), and even linguistics. The three distinct features of synthetic planning that emerge from such comparisons are that 1) the number of rules/"moves" is much larger than in other games (tens of thousands of reaction types versus order-of-ten for chess and few for Rubik's cube); 2) the applicability of any given "synthetic move" depends on context, i.e., on the presence of other chemical groups present in the same molecule, just as the meaning of a word can depend on the context of the entire sentence; and 3) there is no well-defined criterion of a current "synthetic position" which could be systematically evaluated to plan future "moves", which contrasts with the situation for the game of chess. Several important consequences follow. First, the rules the computer needs to be taught have to be not only "locally" correct ("cut this specific bond") but also context-sensitive ("examine other groups in the molecule")—it is largely this context dependence that has rendered previous approaches (either based on bond disconnections or using "analogous" reactions extracted from literature precedents) unsuccessful. In short, the conditional/context rules of chemistry must be coded by human experts, and a large number of them must be coded (>10000) before the machine can begin to compete with a knowledgeable human. Second, the concept of position must and can be algorithmically defined by considering the structural and chemical complexity of the sets of substrates generated in each reaction "move". This, in turn, means that hypothetical pathways must be scored both for the reaction steps performed and for the substrates' complexity. This type of dual scoring is a drastic departure from the examination of disconnections alone and leads us to introduce the concepts of reaction and chemicals' scoring functions. Third, there must be algorithms that use the knowledge-base and the scoring functions to navigate the synthetic space intelligently, not only going "forward" but also reverting from hopeless positions to eliminate what Corey called the "combinatorial explosion" of possibilities. In Sections 3.3 and 3.4 we will discuss approaches that finally meet these requirements while also taking into account full stereochemistry, regiochemistry, protecting group information, and even some rudimentary quantum mechanics.

Although the examples of actual computer-designed syntheses we will see in the text are already non-trivial and reflect the computer's capability to strategize with payoff almost as good as a highly-trained human chemist, there persist multiple challenges and also exciting opportunities for future research—we narrate both in the third part (Section 4). For instance, there have been interesting developments in predicting outcomes of stereoselective reactions, reaction yields, and even reaction conditions. New measures of synthetic complexity are also emerging that can allow for rapid assessment of "synthesizability" in large libraries[16] of "virtual molecules," which are nowadays routinely created in pharmaceutical and materials science industries. Last but not least, there are new approaches to predicting new reaction types/mechanisms ranging from quantum-mechanical calculations, to graph-theoretical and machine-learning methods.

Overall, we believe that modern computers can finally provide valuable help to practicing organic chemists. While the machines are not yet likely to match the creativity of top-

Piotr Dittwald studied mathematics and computer science at the University of Warsaw where he also obtained his Ph.D. in computer science. His doctoral research covered investigation of recurrent rearrangements in the human genome and application of computational methods in mass spectrometry. He twice received the START Fellowship from the Foundation for Polish Science. Currently, he is a post-doctoral fellow at the Institute of Organic Chemistry of the Polish Academy of Sciences in Warsaw, where he is developing algorithms for computer-assisted chemical synthesis.

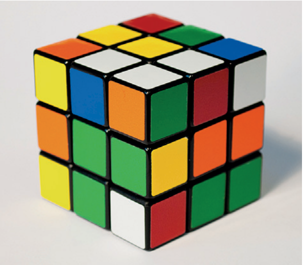
Michał D. Bajczyk obtained M.Sc. degrees in chemistry and in biochemistry from the Jagiellonian University in Cracow. He was twice awarded the TEAM scholarship from the Foundation for Polish Science in 2012 and 2013. He is currently a graduate student at the Institute of Organic Chemistry of the Polish Academy of Sciences in Warsaw with interest in de novo design of multicomponent reactions and physical biochemistry.

Michał P. Startek studied mathematics and computer science at the University of Warsaw, and received his Ph.D. there, with highest distinction, in 2015. He is currently a postdoctoral fellow at the same institution. His current research interests include evolutionary models of the behavior of transposable elements, methods of computational analysis of chemical big-data, and computer-aided planning of chemical syntheses.

Bartosz A. Grzybowski graduated from Yale University in 1995 and obtained his Ph.D. from Harvard in 2000. After over a decade at Northwestern, he recently moved to South Korea where he is now a Distinguished Professor of Chemistry at the Ulsan Institute of Science and Technology and a Group Leader in Korea's Institute for Basic Science. He is also a Professor at the Institute of Organic Chemistry of the Polish Academy of Sciences in Warsaw. He received numerous awards including the 2006 ACS Unilever Award, and the 2013 Nanoscience Prize.

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 1: Comparison of chess, the Rubik's cube and chemical synthesis.[a]

	Chess	Rubik's cube	Chemical synthesis
			
Number of players	Two	One	One
Movements	Small set of moves defined for each piece, some moves may not be allowed for some positions	Rotation of cube's single layer; always the same number of moves allowed	Very large (> 10 000) number of possible moves (i.e., reaction rules); applicable moves depend on the structure of the molecule; database of moves can grow as chemistry advances
Start position	Always the same initial arrangement of pieces on the board; "white" player starts	(Random) rearrangement of the cube	Target that needs to be synthesized
Position	Current configuration of the pieces on the board	Configuration of the cube	Set of substrates/synthons at each step
End position	Check-mate or exceeding allowed time; draws also possible	Each of the six faces of the cube composed of one color	All substrates for target's synthesis judged as "available"
Score of the game	Won/lost/drawn/not finished	Solved/not solved; in addition, the time or the number of moves might be counted (less moves = better score)	Viable synthesis found/not found; viability ultimately confirmed by experimental execution; in addition to "hard" criteria (number of steps, yield) soft criteria such as "elegance" might be applied during evaluation
Complexity	Upper bound for positions with no promotions is $\approx 2\times10^{40}$;[10] common estimate of the average number of moves possible from a given position is 35 translating into $\approx 10^{123}$ possible 80-move games[11]	Over 4×10^{19} possible configurations; more than 2×10^{20} possible sequences of 18 moves[12]	On average, 80.2 distinct reactions can be applied to a non-trivial retron,[13] translating into $\approx 3.5\times10^{28}$ possible 15-step pathways and $\approx 1.2\times10^{57}$ possible 30-step pathways
Maximal number of moves	Theoretically infinite (if none of the players chooses to use threefold repetition or 50-moves rule) though longest recorded tournament game was 269 moves	Based on massive calculations, each Rubik's cube can be solved in no more than 20 moves[14]	Commercially available Halaven is made in 62 synthesis steps[15] which seems to be an upper bound for industrially relevant syntheses
Optimal solution	In general does not exist; in a particular position the winning/drawing strategy might exist	Exists, but is typically hard to identify	In general, no single solution can be objectively deemed as "optimal" as it depends on available substrates and/or the criteria applied (e.g., minimal number of steps, green conditions, no protection groups, etc.)

[a] Images taken from: (left/middle) wikimedia commons, CC-BY-SA-3.0 license; (right) modified from https://www.flickr.com/photos/usdagov/16714715557/, U.S. Department of Agriculture, CC BY 2.0 license.

level total-synthesis masters, they can combine an incredible amount of chemical knowledge and can process it in intelligent ways with rapidity never to be matched by humans. In retrosynthetic planning, even inexpensive desktop machines can consider thousands of matching reaction motifs per second and can identify those that would be difficult to discern even by expert chemists—in fact, even desktop computers can be distinctly superior to humans in their capability to recognize complex rearrangement patterns and multicomponent reactions. Of course, it could be argued that one might be able to recognize these motifs using human intuition. But this is like arguing that we could, using paper and pencil, "eventually" divide two ten-digit numbers to the precision of ten decimal places—why do so if we have a pocket calculator available? Our thinking about all synthesis-aiding programs is that they should be regarded precisely as "chemical calculators," accelerating and facilitating synthetic planning, rapidly offering multiple synthetic

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

options which a human expert can then evaluate and perhaps improve in creative ways.

2. Navigating a Known Chemical Space: Synthetic Planning Based on Literature-Reported Reactions

2.1. Simple and Not So Simple Database Searches

Although the practice of synthetic organic chemistry might not necessarily appear reliant on or even related to the issues of computing, synthetic chemistry was actually one of the first natural sciences that considered massive use of modern information technology. Indeed, already in 1957—that is, one year before Jack Kilby demonstrated the first practical integrated circuit—two Soviet scientists, G. E. Vléduts and V. K. Finn, envisioned[17] an "*information machine for chemistry*" that could store a "*practically unlimited amount of chemical information*" and then process this information to solve various user-specified tasks including "*(…in ascending order of complexity): (i) search for information about an individual chemical compound, (ii) search for chemical compounds possessing a certain given combination of characteristics (including structural indices), (iii) search for the classes of reactions into which a definite individual compound can enter, (iv) search for the class of reactions producing a particular chemical compound, (v) search for the class of reactions which are of the same type chemically and are characterized by a transfer of given structural elements … from the initial molecules into other definite structural elements of the final molecules, (vi) search for the reaction that will take place between given compounds under given conditions, (vii) search for ways of synthesizing a given compound from a definite number of permissible initial compounds, and so on.*" From the perspective of over 50 years one can but admire the authors' vision as they provided quite an accurate blueprint of modern chemical databases and their capabilities. Indeed, today millions of chemists worldwide can use repositories of published reactions such as the aforementioned Reaxys or SciFinder to effortlessly search for specific molecules or substructures, for specific types of transformations, for molecules/reactions sharing structural similarities with a desired compound, and many more. Yet, not all the capabilities from the Vléduts and Finn's list have been, until very recently, realized even with modern computational power. The case in point here is efficient navigation through the known chemical space (i.e., through the millions of published or patented reactions) to combine individual reactions into optimal synthetic pathways back-tracking from a desired target all the way to commercially available substrates (point [vii] on the list). The Reader might object pointing to tools like Reaxys' Auto Plan[18] which appear to produce optimal pathways (up to ten steps) by choosing one (or few) best options at each step, or SciFinder's SciPlanner[19] that allows the user to make optimal (depending on the user's criteria) choices at each step to ultimately arrange them into synthetic plans. We note, however, that if at each synthetic step one chooses the best available option, one does not necessarily end up constructing the best overall sequence. Say we are performing a retrosynthetic search based on published literature and are iteratively moving "backwards" from the target until we ultimately find commercially available substrates. For the sake of illustration let's assume that one step away from the target ("synthetic distance" from the target $d=1$), we find two suitable reactions—one with experimentally determined 80% yield, the other with only 60%. We choose the 80% option but all subsequent choices turn out to be poor (e.g., the best option at $d=2$ gives only 40% yield). The maximum yield we can thus achieve by exploring this "branch" of synthetic possibilities is $80\% \times 40\% = 32\%$. Now, had we chosen the 60% option at $d=1$, we could, perhaps, have gotten to a second step offering a 90% yield and the total yield over the pathway equal to $60\% \times 90\% = 54\%$. While the particular example discussed here is trivial, the general conclusion holds for arbitrary pathways—namely, the optimality (be it in the form of yield, or atom economy, or any other measure applied at each step) of the entire pathway is not known until the pathway is complete—meaning that we first have to propagate complete searches to identify full pathways and only after they are all at hand, identify the best one(s).

This, however, complicates the problem significantly. In the best-at-each-step strategy (known in computer science as "greedy" searching), only one option is selected at each step and the total number of options to explore and evaluate to identify the "best" pathway of L steps is merely the sum of the numbers of options, n_i, evaluated at each step i, $\sum_i^L n_i$—typically, there are tens to few thousand options at each step and such searches are doable even on a desktop computer. If, however, we would like to expand our searches to explore m "best" options at each step, the number of synthetic evaluations to perform increases exponentially as m^L. As long as we consider only few "best" options at each step and the pathways are not too long (in Reaxys' Auto Plan, $L<11$), the numbers are still manageable, at least on a powerful computer—for instance, $3^{10} \approx 60\,000$ for $m=3$ and $L=10$. But if we would like to search for longer optimal syntheses leading to more complex targets, the number of possibilities to consider would rapidly become astronomical—as an example, for syntheses comprising $L=30$ steps (e.g., Fukuyama's synthesis of (+)-manzamine A or Woodward's, Shibasaki's or Magnus' syntheses of strychnine) and m being limited to 5, we would have to evaluate $5^{30} \approx 10^{20}$ individual synthetic steps. Searching through such an enormous space of possibilities requires quite elaborate algorithms and also data structures most appropriate to store chemical information.

2.2. Reaction Databases versus the Network of Organic Chemistry (NOC)[3]

Generally speaking, sets of data/entries between which pairwise relationships/"connections" exist can be represented either as lists of these connections or as networks. As an example, consider the recently launched fast-train connections in Poland. The left part of Figure 1 a lists the specific trains running between Poland's major cities whereas the

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

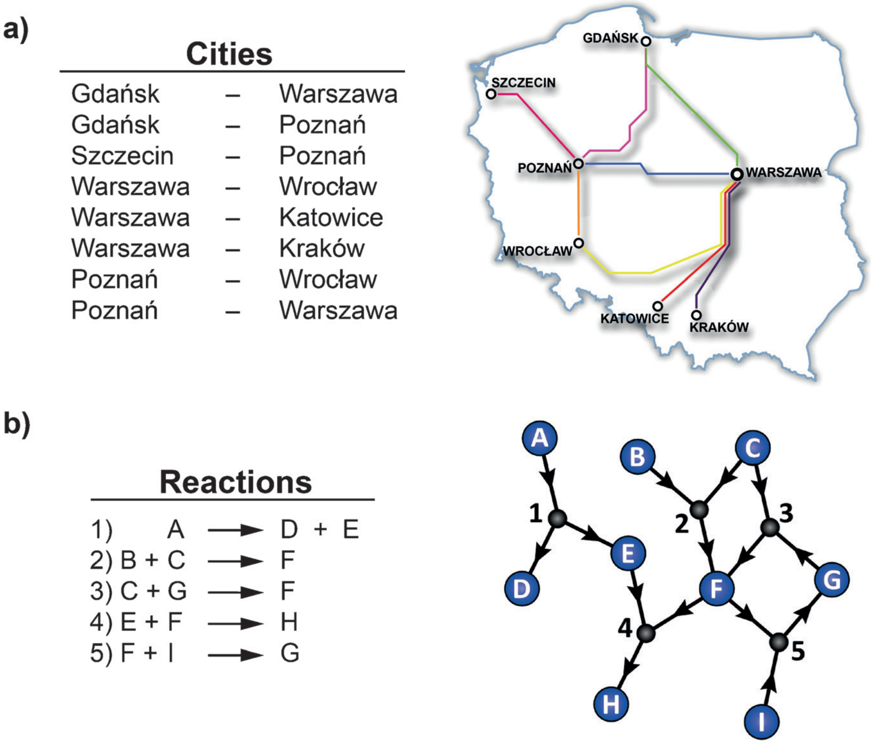


Figure 1. Lists versus networks. a) Inter-City connections of the Polish Railroads presented as (left) a list and (right) a network/map. The network tells us almost immediately that one can travel from Szczecin to Warsaw (via Poznan). b) A list of chemical reactions (left) and its equivalent network representation (right). Notice that two types of nodes (blue for substances, dark for reactions) are used to capture all relationships between substrates and products. If one just drew arrows from all substrates to all products there would be misleading connections like B→F and C→F; in reality both B and C need to be reacted to make F. This type of two-node representation is called a Petri net or a bipartite graph. Figure b reproduced by permission from Ref. [26a].

right part has the same information represented as a network. Clearly, it is much easier to use the network representation to momentarily determine that a connection exists between Szczecin and Warsaw or that Warsaw is the central hub of this railway network. This ease of "visual" inspection is one appealing reason to use network representations in a wide range of applications, from maps, to GPS navigators, to the analysis of big data in sociology[20] and biology.[21] The second and even more important reason is that searching for connections over networks is—with the algorithms we will discuss later, in Section 2.3—computationally more efficient than searching for connections over lists, with the difference becoming more significant as the size of the dataset increases. In synthetic planning, large datasets are queried for multistep connections between the target and available substrate molecules, and so we wish to represent chemistry as a network rather than a list of entries in common chemical databases.

Before proceeding further, it is important to choose the network representation most proper for chemical reactions. With reference to Figure 1b, consider a reaction of type B + C→F. If we draw all relationships between substrates and products (B→F, C→F), we might be introducing into the network chemically nonsensical connections—for example, if the reaction is acylation of some complex molecule B with, say, acetyl chloride (C), then the C→F connection would imply that the large, acylated molecule F can be made from acetyl chloride. The way to avoid such problems and capture all relevant chemical information is to use the so-called bipartite or Petri network representation[22] in which there are two types of nodes, one denoting the substrate/product molecules (blue circles in Figure 1b), and one denoting the reaction operations (black circles in Figure 1b; in our specific example, node labeled "2"). Chemicals B and C can then be imagined as entering a reaction vessel (node "2") to emerge as product F.

With these considerations, any repository of chemical reactions can be translated into a network. We have started the work on such a translation in the early 2000s—today, the Network of Organic Chemistry (NOC) contains on the order of ten million substances and a similar number of reactions connecting them. It is, by all accounts, a very large network (Figure 2a,b), some 1000 times larger than a human metabolome.[23] Despite being created by so many independent agents (i.e., chemists) this vast "universe" of known organic chemistry has been evolving in surprisingly predictable and immutable ways from its inception. For example, the plots in Figure 2c show that the numbers of molecules that have a given number of "incoming" connections/reactions, k_{in} (i.e., the number of times each of these molecules was obtained as a reaction product) and the numbers of molecules that have a given number of "outgoing" connections/reactions, k_{out} (i.e., the number of times each of these molecules was used as a reaction substrate) are both linear on a doubly-logarithmic scale. This mathematical regularity tells us that NOC has the so-called scale-free architecture[24a]—akin to the WWW,[24b,c] the internet,[24d] metabolic networks,[24e] and even societies[24f]—characterized by the presence of highly connected "hub" molecules through which most synthetic traffic takes place (for detailed discussion and the list of such molecules, see Ref. [3b]).

Importantly, to enable *rapid* searches for synthetic pathways, the NOC uses a special data structure tailored for large graphs/networks and one in which both molecules and reactions can be labeled with desirable attributes (molecular masses, solubilities, yields, etc.) to allow for user-specified criteria and/or constraints during network searches. This brings us to the topic of the search algorithms.

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

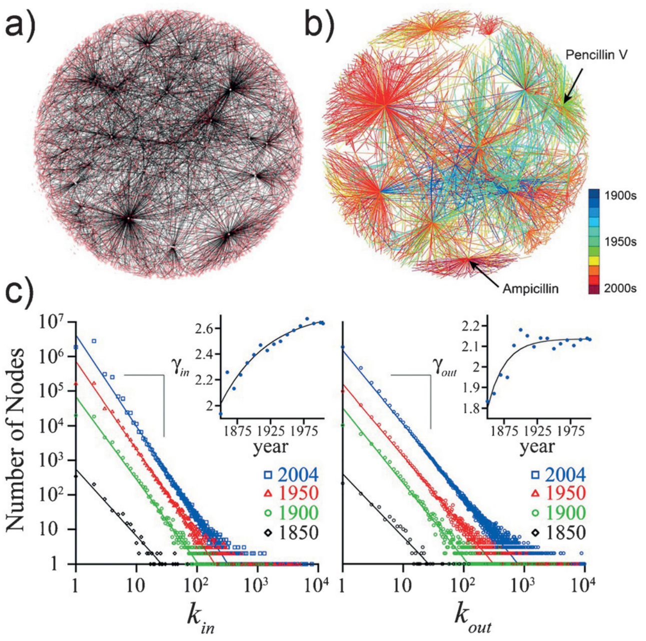


Figure 2. The structure and dynamics of the Network of Chemistry (NOC). a) A small (≈5500 nodes) fragment of NOC where individual nodes represent the molecules and arrows represent reactions. The entire "universe" of known organic reactions is more than 1200 times larger than the sub-network shown here and ca. 1000 times larger than the human metabolic network.[24] The representation in (b) has the reaction arrows colored by the times these reactions were first reported to illustrate "exploding" interest in certain areas of chemistry (e.g., synthetic activity around the Penicillin V node in the 1960s, following the first total synthesis). c) The plots show how many molecules ("nodes", vertical axis) in the NOC have specific numbers of either "incoming" (k_{in}, horizontal axis in the left graph) or "outgoing" (k_{out}, horizontal axis in the right graph) synthetic connections. The linearity of the plots on the log–log scales indicates a power-law distributions, $p(k) \propto k^{-\gamma}$, characterizing scale-free networks (see main text). The insets show that as time passes, the exponents γ of the distributions (i.e., slopes of the log–log plots) asymptotically approach the values $\gamma_{in} = 2.67$ and $\gamma_{out} = 2.14$ close to those characterizing the directed network of the WWW (2.71 and 2.1, respectively). In other words, the NOC and the WWW are topologically similar. Figure reproduced by permission from Refs. [3a,25b]

2.3. *Scoring Functions and Searches for "Optimal" Pathways*

Let's us first estimate in more detail (versus Section 2.1) the complexity of the search problems we wish to attack. Figure 3a shows a realistic view from the "inside" of the NOC—it is a highly connected network with many "branches" emanating from each node. Even intuitively, one can envision the numbers of possible pathways exploding with the numbers of steps taken. In fact, we have previously studied[25a] the dependence of the numbers of synthetic possibilities on the distance from a desired synthetic target (the so-called search depth, Figure 3b). The results summarized in Figure 3c show that while the numbers vary depending on specific target molecules, the average number of synthetic routes to consider within just five steps from the target can be as high as 10^{16}.

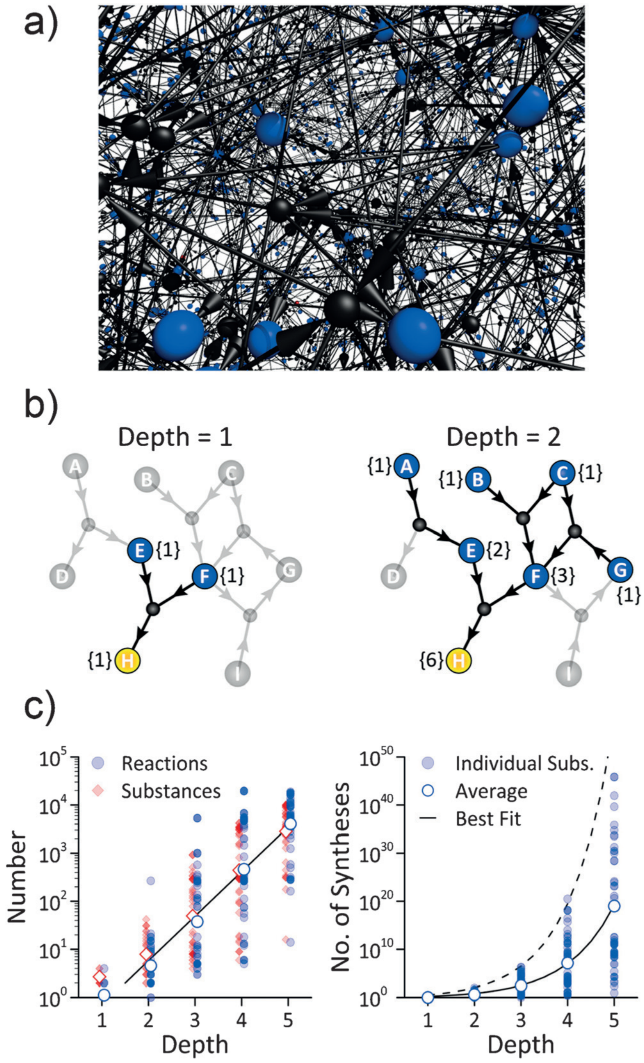


Figure 3. Complexity of the Network of Chemistry. a) A realistic view from the inside of NOC illustrates its high connectivity. b) Counting possible syntheses at two different depths, *d*, from the target product (yellow node denoted H). At depth one, there is only one possible synthesis using E and F as substrates. At depth two, there are six possible syntheses of the target product. To see this, note that substance F can be obtained in three ways: 1) it can be purchased and used as a starting substrate, 2) it can be synthesized from substrates B and C, or 3) it can be synthesized from substrates C and G. Similarly, substance E can be obtained in two ways. Depending on how substances E and F are sourced, there are six possible syntheses for product H. c) Based on network searches in the vicinity of 51 different target substances (for details, see Ref. [25a]) the number of individual reactions (blue) and substances (red) relevant to the synthesis of each target increases exponentially—as $\approx (8.5)^d$—with increasing distance from the target (left). The number of possible syntheses (i.e., plans combining individual reactions) grows even faster—here, as $\approx (1.4)^{(2.7)^d}$ as illustrated by the solid black curve (right). Here, the transparent markers correspond to results for each of 51 substances; the open markers represent the geometric mean of those data; the solid curve is the least-squares fit to the data; the dashed line is an upper-bound for the estimates. Figures (b) and (c) are reproduced by permission from Ref. [25a].

With this complexity in mind, we aim to identify synthetic pathways that would produce a desired target molecule from

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

commercially available substrates while optimizing the global (as opposed to local, "at-each-step," cf. Section 2.1) pathway score. The commercially available substrates along with their actual prices per unit mass[25a] come from vendor catalogs—in its standard form, NOC is connected to the Sigma–Aldrich selection though other catalogs can easily be interfaced as text files. The "score" for any synthetic pathway can depend on the attributes of the reactions (e.g., yields, labor cost of executing reactions) and/or on the attributes of the molecules involved (in particular, on the prices of the reactants, but also network-derived measures such as connectivity within the network, k_{in} and/or k_{out}, see Figure 2c,d). As discussed at the end of Section 2.2, these attributes accompany the nodes/molecules and reactions in the NOC's graph-database format, allowing us to set up various types of "scoring" functions according to which the pathways are evaluated. We illustrate two general types of such functions:

2.3.1. Cost Functions

Here, we aim to find pathways for which real, monetary cost is minimal. For this type of searches, the scoring metric is the total cost of a synthetic pathway, C_{tot} which can be expressed as a sum of the cost of individual reactions (including labor, overhead, purification procedures) and the cost of the commercially available starting materials (which depends on the vendor list connected to the NOC). With a reasonable approximation that the labor cost of carrying each reaction per unit mass is roughly constant, $C^{\mathrm{o}}_{\mathrm{rxn}}$, the simplest way to write such a function[25a] is $C_{\mathrm{tot}} = C^{\mathrm{o}}_{\mathrm{rxn}} N_{\mathrm{rxn}} + \sum_i C_{\mathrm{sub}}(i)$, where N_{rxn} is the number of steps/reactions in the pathway. Naturally, more elaborate functions can also be used in which the component costs are corrected by experimental reaction yields, γ, to take into account the "efficiency" of individual transformations. Also, we note that $C^{\mathrm{o}}_{\mathrm{rxn}}$ is a practically important parameter to consider, as it can be used to specify how expensive labor is compared to the cost of purchasable materials. For example, if one is operating in an economy in which labor is relatively inexpensive while substrates are pricey, $C^{\mathrm{o}}_{\mathrm{rxn}}$ should be set to a low value; in this scenario the scoring function will favor longer pathways leading to cheaper substrates. If, however, labor is expensive (as is the case in the U.S., German, or Swiss chemical laboratories), $C^{\mathrm{o}}_{\mathrm{rxn}}$ should be set high, and the function will favor shorter pathways while using more expensive substrates.

2.3.2. Popularity Function

A practical synthetic query might be to look for syntheses that use very popular chemicals, as such chemicals are generally easy to handle and reliable. In network formalism, the synthetic popularity can be measured by the connectivity of the molecule as quantified by the k_{in} and/or k_{out} indices discussed in Section 2.2 (cf., Figure 2c,d). Based on such indices, the popularity scoring function, P_{tot}, could be such as to minimize the sum of the inverse connectivity indices, $\sum_i 1/k(i)$.

2.3.3. Search algorithms

Of course, the above examples of the scoring schemes can be modified and it is straightforward to define combinations of the above functions or add more variables—we will see some of these capabilities later on, in Section 2.5 where we will discuss Synthesis Optimization with Constraints (SOCS). Irrespective of the function used, however, one must have an algorithm that would traverse the graph efficiently to construct the pathways to be scored. A good candidate for such a traversal is a variant of the so-called breadth-first-search (BFS)[26] algorithm which recursively propagates on the NOC starting from the target (cf. pseudo-code in Figure 4a). For example, to find the pathway with the lowest

a)

```
MinCost(substance s, depth d)
  if s.cost(d) < 0  // substance not yet visited
    if s.type == substance
      s. cost(d) = s.purchase_price
    else
      s. cost(d) = INF  // infinite cost
    if d < d_max
      for each reaction r ∈ {incoming reactions of s}
        if r.mrk(d) == 0 // reaction not currently being explored
          if r.cost(d) < 0 // reaction not yet visited
            r.cost(d) = c°_rxn
            r.mrk(d) = 1
            for each substance u ∈ {reactants of r}
              r.cost(d) = r. cost(d) + MinCost (u, d + 1)
            r.mrk(d) = 0
          if r.cost(d) <  s.cost(d)
            s.cost(d) = r.cost(d)
  return s. cost (d)
```

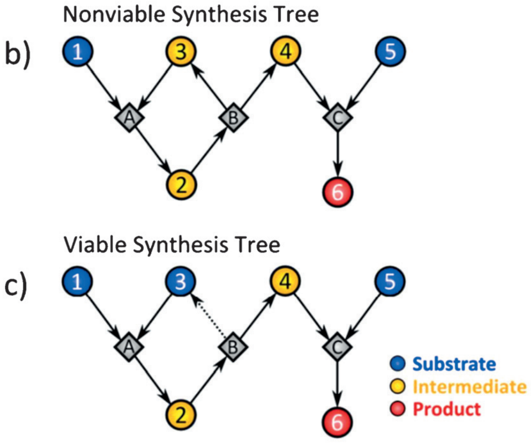


Figure 4. a) A pseudocode of a recursive algorithm to search for the lowest-cost pathways. b,c) Non-viable versus viable synthetic pathways. Reactions that can be performed using commercially available substrates or substances derived from such substrates are termed ***viable***; similarly, substances that can be synthesized from commercially available substrates are termed ***makeable***. Example in (b) is a nonviable synthesis plan. None of the three reactions are viable since intermediates 2, 3 and 4 are not makeable. Example in (c) is a viable synthesis plan. Reaction A is viable because all of its reactants are commercially available substrates; consequently, substance 2 is makeable. Reaction B is viable because its reactant is makeable; consequently, substance 4 is makeable. Finally, reaction C is viable because each of its reactants is either makeable (4) or a commercially available substrate (5). Therefore, the target (6) is makeable. Figure are adapted by permission from Ref. [25a].

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

monetary cost (see 2.3.1), the first "backward" step of the algorithm examines all reactions leading to the target and calculates the minimum cost for each of them. This calculation, in turn, depends on the minimum costs of the associated reactants, which may be purchased or synthesized. In this way, the cost calculation continues recursively, moving backward from the target until a critical, user specified search depth (i.e., maximal allowed length of the pathway) is reached.

The pathways thus generated need to be examined for synthetic viability. This is an important issue since the network is not a simple "tree" and there is a possibility of loops and interconnecting branches. An illustrative example is shown in Figure 4b and 4c: even though the pathways are similar and both have commercially available substrates as end points, only the one in Figure 4c is viable. The pathway in Figure 4b is not viable since intermediates 2, 3, 4 cannot be synthesized from starting materials (e.g., 2 requires 3 and 3 requires 2, none of which are commercially available). The algorithm to assess synthetic viability of the pathways is included in the supporting information to Ref. [25a].

In Refs. [25a–c] we showed that the searches give some interesting and synthetically relevant results. In another example in Figure 5, the cost algorithm finds the optimal synthesis of the zolpidem, used for treatment of insomnia and brain disorders. For the high-labor-cost conditions, $C_{rxn} = 7.5$, the algorithm uses relatively expensive substrates (especially, propiolic acid) but completes the synthesis in just two steps via a three-component reaction. However, if the cost of labor is low ($C_{rxn} = 0.075$), the optimal synthesis comprises seven steps starting from the common and inexpensive 2-aminopicoline and p-methylacetophenone.

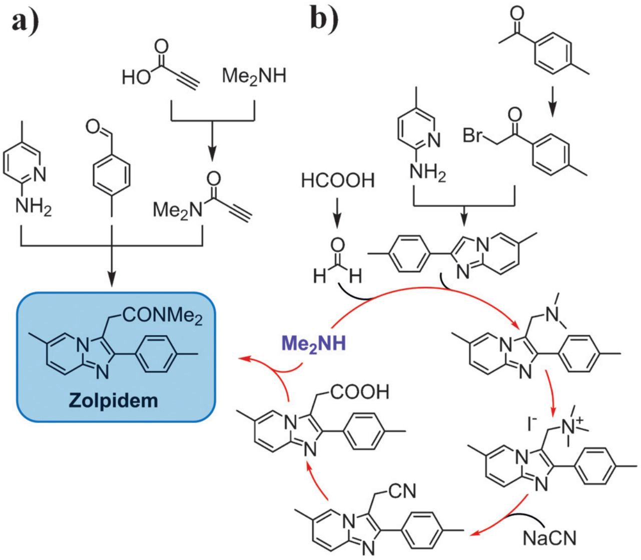


Figure 5. Two different optimal syntheses of zolpidem for relatively low (right, $C_{rxn} = 0.075$) and high (left, $C_{rxn} = 7.5$) cost of labor.

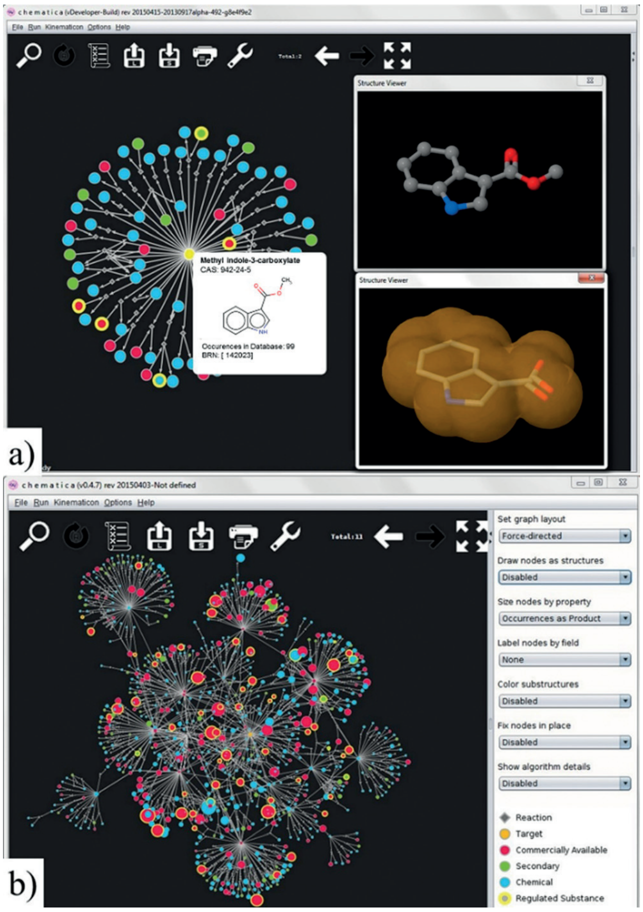
Noteworthy, one of the building blocks (dimethylamine, denoted violet) is used not only for final amidation but also for the introduction of an appropriate leaving group necessary for homologation of Mannich adduct prepared in the preceding step. We observe that identifying this pathway via traditional, one-step-at-the-time searches would be extremely improbable. This is so because, as one back-propagates the searches from the target, the "left" and the "right" subtrees in the synthetic plan diverge and the chance of finding an inexpensive common intermediate connecting the two "branches" is very low.

Notwithstanding such elegant examples, the underlying BFS (breadth-first search) or DFS (depth-first search) algorithms are limited in two fundamental aspects. First, their speeds might be satisfactory for relatively simple molecules (seconds to minutes), but for larger targets and longer syntheses (tens of steps), the number of possibilities to consider cannot be handled even by relatively large computer clusters. Second, the data structures and algorithms lack common, optimal system architecture, often reducing speeds of the tasks performed. These needs are met in a unified software environment we called Chematica.

2.4. NOC Searches in Chematica

Chematica supports various types of NOC searches. The simplest search, called the Network Travel (see Movie S1 in the Supporting Information) displays reactions in the NOC which, depending on user's preference, lead either to (Figure 6a) or from (Figure S2a) the molecule of interest (here, methyl indole-3-carboxylate). The small, diamond-shaped reaction nodes have basic information about the particular reaction (e.g., literature sources), green and blue nodes denote known molecules (with the difference that green stands for reaction by-products), red nodes denote commercially available chemicals, and the yellow halos denote regulated and or toxic substances. Each or all of the molecule nodes can be visualized as 2D molecular structures (Figure 6 and also Figure S2b) or as 3D models on which various types of analyses (geometry optimizations, Connolly surfaces, etc.) can be performed (sub-windows in Figure 6a). In addition, each molecular node has associated with it information about its "synthetic popularity" (i.e., molecule's connectivity in the NOC as quantified by the k_{in} and k_{out} indices discussed in Section 2.2). The sizes of the nodes can then be made proportional to various molecular properties—including the said "synthetic popularity" (see Figure 6b). Also available are the time trends of how this popularity has evolved in time, often reflecting currently "hot" areas of chemistry; cf. Figure S3a). Each of the "daughter" nodes on the display can also be further expanded (see Movie S1), which within few steps can give a complicated web of connections—pictures like the one in Figure 6b illustrate why manual searches are quite hopeless, and why it is important to have implemented automatic search routines using the previously discussed cost and/or popularity functions to identify optimal synthetic pathways.

Figure 7a illustrates the search for the minimal-cost synthesis of an anticancer agent, camptothecin, up to $L = 10$ steps long and with the price of labor set commensurate with the price of reactants. The identified pathway is displayed


a)
chematica (vDeveloper-Build) rev 20150415-20130917alpha-492-g8e4f9e2
File Run Kinematicon Options Help
Total:2
Structure Viewer
Structure Viewer
b)
chematica (v0.4.7) rev 20150403-Not defined
File Run Kinematicon Options Help

Methyl indole-3-carboxylate
CAS: 942-24-5
Occurences in Database: 99
BRN: [142023]

Set graph layout: Force-directed
Draw nodes as structures: Disabled
Size nodes by property: Occurrences as Product
Label nodes by field: None
Color substructures: Disabled
Fix nodes in place: Disabled
Show algorithm details: Disabled

◆ Reaction
● Target
● Commercially Available
● Secondary
● Chemical
● Regulated Substance

Figure 6. Basic network traversal in Chematica. In the specific example, screenshot in a) shows substances that can be made directly from methyl indole-3-carboxylate. The nodes can be displayed as molecular structures and also as 3D models on which basic modelling can be performed (for instance here, sub-windows on the right illustrate geometry optimization and Connolly surfaces). Blue nodes denote products; green nodes are minor/side-products, red nodes denote commercially available substances, yellow halos denote regulated substances (e.g., on the DHS or EPA lists Chematica interfaces with). For the network of reactions leading to (rather than "from") and also other display options, see Figure S2. b) Each of the "daughter nodes" can be further expanded (see Movie S1). Within just few steps the network becomes quite complex. In the display mode shown, the sizes of the nodes are proportional to their "synthetic popularity"—that is, connectivity of the molecules in the NOC. Some of the network-hub molecules (i.e., largest nodes) denote molecules that are used in tens of thousands of syntheses!

using molecular structures, which can be toggled into the node representations (Figure 7b; node coloring as in Figure 6). The numbers displayed next to the nodes estimate the prices per gram calculated based on the actual dollar prices of substrates (here, from Sigma–Aldrich catalog). By changing the labor-to-chemicals cost ratio, the optimal pathways usually change quite dramatically. In the present example, setting labor cost at 20 times more than the cost of substrates heavily favors shorter pathways but starting from more expensive substrates—Figure 7c shows molecules involved in such a pathway with the network representation and pricing shown in Figure 7d.

For the relatively short pathways we have seen so far, the traditional BFS searches adapted to synthetic planning (cf. Figure 4) and combined with the efficient graph-database structure of the underlying the NOC were sufficiently rapid to yield answers within seconds. The same routines, however, are not enough to search for optimal syntheses of complex targets for which the number of possibilities becomes enormous. Since the same limitation would be encountered with other common search algorithms (depth-first searches,[27a] Dijkstra,[27b] etc.) the problem might seem unsurmountable—still, it can be overcome by taking inspiration from chess-playing programs. Therein, one of the algorithmically efficient strategies is to keep a "library of end-games" pre-calculated so that when the program estimates potential moves, the tree of possibilities does not have to be fully expanded—instead, when the program encounters one of the known end-games, it already has an optimal solution for it available (i.e., it has to perform analysis only up to encountering this end-game). In our case, Chematica first calculates the optimal, relatively short M-long pathways to all the compounds in the NOC. When it is then queried to find a long, N-step ($N > M$) pathway to a given target, it only has to perform searches up to the depth of $N-M$, and the M-long endings are already available, in effect accelerating the searches. While the algorithmic details are quite complicated, a judicious choice of M allows for the overall search speeds to be accelerated by several orders of magnitude.

With this algorithm, even very long routes to very complex molecules are identified within few seconds. One example is illustrated in Figure 8a where the program finds the least expensive, up to 50-steps-long synthesis of paclitaxel (Taxol).[28a,b] The entire search (considering 345 million sequences of possible steps and over 400 million combinations of participating substances) performed on a four-core desktop computer takes only seven seconds (cf. real-time Movie S2). We wish to re-emphasize that in this and all other searches of the NOC there is no question as to the feasibility of the pathway since all individual reaction steps have been performed experimentally and published in the literature. What the algorithm does is to put together these individual reactions into an overall optimal pathway. In some cases, optimal pathways reflect total syntheses of individual groups—for instance, most of Taxol's synthesis in Figure 8a follows the strategy by Danishefsky (arrows colored red) though the algorithm finds less costly alternatives for the syntheses of the starting materials like the Wieland–Miescher ketone,[28c] 2-methylcyclohexadione,[28d] or the silylating and mesylating reagents which are simple, but quite expensive. A more general outcome, however, is that the optimized syntheses are "patchwork" of approaches from different groups, sometimes working on the same target but sometimes completely unrelated. This is illustrated in Figure 8d,e where the cost-optimized route to vardenafil (medication used to treat erectile dysfunction) is assembled from reactions published as early as 1952 (alkylation of benzamide with ethyl bromide) to 2008 (sulfonylation of vardenafil precursor). This time-range and the "structure" of the synthetic route are visualized by coloring the reaction arrows and/or providing date labels (Figure 8b,d).

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

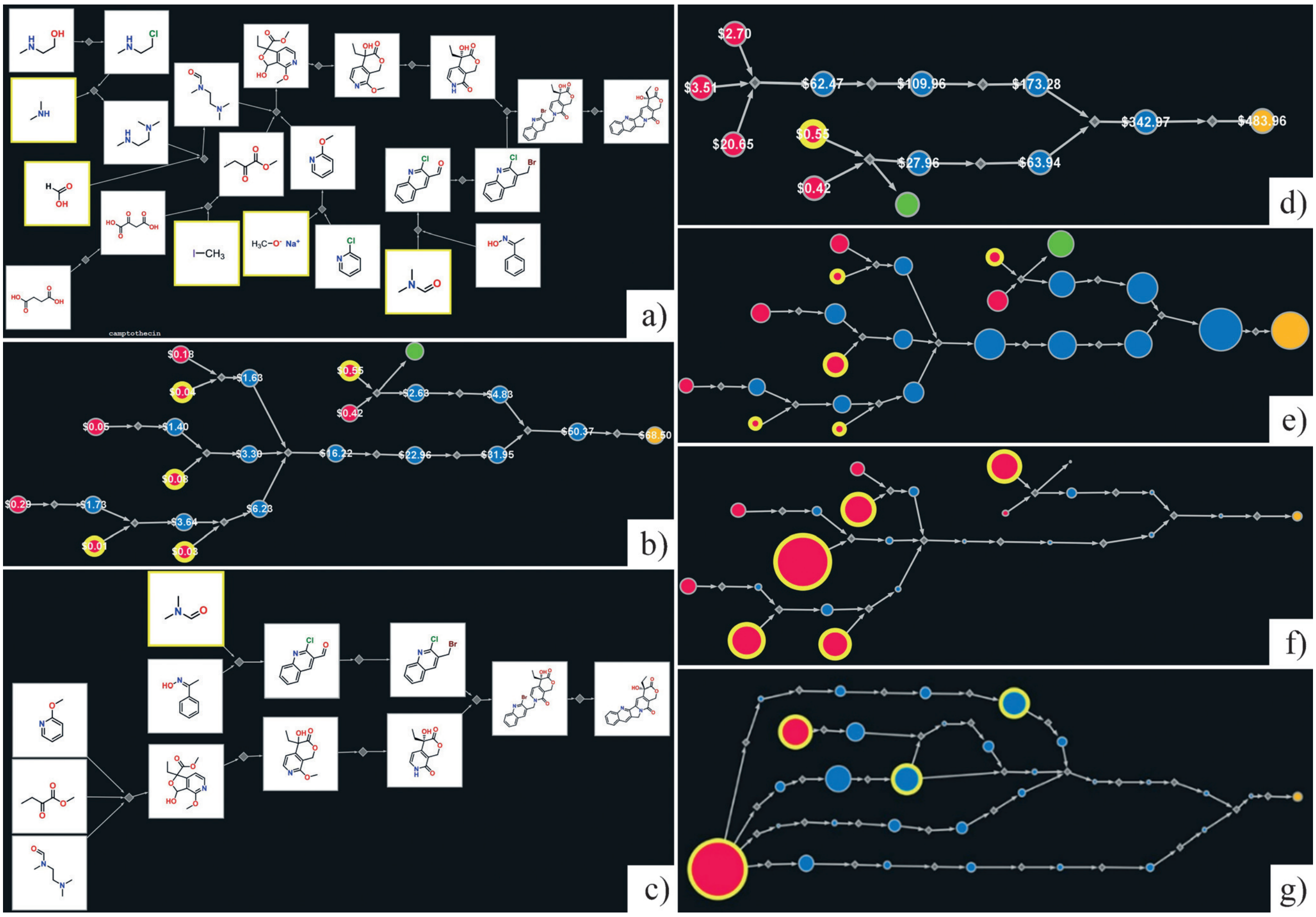


Figure 7. Chematica's snapshots of the optimal pathways based on literature-reported reactions and leading to camptothecin. Display based on a) molecular structures and b) nodes for the cost-optimal pathway up to ten steps long. In (b), the numbers next to the nodes give dollar prices per gram calculated from the realistic prices of Sigma Aldrich substrates (red nodes). c,d) Cost-optimized pathways but with labor cost set to 20 times that of chemicals. While the pathways are shorter, they start from more expensive reactants (to save on labor) and the overall price of the synthesis in (d) is ca. six times higher than the price in b). e,f) Pathways identical to those in (a) and (b) but with nodes sizes proportional to e) molecular weights of the chemicals and f) their popularity. This kind of node sizing allows for rapid assessment whether synthesis progresses from smaller to larger and from more to less popular molecules, as is, indeed, desirable (otherwise we would be making smaller targets from large substrates and common chemicals from specialized/unpopular substrates). In the examples shown, both conditions are met. Finally, in addition to minimal-cost, several other search options are available. One example is shown in Figure 7g where the optimality of the synthetic plan is defined by the popularity (i.e., network connectivity) of all substances involved being maximized "globally" over the entire pathway—this type of a search identifies syntheses that rely of robust (i.e., popular) chemistries, although such pathways might not be characterized by a minimal dollar cost. Color-coding of the nodes: blue = intermediates; green = minor/side-products, red = commercially available; yellow halos = regulated substances. Detailed reaction information is available upon clicking on the "diamond"-shaped reaction nodes. Structures of molecule miniatures in (a) and (c) redrawn for better contrast.

2.5. Synthesis Optimization with Constrains (SOCS)

Monetary cost or popularity of substances involved are just two of several factors that might be taken into account during synthetic planning. For example, one might also wish to design a route that, in addition to being the most economical one, involves a specific desired intermediate (e.g., easily accessible popular molecule), or avoids all substrates and intermediates that are regulated and/or toxic, or uses only substances that are water soluble (for green chemistry[29]). Such requirements translate into evaluating extraordinarily large numbers (again, billions) of possible synthetic pathways with multiple optimization criteria/constraints being applied *simultaneously*. Such Synthetic Optimization with Constraints (SOCS) is clearly beyond human cognition both in terms of the number of possibilities to consider as well as complex logical conjunctions that would—if performed manually—entail cross-referencing synthetic feasibility, cost catalogues, toxicity data, lists of regulated substances, or solubility values. On the other hand, computers can handle multiple logical operations (ORs, ANDs, IFs) on various chemical criteria/constrains with ease, at the cost of only few extra lines of code.

In addition to the cost versus labor parameter discussed earlier, Chematica's SOCS scheme supports various types of constraints including the maximum number of reaction products, the time span of the reactions considered, solubility of the substances involved, avoidance of user-specified intermediates, or avoidance of the any regulated (i.e., toxic or dangerous) chemicals (for details, see SI, Section S4).

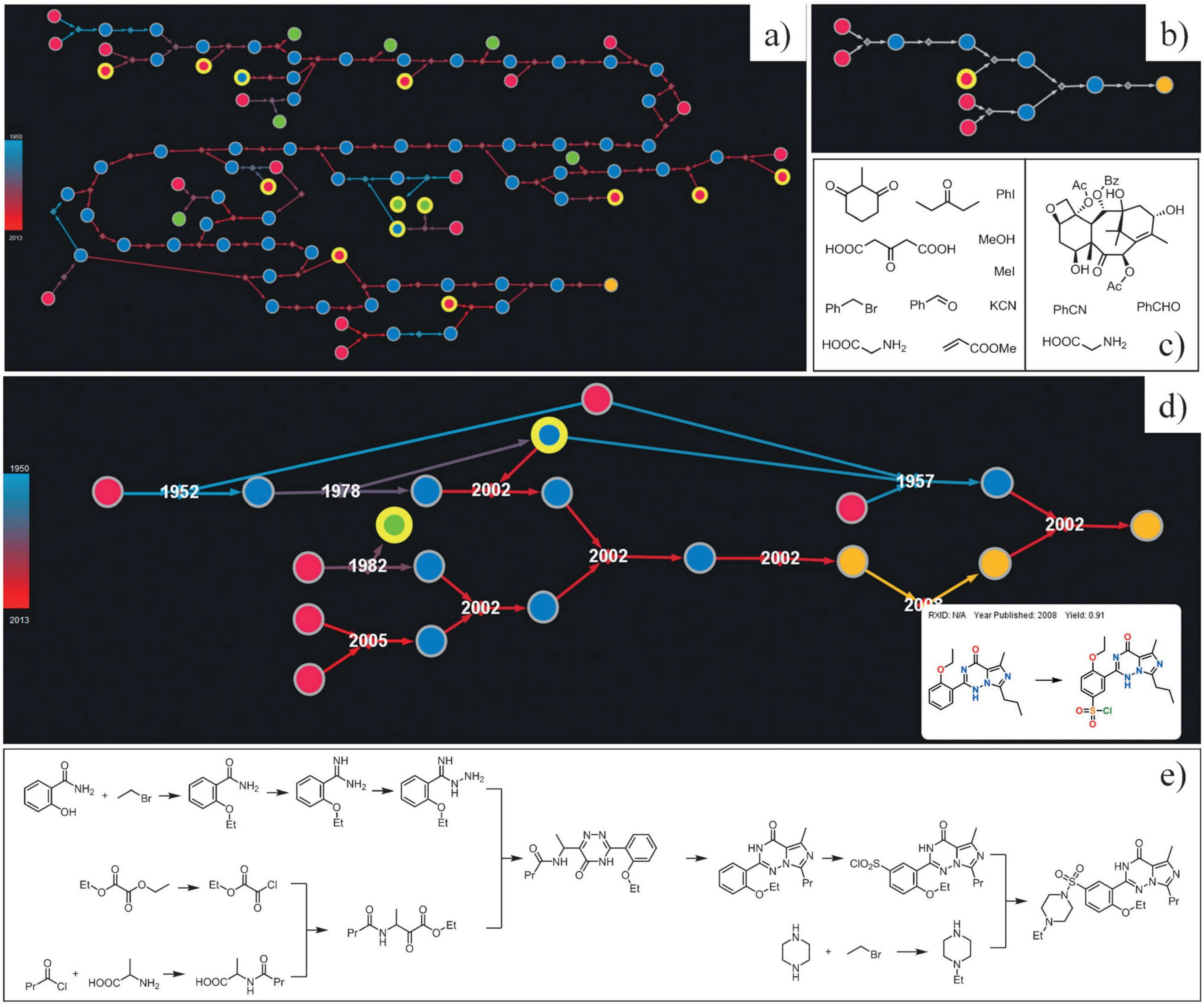


Figure 8. Cost-optimized syntheses of paclitaxel (Taxol) and vardenafil. a) Node representation of the lowest-cost synthesis of paclitaxel limited to *L* = 50 steps (cf. Movie S2). Colors of the reaction arrows correspond to the years in which a particular reaction was reported. Most of the pathway shown (red arrows) is based on Danishefsky's synthesis from 1995.[28a,b] b,c) Increasing the cost of labor by a factor of ten renders the long syntheses very costly—consequently, the lowest-cost solution the algorithm finds starts from the naturally occurring and commercially available taxane (baccatin III; cost from Sigma–Aldrich several thousand dollars per gram). This route resembles first industrial method of paclitaxel's production[28e] further elaborated in Ref. [28f]. c) Sets of commercially available starting materials used for synthesis of paclitaxel in (a) and from 10-deacetylbaccatin III in (b). Reagents used for protections are omitted for clarity. d,e) Cost-optimized synthetic route leading to vardenafil. The optimal pathway identified spans chemistries from 1952 to 2008. Note a cost-reducing use of the same reagent in two different steps (ethyl bromide, top middle red node). Node coloring is the same as in Figure 7. Structures of molecule miniatures in (d) redrawn for better contrast.

For example, Figure 9 compares the synthesis of the gabapentin anticonvulsant drug cost-optimized without any constraints, and the synthesis of the same compound with the constraint that no toxic/regulated substances (nodes in yellow halos) be used. In the absence of any restrictions on regulated chemicals, the most cost-effective pathway (Figure 9a) starts from cyclohexanone (available in large quantities but regulated) undergoing Knoevenagel condensation with dimethyl malonate (step a) to give unsaturated diester which is then treated with inexpensive but extremely toxic potassium cyanide. Michael addition ensues (step b) followed by the reduction of nitrile to primary amine with nickel catalyst and gaseous hydrogen (risk of explosion), followed by aminolysis of ester to spiro-amide (step c). Synthesis of the target molecule is completed via decarboxylation and hydrolysis of the thus prepared amide (step d). When the algorithm is told to find the most cost-effective pathway while avoiding any regulated chemicals, it proposes a plan shown in Figure 9b, in which no regulated (cyclohexanone), toxic (cyanides) or explosive (hydrogen) chemicals are used. Naturally, the synthesis with an added constraint is more expensive than the "unconstrained" one (by some 40%)—in fact, satisfying any type of a constraint will always entail a trade-off of an increased price.

Another set of examples of SOCS is illustrated in Figure 10 and shows syntheses of ketoprofen optimized with

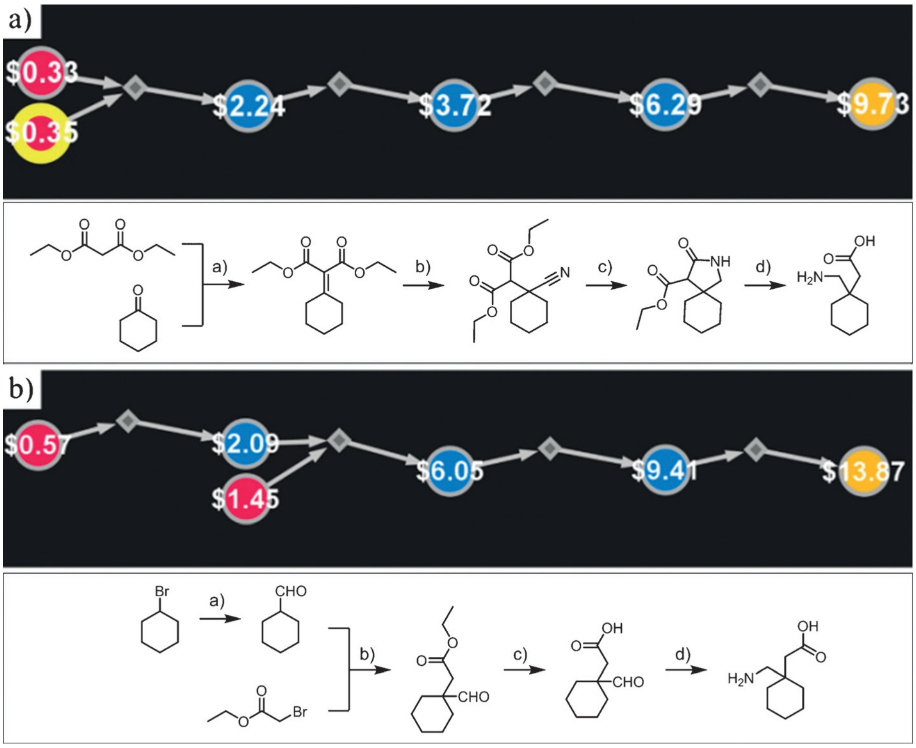


Figure 9. Cost-optimized synthesis of gabapentin anticonvulsant drug a) without any additional constraints and b) with the constraint that no toxic/regulated substances (in yellow halos) be used. For the reagents used, see main text. To avoid regulated substances in the pathway, three regulatory databases were applied (Australia Group, Department of Homeland Security, Environmental Protection Agency List of Regulated Chemicals). Prices per gram are estimated using the actual catalog prices from Sigma–Aldrich.

different constraints. The route with reaction nodes labelled red is cost-optimized while constraining the reactions to only those having one product—any side-products are undesirable in industrial settings, and so the hope here is that the constraint would produce an industrially relevant synthetic plan. The proposed plan avoids some intermediates commonly found in other pathways (e.g., benzaldehyde, ethylbenzophenonpropionate, and benzoyl chloride) and closely resembles an industrial method of ketoprofen synthesis[30a] patented by Rhône–Poulenc.[30b] The cost-optimized synthesis denoted by green markers allows for multiple products but bans any regulated substances. The proposed convergent pathway starts with the preparation of α-bromoacrylate through DMSO-mediated dehydrohalogenation of α,β-dibromopropanoate[30c] and 3-iodobenzophenone via solvent-free iodination of benzophenone[30d] (both reactions are mild and selective), followed by the coupling of an organozinc compound (derived from α-bromoacrylate) with aryl iodide.[30e] The result-

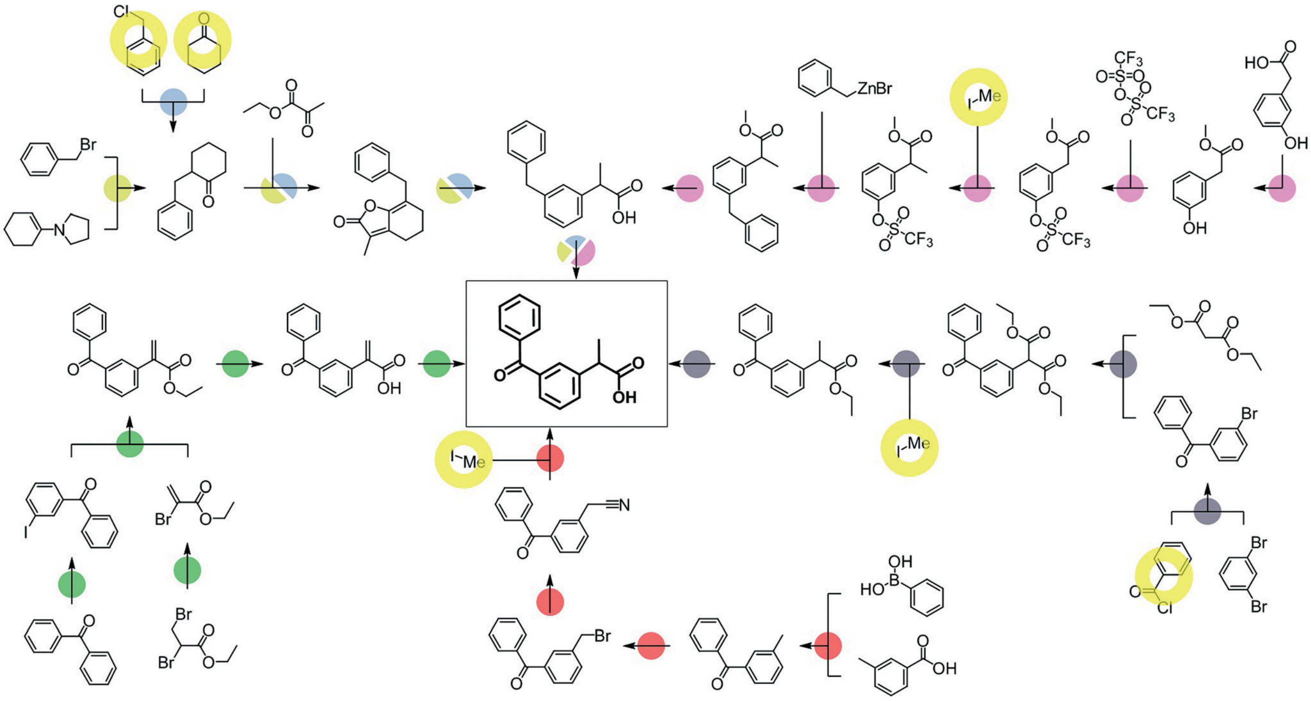


Figure 10. SOCS at work. Cost-optimized syntheses of ketoprofen, each designed by Chematica within few seconds with various additional constraints (see main text for detailed description). Yellow halos denote regulated substances.

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

ing unsaturated ester is then hydrolyzed and hydrogenated to give ketoprofen. The pathway denoted pink is a "contemporary" synthesis limited only to reactions published after 1998—with this constraint, the program identifies as the most cost-effective a synthesis published in one publication,[30f] except for the last step.[30g] The "blue" pathway[30h,g] is cost-optimized for high labor-to-chemicals ratio and with the aim to improve atom economy but starts with regulated benzyl chloride and cyclohexanone. By adding an additional constraint of avoiding any regulated substances, these two undesirable chemicals are replaced with unregulated benzyl bromide and cyclohexenyl pyrrolidine (the "greenish-yellow" pathway). Finally, the "grey-brownish" pathway is an example of how just excluding few substances (e.g., benzaldehyde, 3-chlorobenzophenone and 3-methylbenzophenone) can produce an entirely different cost-optimized route, this time closely following the one published in Ref. [30i].

To take-home message from this section is that the very meaning of an "optimal synthesis" depends on the criteria one imposes. The advantage of computers over humans in searching for such syntheses in the NOC is that the former can much more rapidly combine multiple sources of information and impose desired search constraints. While other types of searches and analyses over known reaction space are also possible and useful (e.g., see SI, Section S5 for simultaneous optimization against multiple targets[25a] and Section S6 for rewiring synthetic networks by replacing sequences of steps by one-pot "shortcuts"[25b]), they are by definition unable to suggest any novel synthetic strategies and/or pathways leading to targets that have not yet been synthesized. Teaching the computers to do such de novo design has been the grand challenge of computer-assisted synthetic planning to which we will now turn. We will continue to use network searches, scoring functions and constraints—this time, however, the networks will not be a priori known and static but, instead, will expand continuously as the computer makes synthetic choices guided by the chemical rules it has been taught.

3. Computer-Aided De Novo Synthetic Design

3.1. The Great Expectations

Exactly half a century ago, in 1965, a team of Stanford computer scientists and chemists (Edward Feigenbaum, Bruce Buchanan, Joshua Lederberg, and Carl Djerassi) initiated the so-called Dendral project[31] aiming to 1) determine structure of organic molecules from spectroscopic data and then 2) use artificial intelligence approaches to make computer plan synthetic routes—even though they never completed the second task, their pioneering effort marked the beginning of computer-assisted synthesis. Already two years earlier, in 1963, Vléduts proposed to develop software that would enable chemists to use information stored in databases and certain "chemical analogies" to navigate "backwards" from a desired target molecule to its precursors.[17] This proposal slightly preceded Corey's seminal 1967 paper in which he expounded the philosophy of "retrosynthetic analysis" and codified its key rules.[32] Since excellent books[33a,b] and reviews[33c,d] on this topic have been written, we only briefly mention that Corey's contribution went significantly beyond just proposing moving iteratively "backwards" on a growing tree of synthetic possibilities, and he was able to identify heuristics guiding the choice of specific synthetic disconnections (i.e., rules suggesting which bonds should be "cut"). Corey's problem-solving technique revolutionized the field making the design of new pathways more systematic while also providing the rules that could be—or at least it was thought so at that time—taught to the computers. Indeed, already in 1969 Corey and Wipke presented the first computer-aided synthesis design software called OCSS for Organic Chemical Simulation of Synthesis.[34] It was short-lived and the project split into two directions—LHASA[5] under Corey's supervision and SECS developed by Wipke.[35] LHASA (for Logic and Heuristics Applied to Synthetic Analysis) has been significant, among its other aspects, as one of the first retrosynthetic programs using a graphical interface to input and display chemical structures. Technically, it can be classified as a semi-empirical retrosynthetic planning software relying on various types of heuristic transforms written in an English-like chemical language called CMTRN (Chemistry TRaNslator) and also several design strategies as well as some group-protection data.[5b] One of the major drawbacks of the program has been its limited ability to deal with stereochemistry[36,33d] and its interactive (step-by-step) rather than automated nature to find full pathways. To the best of our knowledge, LHASA has not been under active development for several years[5b] and the lhasa.harvard.edu web-page appears inactive at the time of writing this Review.

After LHASA, there were numerous efforts to create other types of synthesis planning programs. Since the history of their development—and, unfortunately, ultimate demise of most of them—is very eloquently described in Professor Philip Judson's excellent book provocatively titled "*Knowledge-based Expert Systems in Chemistry: Not Counting on Computers*",[1] we will provide only a brief summary of the key approaches.

The aforementioned Simulation and Evaluation of Chemical Synthesis (SECS)[35] was developed by Todd Wipke largely building on the LHASA approach but extending its knowledge base. Although it received a substantial backing from a consortium of Swiss and German pharmaceutical companies, it was eventually disconnected for reasons that are not entirely clear.

SYNLMA[7] was an effort by P. Y. Johnson's group from the Illinois Institute of Technology and was significant because it separated the knowledge-base from its "reasoning component" based on logical operations to be applied during retrosynthesis. Unfortunately, the program ran into the "combinatorial explosion" problem generating excessively large retrosynthetic trees which it could not meaningfully prune. It disappeared from the scene already in 1989.

SYNCHEM[8] and its successors were under development at Stanford/Stony Brook already at the time of LHASA's initial publication, but the program came to light only in 1977. The truly innovative aspect of this approach—especially at the times when modern computing was in its infancy—was that it attempted to construct and explore (with BFS-like

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

searches) full retrosynthetic trees leading to few-thousand memory-stored commercial products and using on the order of few hundred expert-coded (but general-level) transforms. Unfortunately, the transforms were often coded only after a human chemist inspected the target molecule, and there were additional problems with the transforms' applicability in specific molecules. The generated strategies were found to be too "short-term" without a significant enough accounting for the paths' histories.[33d] As in all previous programs, stereochemistry was a major problem and regiochemistry was not considered. The last publication on SYNCHEM was in 1998[8b] and it described efforts to parallelize the code. Afterwards, SYNCHEM seemed to have joined other retrosynthetic programs in the Valhalla of computational chemistry.

SYNGEN was a program developed by Jim Hendrickson and his team at Brandeis—Jim holds a special place in this narrative as he was perhaps the only senior colleague who, back in 2001, encouraged the back-then postdoc B.G. to work on chemical networks and retrosynthesis. Hendrickson's SYNGEN program was developed in the 1970s and 1980s[36] and placed emphasis on the identification, supported by various types of heuristics, of reasonably-sized retrosynthetic trees containing the best, highly convergent routes. The synthesis was simplified to focus on skeletal construction and ignore refunctionalization as that would produce the shortest synthetic routes.[36d] The starting materials were identified using an empirical observation that three out of every four bonds in the target come from the starting materials. This enabled the computer to generate possible bond sets of the synthetic plan which could be refined by analyzing the synthetic space for paths leading to synthons containing the unchanged bonds. The program ran into issues when functionalization was lost before finding the full pathway, and an additional program, FORWARD, was developed to reintroduce functionalization in the synthetic direction. This effort, however, was never completed.

Turning into conceptually different types of approaches,[37] no story on computer-aided synthesis would be complete without mentioning the contributions of Ivar Ugi who introduced the concepts of logic-oriented synthetic planning relying on fundamental chemical knowledge to evaluate feasibility of not only known reactions but also potentially novel ones. In the 1980s and 1990s Ugi and co-workers developed programs like IGOR and IGOR2[38,31b] in which molecules were represented as bond-electron (BE) matrices and reactions as R-matrices obtained by subtraction of substrate and product matrices. The interactive (i.e., one reaction at a time) analysis of potential reactions relied on the rearrangements of valence electrons stored in the matrices and the selection of feasible versus nonsensical reactions was further guided by calculations of quantities such as reaction enthalpies. As reviewed in reference [31b] IGOR was able to identify several novel pericyclic reactions and a novel rearrangement of α-aminoalkylboranes to the corresponding β-dialkylaminomonoalkylboranes, which were later experimentally verified. On the other hand, the program was not really used in multistep synthetic planning perhaps due to the fact that operations on BE- and R-matrices are computationally costly and only limited numbers of reactions could be explored. The ultimate outcome—quite unfortunate to the development of the field—is that with the retirement and then passing away of Prof. Ugi in 2005, the effort naturally dissipated and nowadays even the acronym IGOR2 is used in algorithm design for Cabell-based software for "inductive synthesis of functional programs," not molecules.[39]

Another notable effort from the 1990s is the WODCA program developed by Johann Gasteiger's group.[40] Akin to IGOR, this approach breaks away from the dogma of synthon-based retrosynthetic planning and functional-group approaches. Instead, it focuses on the fundamental properties of bonds (e.g., polarity, inductive effects, resonance, polarizability) to suggest which bonds are suitable for retrosynthetic disconnections. Another difference is that it allows the user bi-directional analysis in which common substrates stored in computer's memory can be matched onto the target to suggest routes directing the chemist to these targets. Since the program relies on matrix notation, the analyses of molecules are necessarily slower than with alphanumeric representation such as SMILES. This, however, does not beat WODCA's objective as it is not per se a tool to automatically design syntheses but rather to assist the chemist in synthetic planning.

The CHIRON[9] program was developed by Stephen Hanessian at the University of Montreal and also uses the idea of mapping available substrates onto the user-specified targets, directing the user towards syntheses that maximize the overlap. The distinctive feature of this approach is that it takes care of stereochemistry during mapping. On the other hand, the program is not searching for complete retrosynthetic trees and can therefore be classified as an interactive tool whose purpose, like WODCA's, is to aid a human chemist in synthetic planning. The last paper on CHIRON was published in 2005.[9b]

Finally, the idea of using similarity underlies the ARChem Route Designer[41] developed by SymBioSys. This approach departs drastically from the concept of expert-coded reactions and instead relies on reaction transformations/reaction "rules" (close to 100 000) machine-extracted from similar literature examples (though it also supplies a set of around 50 hand-generated rules). The program explores relatively short reaction trees exhaustively but does not account for stereochemistry and/or regiochemistry. In a similar genre, the InfoChem's IC_{SYNTH} relies on the reaction cores extracted from various databases[42a] which are then used for construction of synthetic suggestion trees under user control. While the suggestions are based on "analogous" reactions performed on different compounds (cf. Section 3.2.2), they can complement the intuition of a practicing chemist thus serving as a synthetic "idea generator".[42b] Unlike many other programs described earlier in this section, both ARChem and IC_{SYNTH} are commercially available.

3.2. Failure Analysis and Key Challenges

Despite numerous attempts, none of the approaches discussed in the previous section seems to have delivered

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

a software that would be relevant to everyday practice of expert organic chemists. Perhaps the challenge has been tackled too early, when the computers were in their infancy and many of the requisite algorithms were simply not yet properly developed. Whatever the reasons might have been, it is quite unfortunate that while computers revolutionized so many fields of research starting in the 2000s, chemistry largely abandoned the problems of synthetic planning—after all, if Corey failed, who should even try? Part of the problem might have been the oversimplification of the challenge in the early approaches. The insufficient computational power of the 1970s and 1980s required the researchers to take "shortcuts" by imposing various types of synthetic heuristics and simplifications—with the hindsight of today's knowledge, we know that certain complex types of computational problems cannot be simplified too much. A good example is Deep Blue and other chess playing programs that beat the human champions not by using few heuristic rules but by exhaustive searches of all viable options. In the same spirit, synthetic planning cannot be done by teaching computer few hundreds of general rules or working by analogy to literature-reported reactions. Computer has to be taught an enormous number of precise chemical rules, trained how to use them, and be able to explore billions of synthetic options before its true power manifests itself. Let us first reexamine these and some other factors that make synthetic planning such a formidable problem.

3.2.1. The Importance of "Sparse Events"

Like Polish grammar, organic synthesis is full of exceptions. This statement is quantified by the plot in the left portion of Figure 11 in which we extracted unique reaction types/cores (i.e., substructures changing in the reactions) from over 1200000 literature-reported reactions, ranked them according to how many times they were observed (i.e., most popular reaction type was ranked #1, second most popular #2, and so on) and then constructed a frequency versus rank plot. Note that the plot is linear on the doubly-logarithmic scale. Similar power-laws are also observed on "smaller-scales," for specialized branches of chemistry like formation of aromatic heterocycles (90000 reactions, right portion of Figure 11). The point of these analyses is that that presence of a power-law indicates the importance of the low-occurrence, "black swan" events in the process underlying the distribution. What this means for us is that in chemistry some relatively rare/specialized reactions can be of crucial importance in a particular synthesis (see examples in SI, Section S7). While the problem of synthetic black-swans pertains mostly to complex or "exotic" targets we must be able to deal with such cases if our goal is to construct a truly expert system (as opposed to a "toy" dealing only with simple molecules). Unfortunately, this means that we need to teach the computer not hundreds general transforms but probably tens of thousands, including the specialized cases.

3.2.2. No Easy Automation

It would be very tempting to extract these tens of thousands of reaction types automatically from databases of known/published reactions. In fact, this is how we initially approached the problem and this is what some other software packages use as their knowledge-base.[41,42] In such approaches, the computer 1) extracts the group of atoms/bonds that change in every reaction stored in the database and 2) possibly adds some predetermined number of flanking atoms to this "core" to define a unique synthetic transform. Unfortunately, there are major problems with such automatic extraction. Say we extracted the Friedel–Crafts reaction motif/core spanning an aromatic carbon to which an alkyl or acyl moiety is being attached—the key problem in using such a transform in subsequent synthetic planning is that it does not account for the effects of other substituents (possibly present on the aromatic ring) that do not per se participate in this aromatic substitution reaction but, as we know, dictate its outcome. One can try to extend the "core" few atoms "to the left" and/or "to the right" from the reaction center, but this is always arbitrary and indeed futile given all types of possible substitution patterns, substituent types, or aromatic systems on which Friedel–Crafts reaction can be performed. This is just one relatively trivial example and there are many more—summarized in the SI, Table S1 in Section S8—where the automatic extraction suffers from problems ranging from simple errors in underlying database entries, to the inability to account for steric and/or electronic effects, reactivity conflicts, and reactions'

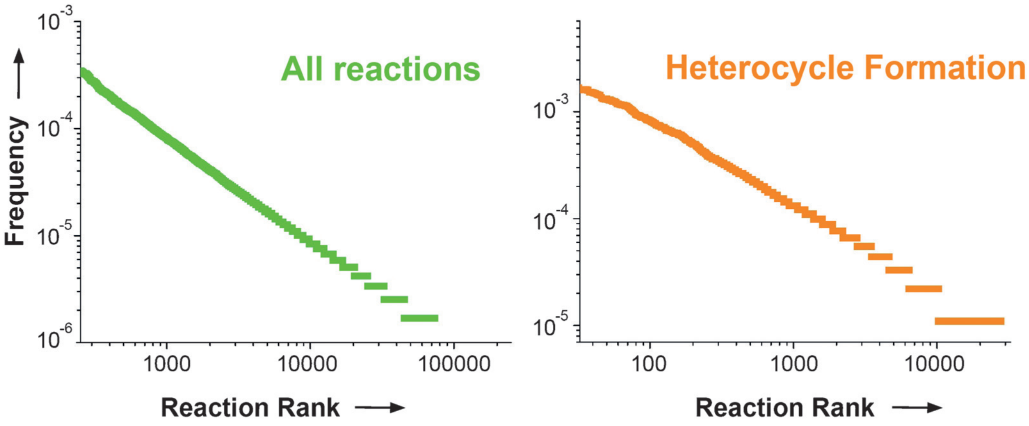


Figure 11. The frequency–rank plots of distinct reaction types. The left plot is based on the analysis of 1.2 million literature-reported reactions randomly chosen from the NOC. The right plot is for the reactions forming aromatic heterocycles. In both cases, the distributions are power laws (i.e., linear on a doubly-logarithmic scale) indicating the relative importance of reactions that occur infrequently. Reaction rank = 1 indicates the most popular reaction, 2 is for the second-most popular, etc.

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

stereo- or regiochemistry (cf. Section 3.2.4). To do things right, the reactions must be coded by human experts carefully delineating which substituents are or are not allowed, and considering both steric and electronic factors, and more. This expert-based approach is actually not an exception in teaching computers to solve complex problems—indeed, Deep Blue was able to score chess positions because it was "taught" an incredible number, 700000, of grandmaster games; Mathematica began to do its wonders of symbolic mathematics only after it has been "taught" by humans a certain number of rules, heuristics and algorithms, some of which took years to develop and volumes to describe (e.g., Risch's algorithm for indefinite integration alone took over 100 pages of Ref. [43]).

3.2.3. The Importance of Molecular Context

Perhaps the most important reason why retrosynthetic planning is so challenging is that it is not enough to establish whether a given bond can be cut (all *individual* bond types can somehow be cut) but rather whether it can be cut in a specific molecule. No matter what the rules for bond disconnection might be—Corey's heuristics of strategic disconnections, Gasteriger's choices based of bond properties, preference of bonds having maximum information content,[44] etc.—they should come with additional information of molecular *context*: in particular, which other groups in the reacting molecule(s) need to be protected during a putative reaction, and which groups present unsurmountable reactivity conflicts. Figure 12 has one simple example whereby the synthesis of ketones from Weinreb–Nahm amides can proceed with the substrate shown in Figure 12a but not with the one shown in Figure 12b (in which a more cross-reactive aldehyde group is present yielding an alcohol rather than a ketone product). Noting that there are a myriad of other similar examples, the general way to express the retrosynthetic rules is therefore in the form of conditionals "*reaction X can proceed if groups Y,Z are not present*" or "*reaction X can proceed if groups Y,Z are appropriately protected*". We note the importance of context is analogous to a language whereby the same words have different meanings depending on the rest of the sentence—indeed, we recently showed that organic synthesis and linguistics share some formal analogies (see Ref. [44] for details).

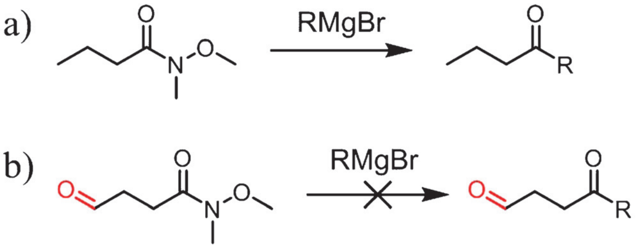


Figure 12. The importance of molecular context illustrated for the synthesis of ketones from Weinreb–Nahm amides. In example a) Grignard reaction leads directly to a ketone. Usage of Weinreb–Nahm amides allows for avoiding over-addition. b) Due to the presence of a more reactive (aldehyde) group, the same Grignard reaction will yield an alcohol rather than a ketone.

3.2.4. Accounting for Stereochemistry and Regiochemistry

In the survey of retrosynthetic programs in Section 3.1 we have seen that the vast majority of them do not consider stereochemistry or regiochemistry. This is not an accidental omission but rather an inherent problem with data structures. For example, in matrix notation it is easy to code the connectivity of the atoms but not absolute configurations. The problem is present even in the modern SMILES[4]/SMARTS[45] notation, where the configurations can be assigned easily within individual molecules but keeping track how they change during reaction is often problematic.

3.2.5. Lack of Well-Defined Synthetic "Positions"

A feature that enabled chess programs to become so successful is that it is possible to make predictions about the game's outcome on the basis of the current position defined by the arrangement of the pieces on the chessboard. In synthesis, the "position" is ill-defined although we intuitively feel it must have something to do with the complexity of substrates generated in a given step. Unlike in chess, however, where history of prior moves does not matter, synthetic positions also carry with them a "cost" associated with the number of steps already performed. A preferable scenario is to get to the "position" of simple substrates in the least possible number of synthetic moves.

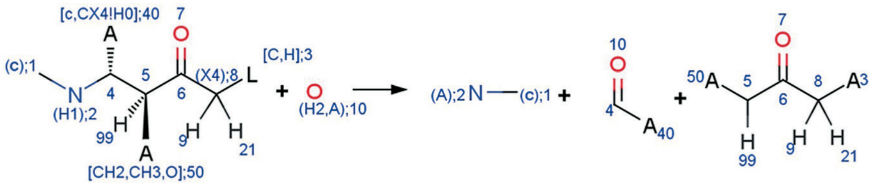
3.2.6. Size of the Search Space and Lack of Intelligent Algorithms

The encouraging news is that the number of possibilities to consider in retrosynthetic planning—while still very large, on the order of 10^{30}–10^{50} for long sequences (Table 1)—is much lower than the number of potential chess games (ca. 10^{230} for 80-move games, Table 1). Since computers somehow play chess it is then reasonable to assume that we should be able to teach them how to solve the smaller-scale synthesis problems. However, one of the major roadblocks to date has been the lack of suitable algorithms with which to explore the vast space of synthetic possibilities. Most programs from Section 3.1 that attempted any retrosynthetic-tree expansion at all, did so in an exhaustive fashion or via basic BFS-type searches, both of which are unfeasible for such an enormous search space. To emulate human synthetic logic, the algorithm should not simply move "forward" but be able to revert from unpromising pathways, explore local alternatives, and if these fail, switch to completely new strategies (as we will, indeed, see in Section 3.4.3).

3.3. Syntaurus

Over the past decade, our group has been developing software that addresses most of the abovementioned challenges. The program that emerged from this effort is called

Syntaurus, a conjunction of *syn*thesis and *taurus*, Latin for a bull.



```
rxn_id: 8382,
name: "Proline-catalyzed Mannich Reaction",
reaction_SMARTS:[c:1][NH:2][C@H:4]([c,CX4!H0:40])[C@:5]([#1:99])([CH2,CH3,O:50])[C:6]
(=[O:7])[CX4:8]([#1:9])([#1:21])[#6,#1:3].[OH2:10]>>[c:1][N:2].[*:40][C:4]=[O:10].[*:50][C:5]([
#1:99])[C:6](=[O:7])[C:8]([#1:9])([#1:21])[*:3]"
products:["[c][NH][C@H]([c,CX4!H0])[C@]([#1])([CH2,CH3,O])[C](=[O])[CX4]([#1])([#1])[#
6,#1]", "[OH2]"]
groups to protect: ["[#6][CH]=O", "[CX4,c][NH2]", "[CX4,c][NH][CX4,c]", "[#6]C([#6])=O"]
protection_conditions_code: ["NNB1", "EA12"]
incompatible_groups: ["[#6]O[OH]", "c[N+]#[N]", "[NX2]=[NX2]", "[#6]OO[#6]",
"[#6]C(=[O])OC(=[O])[#6]", "[#6]N=C=[O,S]", "[#6][N+]#[C-]", "[#6]C(=O)[Cl,Br,I]",
"[CX3]=[NX2][*!O]", "[#6]C(=[SX1])[#6]", "[#6][CH]=[SX1]", "[#6][SX3](=O)[OH]",
"[CX4]1[O,N][CX4]1", "[#6]=[N+]=[N-]","[CX3]=[NX2][O]"]
typical reaction conditions: "(S)-proline. Solvent, e.g., DMSO",
general references: "DOI: 10.1021/ja001923x or DOI: 10.1021/cr0684016 or DOI:
10.1021/ja0174231 or DOI: 10.1016/S0040-4020(02)00516-1"
```

Figure 13. Proline-catalyzed Mannich reaction as coded into Syntaurus. The program's knowledge base comprises ca. 20000 reaction expert-coded records like this. For this specific example, SMARTS sub-record specifies, for example, that atom numbered 40 can be an aromatic carbon ("c") or an alkyl group with at last one hydrogen atom present (coded as "CX4!H0"). This constraint excludes bulky/branched substituents[53a] known to to give lower yields and enantioselectivities. Furthermore, the list of admissible substituents at position numbered 50 includes aliphatic carbon bearing two or three hydrogens or an oxygen atom (list coded as "[CH2, CH3,O]"). Substituents on atom numbered as 8 (with two hydrogens explicitly coded) limit possible outcomes to primary or methyl carbons. "@" signs at carbons 4 and 5 allow to maintain (with operations described in Section 3.4.3) actual reaction stereochemistry limiting reaction products to one of the possible *syn* diastereomers. The way the transform is coded also preserves reaction regiochemistry while placing oxygen (e.g., in hydroxyacetone) as a substituent on carbon 5, which is consistent with experimental results.[53b] Finally, because all known proline-catalyzed direct Mannich reactions use only aromatic amines, the substituent on the amine moiety is described as aromatic carbon (coded as "c"). In addition to the above information, the reaction record includes the list of groups that, if present on any of the substrates, must be protected, the conditions code specifying optimal protecting groups (if protection is needed), the list of groups that are incompatible/cross-reactive with the reaction, the typical reaction conditions (these might be fine-tuned in specific syntheses), and some illustrative literature references describing this type of chemistry.

At the heart of Syntaurus is, of course, a collection of reaction transforms that are applied during retrosynthesis of desired molecules. Initially, we relied on transforms machine-extracted from literature precedents and categorized into ca. 115000 unique reaction classes. While this method of knowledge-base creation proved straightforward (the entire extraction from the NOC took only few weeks), it unfortunately led to utterly meaningless synthetic predictions—as we have discussed in Section 3.2.2 and Table S1 (see SI, Section S8), the fundamental reason for this failure is that machine-extraction assumes that synthetic planning can be done "by analogy," even though the original literature-reported precedent and any given synthetic target of current interest may have very different "contexts" (i.e., protection requirements, group incompatibilities, etc.; cf. Section 3.2.3). This problem might be negligible in some simpler molecules, but in general it is a fatal flaw of the automated extraction approach which cannot replace expert chemical knowledge—this, however, means that thousands upon thousands of reaction types have to be hand-coded by experts and carefully tested for errors. That is why creation of Syntaurus' knowledge base took so many years ultimately resulting in ca. 20000 expert-coded (and expert-tested) transforms, ranging from simple S_N2 all the way to complex rearrangements and cascade reactions. A typical entry in our database is illustrated in Figure 13 and contains the reaction motif (here, proline-catalyzed Mannich reaction) coded in the SMILES/SMARTS notation along with the list of groups to protect, list of groups that are incompatible with a given reaction, the typical/suggested reaction conditions, and some references which describe this type of chemistry. For this and other chemical "rules" to yield meaningful results during synthetic planning, all atom/substituent types must be carefully considered and defined such that stereochemistry, regiochemistry, as well as steric and electronic effects are properly accounted for (also see examples in the SI, Section 9).

The reaction rules are coded in the SMILES/SMARTS format which in recent years has emerged as one of the standard chemical notations and represents the molecules and reactions as alphanumeric strings. This is of critical importance in large-scale computational problems as operations on strings are much faster than on matrices which underlie, for example, .mol files (cf. SI, Section S10). There are, however, two major problems with SMILES/SMARTS. The first one deals with stereochemistry, which in individual molecules is specified by symbols @ and @@. In simple reactions and using software like RDKit,[45b] the stereochemistry of reactions coded with all atom mappings can usually be ascribed correctly (e.g., @ changing into @@ or @@ changing into @ denote configuration inversion). However, for more complex reactions involving multiple stereocenters (especially, proximal ones), there have been no satisfactory algorithms. The second problem is similar and deals with symbols // and /\ specifying regiochemistry of double bonds in individual molecules—unfortunately, there have been no methods to keep track of these symbols over reactions' SMARTS and to ascribe proper reaction regiochemistry. To overcome these problems, we developed a software module which passes between the retrons and the synthons not only the (@, @@) and/or (//,/\) information, but also appropriately ordered (by the masses of substituents) lists of bonds neighboring each atom mapped in the transform; if the specific atom is involved in the bond making or breaking, the ordering of its neighboring bonds is generally different in the retron and in the synthon. While constructing the lists, it is important to add any missing hydrogens, which aids proper ordering by avoiding ambiguity (in corner cases, next-nearest bonds might need to be taken into account). Ultimately, upon the execution of a transform, its stereo/regiochemistry is determined by the consensus of the bond list order and the stereo/regiochemistry symbols present in the SMARTS notation.

It should be noted, however, that for some classes of reactions, even detailed specification of the reaction "core"/motif is still insufficient to predict where the reaction will occur in a given molecule. The case in point here are aromatic substitutions for which "distant" substituents present on the aromatic system can have a decisive influence on the reactivity of other sites. While for simple benzene system one could, potentially, enumerate all possible *ortho/meta/para* substitution patterns, the number of possibilities to consider for other aromatic systems (fused rings, heterocyclic aromatics) is impossible to account for. On the other hand, if one knows whether the substitution is electrophilic/nucleophilic, it is possible to determine the reactivity of each atom based on its high/low electron density or electron delocalization energy. Syntaurus calculates on the fly the per-atom delocalization energies (see Ref. [46] and SI, Section S11) of aromatic systems and then uses these values to determine where and which substitutions are allowed (electrophilic and nucleophilic substitutions are allowed if delocalization energy is, respectively, below and above certain thresholds).

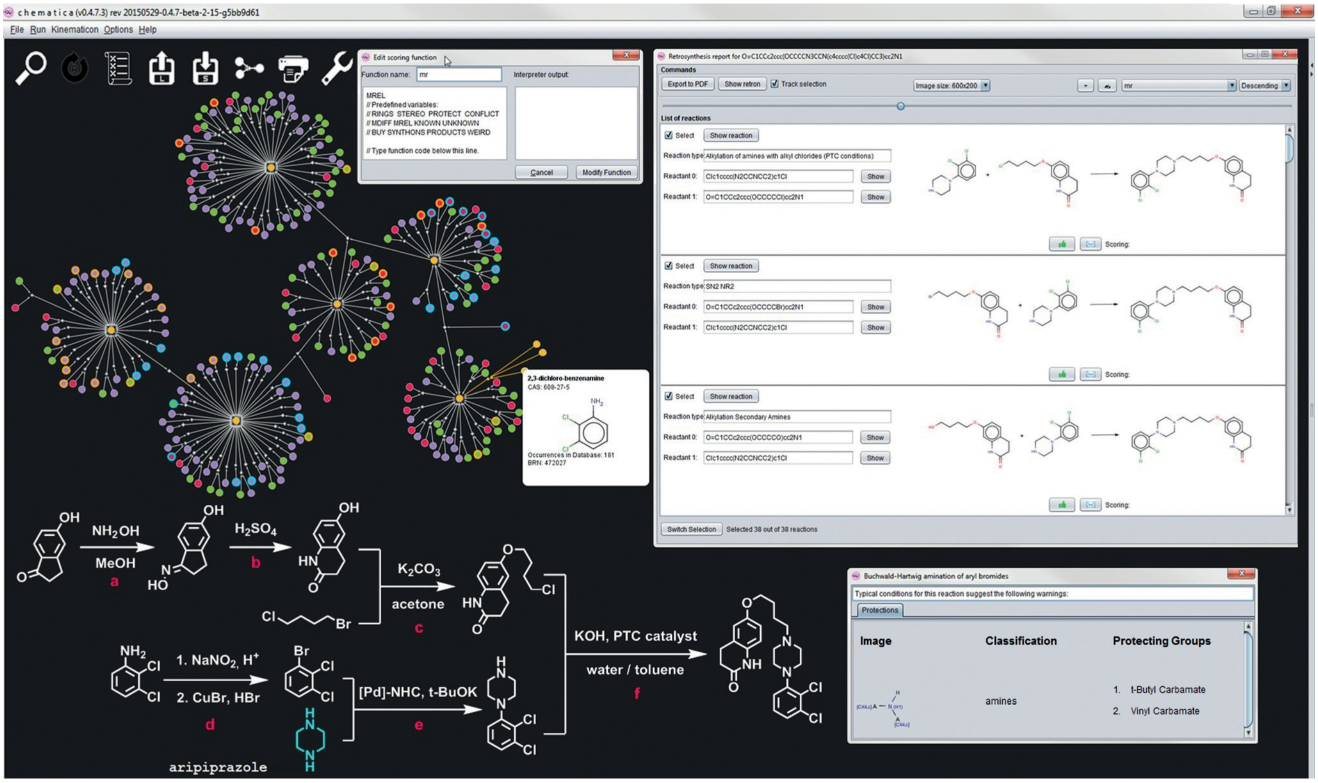
Finally, a database of few thousand "impossible" molecular fragments (e.g., those violating Bredt's rules) is used to avoid any structurally nonsensical outcomes on scales larger than the cores of individual transforms (cf. SI, Table S2 in Section S12 for a small selection).

3.3.1. *Step-by-Step Synthetic Planning in General and in Syntaurus*

The simplest mode of synthetic planning is one in which the user makes choices at each synthetic step. To guide these choices, several prior synthetic-planning programs have used various types of heuristics ranging from strategic bond disconnections in LHASA, to maximization of structural overlap between substrates and the target (CHIRON), to maximization of yields of "similar" literature-reported reaction in ARChem. However, it is highly unlikely that any single heuristic will be universally applicable—for instance, for polycyclic targets one would probably favor disconnections creating new rings while for targets having multiple stereocenters the premium should likely be on stereoselective syntheses. In Syntaurus, we created a script-like language that uses predefined variables to evaluate synthetic steps. For example, variable RINGS defines how many rings are created in a given reaction, STEREO specifies how many stereocenters are created, MREL gives favorable scores to reactions that lead to substrates of similar masses ("cut into equal fragments"), BUY promotes substrates that are commercially available, CONFLICT and PROTECT penalize reactions in which, respectively, group incompatibilities or the need for protection chemistry are detected. From these and some other variables (cf. SI, Section S13), the user can define arbitrary expressions ("scoring functions") that can be used to rank synthetic options.

For example, in Figure 14 (and Movie S3), a simple function MREL (promoting equal-size cuts at each step) was applied to guide retrosynthesis of aripiprazole, a second generation antipsychotic drug. A short and highly convergent synthetic pathway was efficiently found that starts from commercially available 5-hydroxyindanone, 1-bromo-4-chlorobutane, piperazine and 2,3-dichloroaniline. The key retrosynthetic steps are the Beckmann rearrangement guiding preparation of the first building block, amination of aryl bromide, and alkylations of O- and N-nucleophiles. The building block containing a lactam moiety is prepared starting from commercially available 5-hydroxyindanone. Treatment with hydroxylamine chloride (step a) gives an oxime which undergoes Beckmann rearrangement in step b leading to 3,4-dihydroquinolinone (alternatively, this transformation can be carried out under Schmidt's conditions but protection of phenolic oxygen is necessary). Chemoselective alkylation with 1,4-bromochlorobutane then follows in step c. The second, *N*-aryl piperazine building block is prepared from commercially available piperazine and 2,3-dichloroaniline. The latter is transformed (step d) into an appropriate aryl bromide via sequential diazotization/bromination (Sandmeyer reaction), followed by coupling (step e) with piperazine under Buchwald–Hartwig conditions (Pd-NHC catalyst, t-BuOK). For this step, the program correctly detects that one of the nitrogens on the piperazine substrate (colored blue) can present a synthetic problem which (other than adjusting the molar feed ratio or/and careful optimization of conditions) can be avoided by protecting one of the amine groups—the list of the protection groups most suitable for this particular reaction type/conditions is suggested by the program and is displayed in the sub-window in the bottom-right portion of Figure 14. Assuming the protection problem is addressed, the last synthetic step (f) is the alkylation of *N*-arylpiperazine. All along, the design of this synthetic strategy was supported by the ranking windows such as the one shown in the right part of Figure 14 whereby all reactions were scored according to the aforementioned user-specified MREL function.

In the cases of simple molecules like aripiprazol, it is fully conceivable that a competent organic chemist would suggest pathways similar to Syntaurus' without any planning software—though likely not within minutes and perhaps with some difficulty in identifying steps like the Beckmann rearrangement. Still, a well-trained synthetic chemist *could* do it. A more interesting challenge for the computer is in the syntheses that present a challenge even to human masters. In Figure 15 we illustrate one such example, suggested to us by Prof. Dirk Trauner, of an *Epicoccum nigrum* metabolite called epicolactone, which was isolated only in 2012[47a] and for which only a plausible biosynthetic pathway (though not



Edit scoring function

Function name: mr

Interpreter output:

```
MREL
// Predefined variables:
// RINGS STEREO PROTECT CONFLICT
// MDIFF MREL KNOWN UNKNOWN
// BUY SYNTHONS PRODUCTS WEIRD

// Type function code below this line.
```

Cancel | Modify Function

Retrosynthesis report for O=C1CCc2ccc(OCCCCN3CCN(c4cccc(Cl)c4Cl)CC3)cc2N1

Commands

Export to PDF | Show retron | ☑ Track selection | Image size: 600x200 | mr | Descending

List of reactions

☑ Select | Show reaction
Reaction type: Alkylation of amines with alkyl chlorides (PTC conditions)
Reactant 0: Clc1cccc(N2CCNCC2)c1Cl | Show
Reactant 1: O=C1CCc2ccc(OCCCCCl)cc2N1 | Show
Scoring:

☑ Select | Show reaction
Reaction type: SN2 NR2
Reactant 0: O=C1CCc2ccc(OCCCCBr)cc2N1 | Show
Reactant 1: Clc1cccc(N2CCNCC2)c1Cl | Show
Scoring:

☑ Select | Show reaction
Reaction type: Alkylation Secondary Amines
Reactant 0: O=C1CCc2ccc(OCCCCO)cc2N1 | Show
Reactant 1: Clc1cccc(N2CCNCC2)c1Cl | Show
Scoring:

Switch Selection | Selected 38 out of 38 reactions

Buchwald-Hartwig amination of aryl bromides

Typical conditions for this reaction suggest the following warnings:

Protections

Image	Classification	Protecting Groups
	amines	1. t-Butyl Carbamate 2. Vinyl Carbamate

Figure 14. Synthesis design of aripiprazole (sold as Abilify, an antipsychotic drug) in Syntaurus guided by the MREL scoring function. The upper-left portion of the Figure has the raw Syntaurus output of reaction "spiders" representing synthetic possibilities at each step (see also Movie S3). Violet nodes denote unknown molecules, green nodes stand for known molecules (i.e., those whose synthesis has been described and deposited in the NOC), red nodes will denote commercially available chemicals, blue halos signal need for protection, and orange halos denote serious group-reactivity conflicts. The summary scheme of the steps selected is shown below the "spiders". All reactions at a given step can be displayed in the form of a list like the one in the upper-right portion of the figure. In the window shown here, the reactions are options one step away from the target and are ranked according to the MREL function. Finally, the small sub-window on the bottom-right shows information about groups the program suggests to be used for N-protection on piperazine. The conditions in the reaction scheme on the lower-left are all suggested by the program.

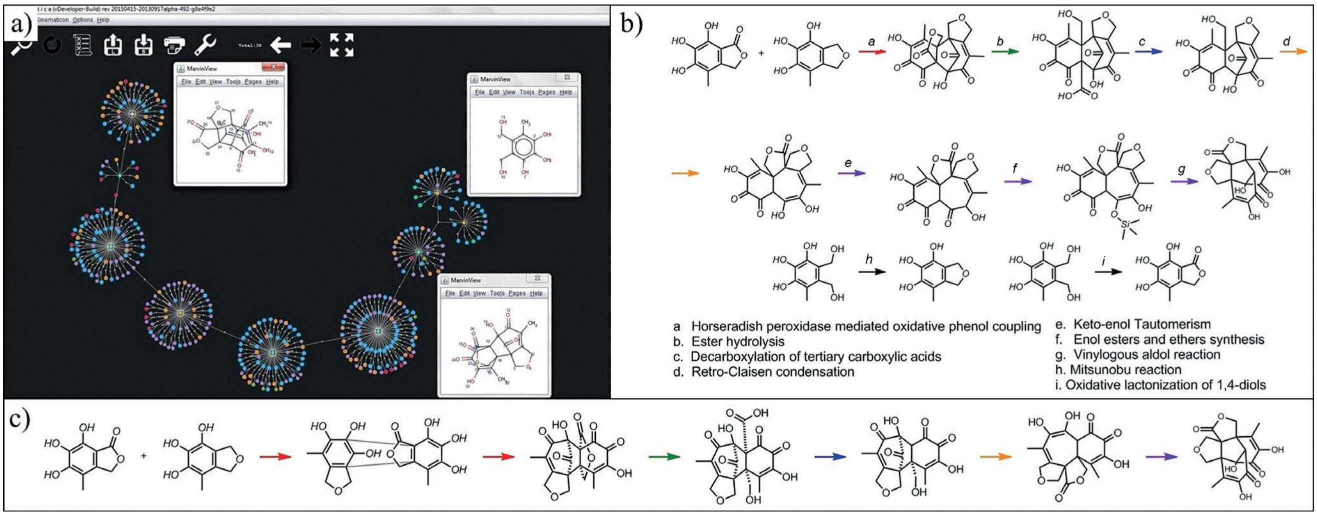

a)
b)
c)
MarvinView
File Edit View Tools Pages Help
HO
OH
O
a
b
c
d
e
f
g
h
i
a Horseradish peroxidase mediated oxidative phenol coupling
b. Ester hydrolysis
c. Decarboxylation of tertiary carboxylic acids
d. Retro-Claisen condensation
e. Keto-enol Tautomerism
f. Enol esters and ethers synthesis
g. Vinylogous aldol reaction
h. Mitsunobu reaction
i. Oxidative lactonization of 1,4-diols

Figure 15. Designing a synthetic route to epicolactone. a) Syntaurus' screenshot of the search tree. Ranking sub-windows guiding the search are included in the SI, Section S14. b) Details of the pathway actually designed with the summary of reaction types employed at each step (we emphasize that all reaction types were coded in a generalized, not target-specific or literature-precedent-based form, see Figure 14). c) A biosynthetic route to epicolactone proposed in Ref. [47b].

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

an actual total synthesis) has been published.[47b] Syntaurus' exploration of epicolactone's synthesis was supported by scoring functions similar to those described in aripiprazol's example, and produced an elegant pathway whose screenshot is shown in Figure 15 a. As before, all of the transforms used were general-scope (i.e., in no way tailored to fit only this specific target) and the full synthesis is shown in Figure 15 b. We observe that the computer suggested choices that took advantage of the target's "symmetry" such that the complex polycyclic structure was effectively built from one (!) starting material. Specifically, in the beginning, main part of the carbon skeleton was created via oxidative coupling of phenolic compounds initiated by hydrogen peroxide (reaction "a") described in literature for polyhydroxylated benzenes.[47c] Subsequent simple transformations (hydrolysis of and ester and decarboxylation, steps denoted "b" and "c") led to the key intermediate. Rearrangement of the skeleton via retro-Claisen condensation[47d] (perhaps the most elegant and hard to see step labelled as "d") provided tautomeric mixture ("e") of enol and ketone of an intermediate activated as a silyl enol ether ("f"). The synthesis was then completed by a well-known vinylogous aldol reaction[47e] (step "g"). The entire computer-assisted analysis leading to the identification of this pathway took just few hours for one of the authors (for more details of the design process, see SI, Section S14). We note that the pathway thus found is closely related to the proposed biosynthesis of epicolactone shown in Figure 15 c. The colors of the arrows in the Figure map the corresponding steps in these two pathways—the only difference in Syntaurus' approach is that it singles out distinct mechanistic steps whereas the proposed biosynthesis combines these steps into "oxidative cascades".

3.4. Automated Searches of Full Pathways

In the examples of the previous Section, computer was able to assist a chemist by rapidly generating and ranking synthetic possibilities at each step. While useful, these capabilities do not address the grand challenge of computer-assisted organic synthesis—that is, fully automated synthetic design of complete pathways. Due to very large number of possibilities on the "expanding" networks of retrosynthetic options, this problem cannot be addressed with any exhaustive types of searches (cf. similar problem encountered in the NOC searches, cf. Section 2). Instead, it requires more intelligent network-exploration algorithms which, in turn, rely on appropriate metrics evaluating and guiding the searches. It is therefore crucially important that we revisit and properly define the concepts of "position" and "scoring functions" applicable to synthetic planning.

3.4.1. Defining "Synthetic Positions"

The fundamental notion of a "position" that enables the chess programs to score current and future arrangements of pieces on the chessboard has so far been missing from computational synthetic design. The previous approaches have largely focused on the disconnection strategies (i.e., "moves") with only intuitive evaluation of synthons' feasibility. It is essential to depart from this paradigm and formally score *both* the reactions and the sets of substrates (i.e., complete molecules not "virtual" synthons) created in each retrosynthetic step; these sets are the "synthetic positions" after each reaction "move". Another feature that distinguishes synthetic planning from chess is that in the former, each performed step adds cost to the overall pathway, whereas in chess it does not really matter in how many steps we got to a given position—all that matters for the ultimate outcome of the game is the current position. In this respect, synthetic planning is more akin to the Rubik's cube where one strives to solve the puzzle in the minimal number of steps (cf. Table 1). It follows that during synthetic planning, computer should score both the sets of substrates generated at each step and the reactions that lead to these sets.

3.4.2. Chemicals' and Reaction Scoring Functions

Following the above logic, synthetic choices the machine makes are to be evaluated by two scoring functions: the Chemicals' Scoring Function (CSF) evaluating the "synthetic positions" (i.e., substrate sets), and the Reaction Scoring Function (RSF) evaluating the "synthetic moves". The sum of these functions, CSF + RSF can be thought of as a measure of the overall difficulty (or "cost") of synthesis and so syntheses minimizing CSF + RSF are sought. In Syntaurus' scripting language, both CSF and RSF can be constructed using predefined variables reflecting structural features of the molecules as well as the features of reactions (e.g., RINGS, STEREO, KNOWN, BUY, PROTECT, CONFLICT; see Section 3.3.1 and SI, Section S13).

1) CSF. The main premise of CSF is that it should favor the simplest possible substrates (such that reactions generating most "complexity" are preferred). The function sums the variables (or constants) characterizing each molecule in the set of substrates. For example, say we define the CSF according to the RINGS variable (number of rings created in a given reaction). Let's assume the target of a particular reaction has two rings and can be made in two reactions from two different sets of substrates. In the first substrate set, there is only one substrate with two rings (CSF = RINGS = 2), whereas in the second set, there are two substrates, one of which has one ring, so that $\mathrm{CSF} = 1 + 0 = 1$. Clearly, according to this CSF's criterion of creating as many rings as possible, the second set of substrates is better (i.e., it has a lower value of CSF). The parameter counting the numbers of stereocenters, STEREO, works analogously. Slightly more advanced in its scope and capabilities is parameter MASS corresponding to the mass of a molecule. Say a reaction cuts the target of molecular weight 400 into two smaller substrates. If CSF = MASS, the score will be independent of where the cut occurs (e.g., $200 + 200 = 300 + 100$). However, if one defines $\mathrm{CSF} = \mathrm{MASS}^2$, the value of CSF will be minimal (i.e., the best) for cutting into halves (e.g., $200^2 + 200^2 < 300^2 + 100^2$). In general, for $\mathrm{CSF} = \mathrm{MASS}^x$, increasing values of x favor cutting into like-sized precursors (which is often good for finding strategic skeletal disconnections), while decreasing x allows for more "peripheral cuts" (like in "decorating" existing skeletons; the

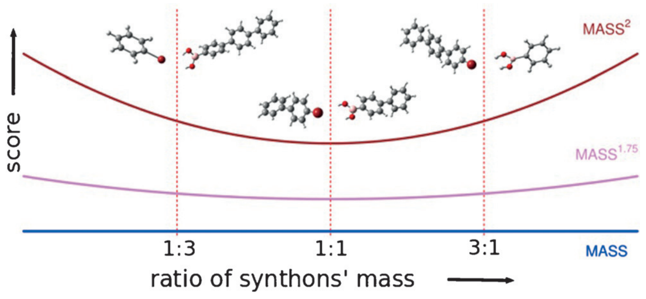


Figure 16. Favoring disconnections into substrates having certain relative sizes by using Chemicals' Scoring Functions involving only one variable. The CSF function here is defined as the sum (over the substrate set) of the masses of substrates, each taken to some power x. Say $x=1$ and CSF $=\sum_{\text{substrates}}$ MASS. Disconnecting the target (here, p-quaterphenyl, MASS ≈ 300) into two equal fragments ($150+150$) will yield the same value of CSF as disconnecting into unequal fragments (e.g., $75+225$ or $225+75$). However, if $x>1$, the function summing MASS^x over the substrates will favor equal-sized cuts—indeed, it can easily be proven that if MASS$_1$ + MASS$_2$ = MASS$_{\text{target}}$ then function MASS$_{\text{substrate1}}{}^x$ + MASS$_{\text{substrate2}}{}^x$ is minimized if MASS$_{\text{substrate1}}$ = MASS$_{\text{substrate2}}$. Because best synthetic positions minimize CSF, such equal-size disconnections will be preferred during synthetic searches. Also please note that higher values of x will give higher CSF values and thus penalize unequal cuts more strongly than lower x exponents (here $x=2$ versus $x=1.75$ curves).

smaller substrates are then often known or commercially available molecules) (Figure 16). Of the remaining variables, SMILES_LEN corresponds to the lengths of molecule's SMILES and is related to its mass as well as overall complexity (since parentheses denoting branching, numbers denoting rings, and @/@@ symbols denoting stereocenters increase SMILES_LEN) whereas variable WEIRD is defined to truncate the searches when very small but "weird-looking" and unknown molecules are encountered (MW < 100, ratio of heteroatoms to carbons > 1.5; if such a small molecule is still unknown in literature, it very likely simply does not exists). Two other useful variables are BUY (+1 if a molecule is commercially available, 0 otherwise) and KNOWN (+1 if a molecule is not commercially available but known in the NOC, 0 otherwise).

2) RSF. The purpose of this function is to favor the shortest possible syntheses and also ones that do not involve serious reactivity conflicts or perhaps need for protection chemistries. In the RSF, one might assign some constant cost of performing a reaction step and a combination of PROTECT (a set penalty for each group that needs to be protected), CONFLICT (penalty for each group incompatibility), or YIELD (theoretically estimated yields)[48] variables. For example, RSF = 30 + 1000·CONFLICT will heavily penalize any reactions in which there are reactivity conflicts whereas RSF = 30 + 1000·PROTECT will heavily penalize reactions requiring protection chemistries.

3) Scoring functions for various types of searches. One of the key advantages of scoring both reactions and the substrate sets is that by setting different values of CSF and RSF one can flexibly fine-tune (or modify) the general search strategy. For instance a simple combination CSF = 0 and RSF = 0 will score all reactions, substrates and pathways equally (at zero) in effect preforming exhaustive searches of the synthetic space. Obviously, this type of a search has no "intelligence" and risks astronomical search times for any but very simple targets. A combination CSF = 0 and RSF = 1 is already much more purposeful. Here, all chemicals have zero score while each individual reaction step "costs" +1—consequently, searches minimizing sums of ones pay no attention to molecules' details and effectively seek shortest synthetic pathways leading to known or commercially available substrates. The search strategy corresponding to this CSF + RSF is similar to classic BFS searches and, given the number of possible synthetic strategies to explore (cf. Table 1), is expected to require very long calculation times. The search times can be drastically reduced by avoiding reactions entailing serious group incompatibilities/conflicts (e.g., by using functions like RSF = 10 + 1000·CONFLICT). Similarly, one can avoid reactions requiring any protection (and construct protection-free, Baran-like pathways) by assigning high weight to the PROTECT variable, e.g., RSF = 20 + 10000·PROTECT.

In the CSF, the weights of the RINGS and STEREO variables should be increased in molecules that have, respectively, many rings or stereocenters; on the flip-side of the coin, these variables should obviously be given zero weight (i.e., be absent from the CSF) if there are no rings/stereocenters in the molecule. In terms of preferring half-and-half versus peripheral disconnections, SMILES_LEN$^{3/2}$ (or MASS$^{3/2}$) is the best compromise and, in general, the exponent should be between 1.2 and 2. As we will see in specific synthetic examples, the most versatile functions (i.e., applicable to the vast majority of targets) are of the form CSF = SMILES_LEN$^{3/2}$ + αRINGS + βSTEREO, where the α and β weights have appreciable weights for targets comprising, respectively, multiple rings, or stereocenters.

3.4.3. "Intelligent" Searches

The RSF and CSF are used to score the searches over the space of available synthetic options. Since in this process, sets of substrates in each step rather than individual molecules are evaluated, the mathematical formulation of the problem becomes more involved and requires a dual-graph representation in which the algorithm traverses a network of substrate sets "overlaid" on the network of specific molecule-to-molecule reactions (Figure 17).

With this representation, we require that the search algorithm be **1) non-local**—that is, able to explore not only one synthetic "branch" of synthetic solutions at a time but consider numerous distinct possibilities simultaneously; **2) strategizing**—that is, able to perform few individual reaction "moves" that might locally appear sub-optimal but could ultimately lead to a "winning" synthetic solution; **3) self-correcting**—that is, able to revert from hopeless branches and to switch to completely different synthetic approaches. In addition, we require that the searches always terminate at either known or commercially available substances (with the threshold molecular weights specified by the user).

The scheme in Figure 18 illustrates how an algorithm meeting these requirements works—its basic features and the

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

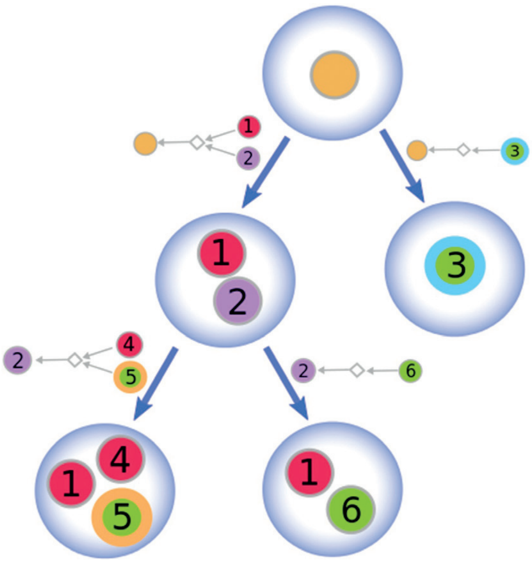


Figure 17. A dual-graph representation underlying Syntaurus' automated searches. The algorithm searches through and scores not individual substrates but sets of substrates ("synthetic positions," here represented by large circles). The relationships between the position nodes constitute one search graph. In addition, since individual molecules rather than "positions" undergo specific reactions, the algorithm must also keep track of the reaction connections between all molecules (here, shown as reaction miniatures next to blue arrows). This creates a second, "overlaid" graph.

use of the priority queue go beyond the specific software and are generally applicable to synthetic planning. In the scheme, nodes represent collections of substrates generated in each reaction "move" (cf. Figure 17), and light gray-nodes are the total space of synthetic possibilities (i.e., "potential" substrate sets, not necessarily visited). The search starts at the central green node denoting our synthetic target (Figure 18 a). All possible "moves" one-step from the target are explored and each is scored (nodes in black halos, numbers correspond to the scores) by the value of CSF + RSF. Of the available possibilities, the algorithm moves to the lowest-score/"cost" node (here, score = 4) and expands it further (Figure 18 b). The other options already scored (nodes with scores 5, 35, 63, 71, 85, 93, 97) are kept in computer's memory as the so-called priority queue (PQ) keeping track of the "current endpoints" of possible pathways. The new move leads to the node with score = 35. We note that this score takes into account the RSF cost of now making *two* steps from the target (and, as before, includes the CSF measure of complexity of the synthons in the current node 35). However, at this point, the cost of performing the reaction along the current pathway (i.e., 35) becomes higher than the score of the other-best synthetic option kept in the PQ—namely, that at the node with score = 5 (Figure 18 c). Accordingly, the search reverts to node with score 4 (blue) but finds no options for "beneficial" local expansion there. Since the local improvement around node 4 did not work, the algorithm 1) keeps node 35 in the PQ (which now

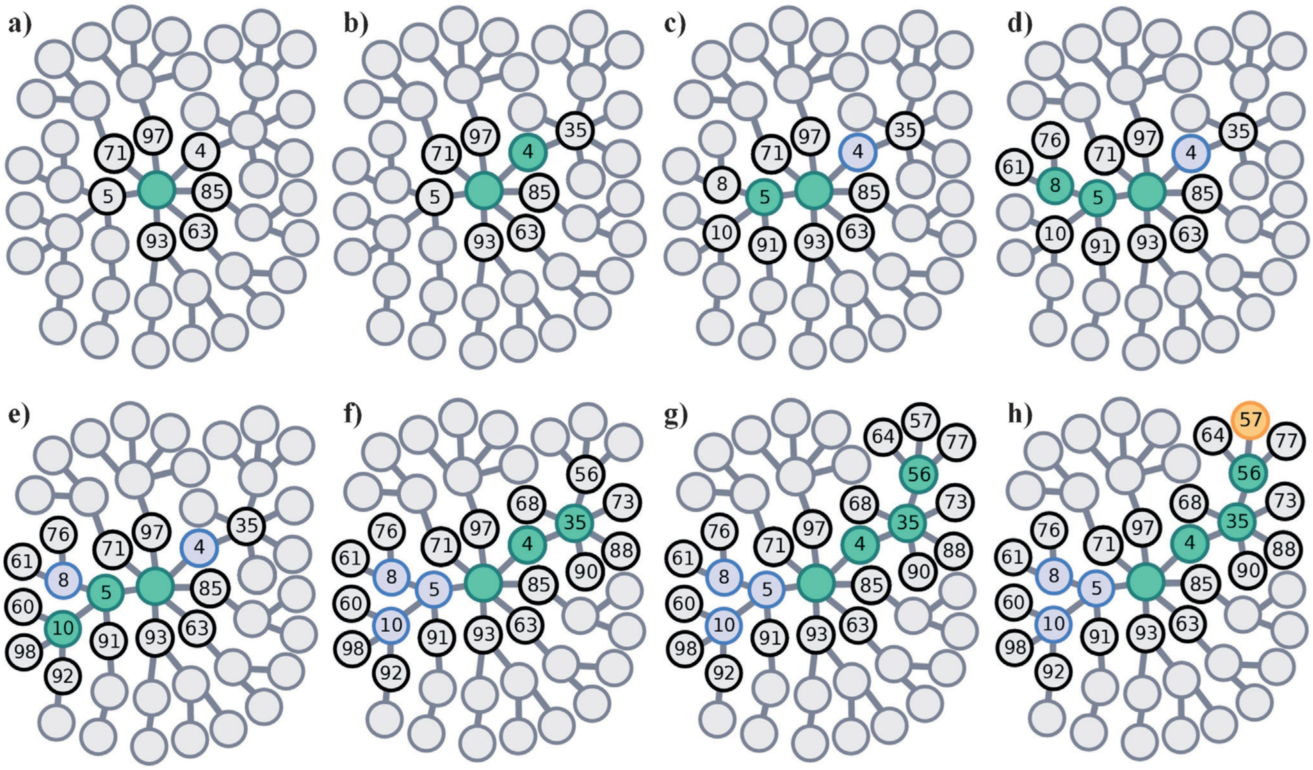


Figure 18. A simplified scheme of Syntaurus' retrosynthetic graph-search algorithm. Each node represents a collection of substrates generated in each reaction "move"; the current description does not explicitly show the "expansion" of individual molecules within each node—such full representation is shown in Figure 17 and is too complex to explain the nuts-and-bolts of how the algorithms works. Light gray nodes are the hypothetical space of synthetic possibilities (i.e., "potential" substrate sets, not necessarily visited). The search continues until all substances in the synthon set are commercially available or known in literature with mass below a certain user-specified threshold.

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

includes nodes with scores 35, 63,71, 85, 93, 97), 2) remembers node 4 as already visited (but not in the PQ since it is not an endpoint of an already-explored pathway), and 3) switches to the completely new pathway with node 5 becoming the best (i.e., lowest-scoring) current option (not included in the PQ anymore). From node 5, the algorithm explores three available options with scores 8, 10 and 91. Of these, it chooses node with score 8 (which is lower than any node in the PQ) and explores into two daughter nodes with scores 61 and 76 (Figure 18d). These, however, are higher in score than the best PQ option (35), so the algorithm reverts "locally" into node 8, finds nothing to explore there, reverts into node 5 (better than 35 in the PQ) and expands into node with score 10 (so far so good) but then encounters three high-score nodes 60, 92, 98 (Figure 18e). Having failed to find any promising syntheses branching from node 5, the algorithm keeps 8 and 10 as already visited, adds 60, 61, 76, 91, 92, 98 to the PQ (which now comprises nodes 35, 60, 61, 63, 71, 76, 85, 91, 92, 93, 97, 98) and moves to the best available PQ option. In this particular example, this best option is the already explored endpoint 35. This endpoint is now expanded into nodes 56, 68, 73, 88, of which the first is better than any entries in the PQ (now 60, 61, 63, 68, 71, 73, 76, 85, 88, 91, 92, 93, 97, 98) (Figure 18f). Accordingly, the algorithm moves to 56 and expands it into options with scores 57, 64 and 77 (Figure 18g). Of these, the node 57 (colored orange) is not only better than other PQ options but finally meets the STOP criterion such that all of its substrates can be bought or are known (Figure 18h). Thus, the first viable synthetic pathway (57→56→35→4→TARGET) is found and scored as the cost of four steps plus the score of the terminal substrates at node 57. The search continues to explore other pathways—in doing so, the algorithm is prohibited into venturing into already identified pathways.

Of course, the above description is quite abstract. For one, we have tacitly assumed that the light gray (not-yet-visited) nodes are there from the very beginning in some form of a static network waiting to be explored (like the NOC). In reality, these available synthetic possibilities form a dynamic network that is not known a priori and expands with every synthetic option explored. Using a specific example, Figures 19a and b show the expanding network (here, with nodes corresponding to substances, not entire substance sets) of possibilities considered by Syntaurus while searching for the syntheses of Donepezil, an anti-Alzheimer drug, using a scoring function favoring substrates of decreasing complexity and strongly prohibiting any cross-reactivity conflicts ($CSF = SMILES\_LEN^{3/2} + SMILES\_LEN$ and $RSF = 10 + 1000 \cdot CONFLICT^2$). The smaller network in Figure 19a is after eight expansions of individual "spiders," and the larger one in Figure 19b is after 35 expansions (ca. 2 min

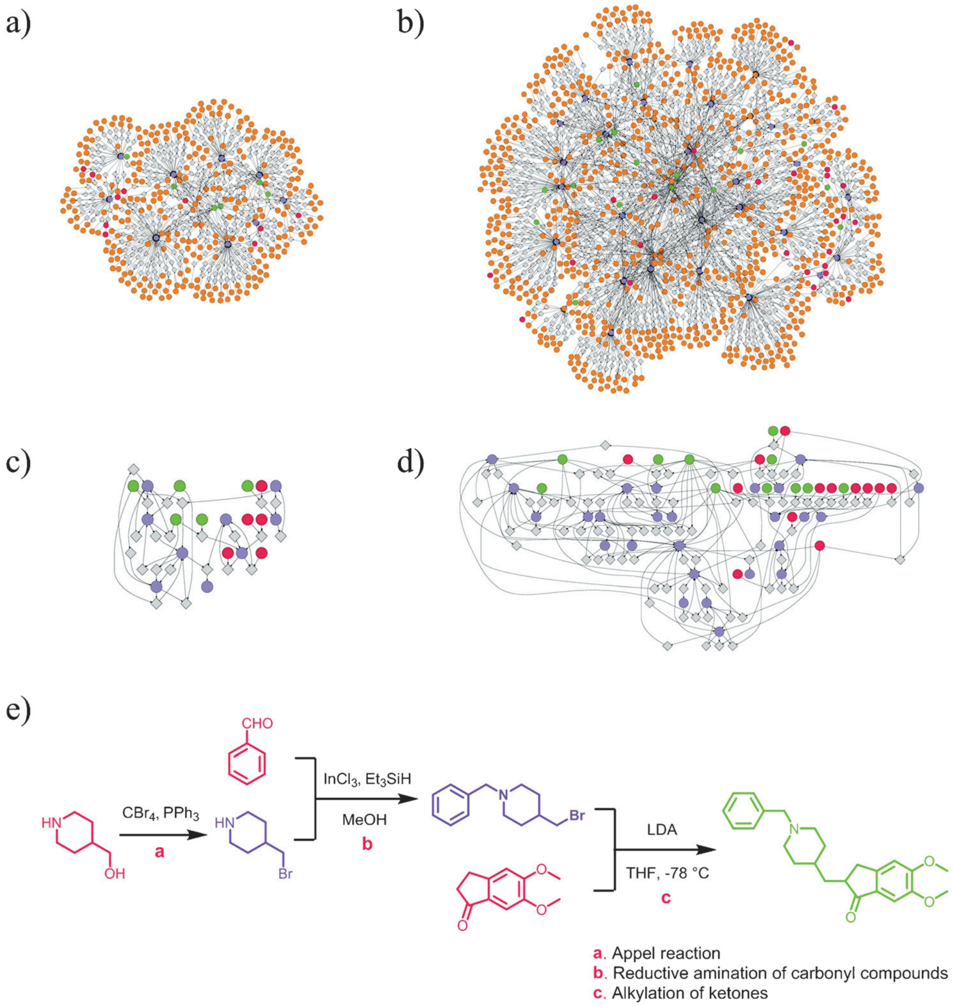


Figure 19. The expanding network of synthetic possibilities considered by Syntaurus while searching for the syntheses of Donepezil (after a) 8 and b) 35 expansions). Circles represent individual molecules: red = commercially available, green = known in the NOC, orange = considered by the algorithms but not yet makeable from commercially available or known precursors, violet = molecules for which a viable synthetic pathways have already been found. In essence, when the first synthesis involving a particular substance is found, the color of the considered node changes from orange to violet. Panels (c) and (d) shows the sub-networks (of (a) and (b), respectively) of the viable syntheses found. e) One of the top-scoring pathways identified. Molecules are colored according to the same scheme as the nodes in the networks (in particular, red = commercially available). The searches were performed using $RSF = 10 + 1000 \cdot CONFLICT^2$ and $CSF = SMILES\_LEN^{3/2} + SMILES\_LEN$.

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

of searching). These networks include all nodes considered by the algorithm (some of which were expanded further and some of which were not, see figure caption)—in contrast, Figures 19 c and d display sub-networks comprising only nodes that are involved in viable syntheses found, ones starting from buyable and/or known substrates. As should be expected, the spectrum of such viable syntheses expands as the algorithm progresses and finds not only more but also better-scoring pathways. One of the top-scoring, concise and chemically reasonable pathways identified is shown in Figure 19 e: After conversion of alcohol to alkyl bromide followed by reductive amination of benzaldehyde, synthesis of the target molecule is accomplished via straightforward alkylation of a ketone.

3.4.4. *Validation and Specific Synthetic Examples*

The ultimate test for every synthetic-planning software is whether it can design viable synthetic plans. One method of validation is to compare program's output against pathways in which all steps had already been experimentally realized. Importantly, this type of validation requires that none of the reaction rules were taught to the machine with the specific target in mind—in other words, the program cannot be tested on the examples on which it was trained. This requirement holds for Syntaurus whose underlying reaction rules are general-scope and are not target-specific (see Section 3.3). The Section S15 (and Movie S4) provides several examples in which Syntaurus' fully automated searches produced—within minutes—pathways leading to relatively non-trivial targets; importantly, all the individual steps in these syntheses have been documented in the literature.

In Figure 20, viable syntheses are designed in similar times but the targets are natural products that have only recently been isolated such that no or only few prior syntheses are available—it is in cases like this that computer-assisted de novo synthetic planning is most useful. For instance, Figure 20 a shows a proposed route of tacamonidine, a natural pentacyclic alkaloid isolated from *Tabernaemontana corymbosa*[49a] and not yet synthesized in the laboratory. First, tryptophol (3-(2-hydroxyethyl)indole) is acylated with succinic anhydride (step a). The program suggests (blue halo on the node) that protection of the free hydroxy group is needed (with methoxymethyl being most suitable for the given reaction conditions). The thus prepared *N*-acylindole derivative is then transformed into acyl chloride under treatment with oxalyl chloride and subsequently acylated under Friedel–Crafts conditions[49b] to give tricyclic intermediate. Further enantioselective alkylation according to Enders' protocol[49c,d] (step c) and reductive amination[49e] (step d) lead to 1,2-*syn*-amine. Iodine-mediated[49f] ring closing (step e) followed by Sharpless asymmetric dihydroxylation leads to the formation of the desired diol, which is then cyclized to give the target molecule.

The example in Figure 20 b is for the synthesis of goniothalesdiol A, isolated from *Goniothalamus amuyon.*[50a] Compounds from this class are popular synthetic targets but syntheses of goniothalesdiol have been reported only recently.[50b–e] Syntaurus identifies the key step (*syn*-selective oxa-Michael addition) and proposes a short synthetic pathway starting from methyl acrylate and but-3-enal. Treatment of these two compounds with Grubbs' II generation catalyst[50f] (alkene metathesis conditions) leads to the formation of methyl (*E*)-5-oxopent-2-enoate. In the following step, the product thus obtained can be reacted with hydroxyacetophenone according to Shibasaki's protocol employing *anti*-selective BINOL-based heterobimetallic catalyst system,[50g] which allows for control over diastereoselectivity and enantioselectivity of the obtained 1,2-diol. Synthesis of the target molecule is then completed in a manner similar to Reddy and Fadnavis' approach by one-pot, Ru-mediated enantioselective reduction/intramolecular oxo-Michael cascade described in literature for the synthesis of centrolobine.[50h]

The third example in Figure 20 c is the synthesis of racemic juvabione (insect juvenile hormone analogue isolated from *Abies lasiocarpa*).[51a] Syntaurus' synthetic pathway begins with conversion of a commercially available, branched unsaturated alcohol into bromide with Br_2/PPh_3 (step a) followed by formation of a Grignard reagent.[51b] Subsequent reaction with a nitrile (step b)[51c] leads to the dienophile needed for the final Diels–Alder reaction (step d; see SI, Section 17 for additional details). Appropriate diene is prepared via methylenation of an unsaturated carboxylic ester (step c).[51d]

Finally, Figure 20 d shows synthesis of a polyhydroxylated natural product isolated from *Cryptocarya latifolia*[52a] and illustrates some nuances that need to be inspected carefully by the human user. The pathway identified by the machine resembles the methodologies described in references.[52b,c] Synthesis of the carbon skeleton is accomplished using two enantioselective aldol reactions (stereocontrol in step a could be achieved using a proline-based catalyst,[52d–f] while proper diastereoselectivity in step c could be ensured using chiral boron enolate[52g] or lithium enolate[52h]). The second aldol reaction seems to be the most challenging step of the proposed pathway; although the use of chiral boron-based auxiliaries was reported to achieve good diastereocontrol, one might anticipate—and consider in detail—that unfavorable 1,5-*anti* selectivity[52i] may influence outcome of this reaction (see Ref. [52j–l]). Acylation of alcohol with acryoyl chloride (step b) and alkene metathesis (step e)[52b] allow for efficient preparation of lactone moiety while the third stereocenter is introduced in step d using Narasaka–Prasad methodology.[52c,j,m]

4. *Challenges and Opportunities*

The synthetic examples of the previous sections suggest that computers are finally capable of designing syntheses relevant to the everyday practice of organic syntheses. At the same time—as we stressed in the title of this Review—it is only "the end of the beginning." and there are still many challenges to be overcome. In this part, we look at these challenges as exciting opportunities for future research and, wherever applicable, highlight the most relevant and promising recent developments.

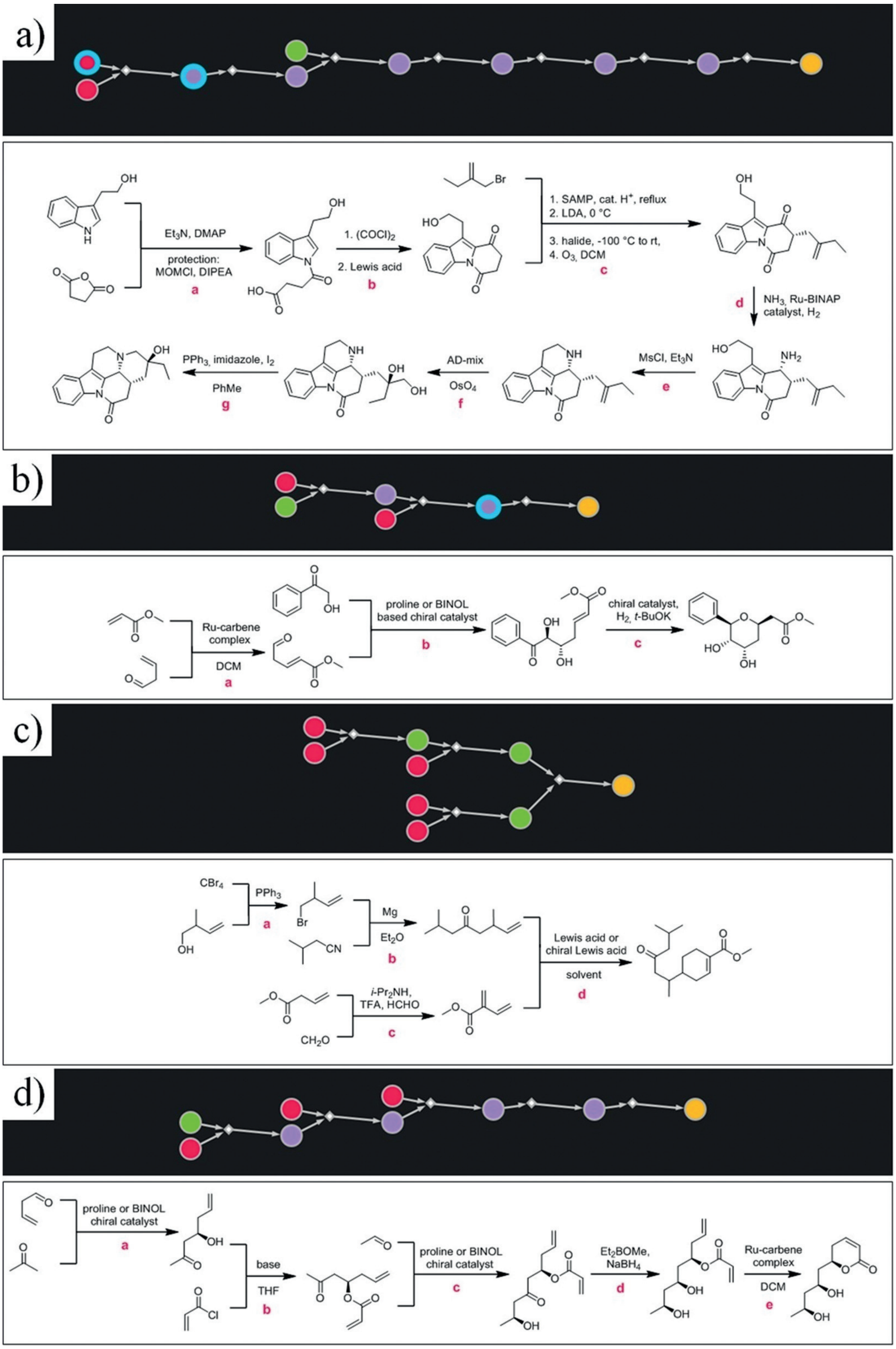

a)
Et$_3$N, DMAP
protection: MOMCl, DIPEA
a
1. (COCl)$_2$
2. Lewis acid
b
1. SAMP, cat. H$^+$, reflux
2. LDA, 0 °C
3. halide, -100 °C to rt,
4. O$_3$, DCM
c
NH$_3$, Ru-BINAP catalyst, H$_2$
d
MsCl, Et$_3$N
e
AD-mix
OsO$_4$
f
PPh$_3$, imidazole, I$_2$
PhMe
g
b)
Ru-carbene complex
DCM
a
proline or BINOL based chiral catalyst
b
chiral catalyst, H$_2$, t-BuOK
c
c)
CBr$_4$
PPh$_3$
a
Mg
Et$_2$O
b
i-Pr$_2$NH, TFA, HCHO
c
CH$_2$O
Lewis acid or chiral Lewis acid
solvent
d
d)
proline or BINOL chiral catalyst
a
base
THF
b
proline or BINOL chiral catalyst
c
Et$_2$BOMe, NaBH$_4$
d
Ru-carbene complex
DCM
e

Figure 20. Syntaurus' synthetic pathways designed automatically and leading to recently isolated natural products. a) Synthesis of tacamonidine[49a] identified using CSF = SMILES_LEN$^{3/2}$ and RSF = 40 + 50·PROTECT + 1000·CONFLICT. Program-suggested "generic" conditions are shown over the reaction arrows. Note: due to the presence of terminal double bond, conditions used for removal of Enders' auxiliary should be modified to avoid oxidation to ketone.[49d] b) Synthesis of goniothalesdiol A[50a] identified using CSF = (RINGS + STEREO)·SMILES_LEN^2 + (RINGS + 10·STEREO) and RSF = 20 + 20·CONFLICT. c) Synthesis of racemic juvabione[51a] found with CSF = SMILES_LEN^2 and RSF = 100 + 5·PROTECT + 10·CONFLICT. d) Synthesis of a polyhydroxylated natural product isolated from *Cryptocarya latifolia*.[52a] The pathway was identified using CSF = 10·RINGS + 10·STEREO and RSF = 5 + 20·PROTECT + 50·CONFLICT. Color coding of nodes: red = commercially available; green = known in the NOC; violet = unknown, yellow = target; blue halos = protection required.

4.1. Streamlining the Searches, Universal Scoring Functions, and "Synthesizability Measures"

Algorithms such as those described in Section 3.4.3 are capable of rudimentary strategizing by choosing synthetic moves that "locally" might look sub-optimal but lead to globally optimal pathways. For instance, in the synthesis of (−)-curvularin in Figure S27 d, the algorithm performs three steps (d–f) which do not seem to offer any obvious immediate gains in terms of building molecular complexity but are important for the overall synthetic strategy leading to the key aryne/ketoester disconnection in step g.

However, for very large networks of synthetic possibilities (e.g., during planning syntheses of large/complex targets) finding multi-step strategies might require excessive numbers of individual back-and-forth probing moves. An idea dating back to LHASA[5b] is that in such cases, the searches can be streamlined by hard-wiring certain common reaction sequences into pre-coded "strategies" (Figure 21 a)—if a reaction corresponding to the first step in a strategy is encountered during synthetic planning, then the program should automatically consider the subsequent strategy step(s). Strategies can indeed shorten the search times. This is illustrated by the example in Figure 21 b, where the use of a strategy comprising Mannich reaction followed by deoxygenation of ketone (cf. fifth row in Figure 21 a) shortened by the factor of three the search time to design synthesis of *N,N*-dimethylbispidine (a scaffold of compounds investigated as drug candidates for treating cardiac arrhythmias and structurally related to cytysine used for tobacco smoking cessation). On the other hand, it must be remembered that introduction of too many strategies may excessively bias the searches into certain branches of synthetic possibilities thus limiting the diversity of pathways generated. Based on our experience, the number of strategies on the order of tens to hundreds seems optimal, but a systematic study would be needed to compare the gains in search speeds versus the restrictions strategies impose on the diversity of synthetic solutions.

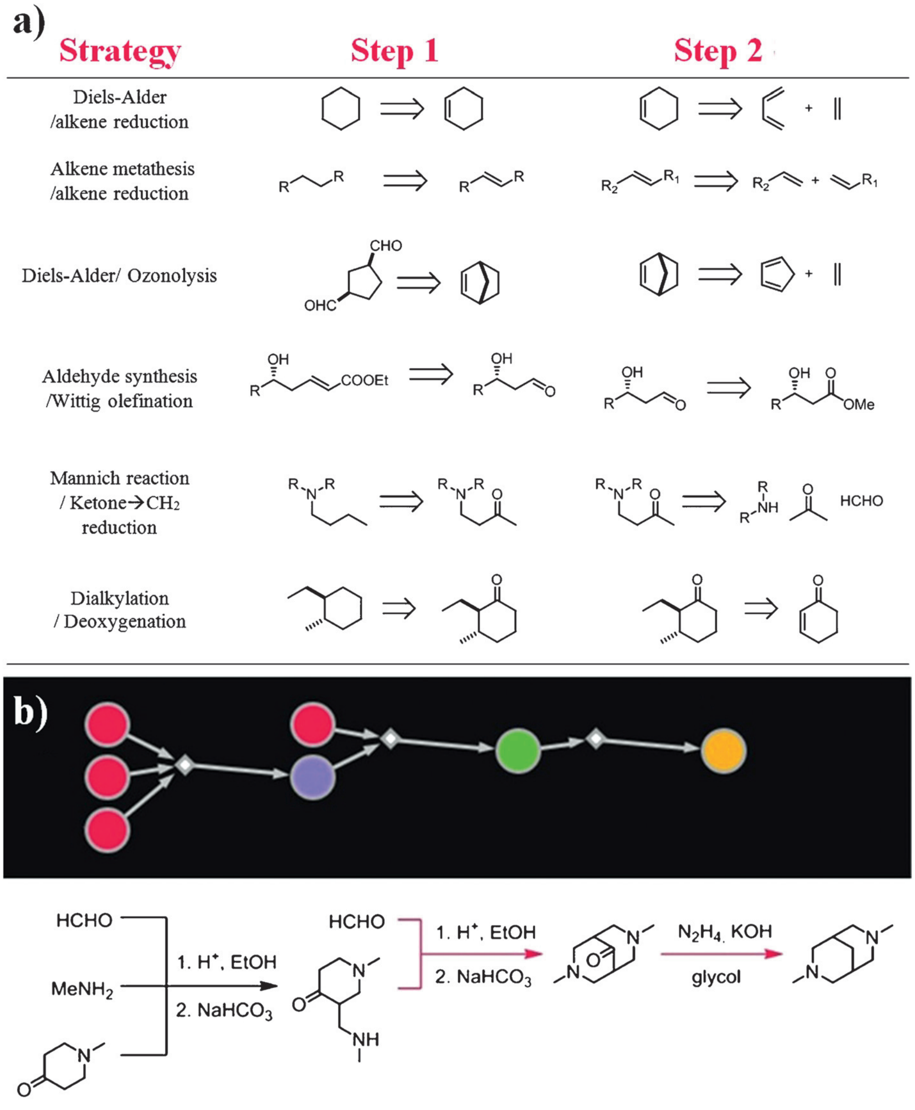

a)
Strategy
Step 1
Step 2
Diels-Alder /alkene reduction
Alkene metathesis /alkene reduction
R
R
R
R
R2
R1
R2
R1
Diels-Alder/ Ozonolysis
CHO
OHC
Aldehyde synthesis /Wittig olefination
OH
COOEt
R
OH
O
OMe
Mannich reaction / Ketone→CH2 reduction
R
N
R
HCHO
NH
Dialkylation / Deoxygenation
O
b)
HCHO
MeNH2
1. H+, EtOH
2. NaHCO3
HCHO
1. H+, EtOH
2. NaHCO3
N2H4, KOH
glycol
N
NH
O

Figure 21. a) Examples of two-step strategies. Steps are numbered in the retrosynthetic direction. b) A Syntaurus-designed synthesis of *N,N*-dimethylbispidine. From the program's perspective, the introduction of a carbonyl group in the first retrosynthetic step does not immediately seem beneficial as it leads to a substrate of higher complexity. Use of one of in-built strategies (here, in the retro direction: "after introduction of a ketone, explore Mannich"; steps denoted in pink, see also fifth row in (a)) reduced the number of search iterations from 151 to 48. All parameters were identical in the searches with and without strategies.

Another question related to search efficiencies is which types of CSF and RSF functions should be used to yield solutions in the shortest times possible. Analyses like the one illustrated in Figure S28 (Section S17) suggest that disconnections into equally-sized fragments should be preferred (though with leeway for unequal-size cuts), and that variables reflecting molecular complexity are important (e.g., length of a molecule's SMILES is performing better than, say, mass of the molecule). The latter criterion connects with the recent progress in defining information-rich metrics of molecular and synthetic complexity. Notably, Li and Eastgate have recently proposed[53a] a complexity index that combines "intrinsic" molecular characteristics (Randic's molecular topological index[53b,c] and the number of heteroatoms on and in aromatic rings) with "extrinsic" measures of synthetic complexity (number of stereogenic centers established during the synthesis, total number of steps, ideality of the route)[53d] that evolve as synthetic knowledge advances. They showed that this composite intrinsic–extrinsic index correlates well with the perception of expert chemists pooled to rank the difficulty of syntheses of several complex targets.

The "intrinsic" measures such as those proposed by Eastgate and also by Gasteiger[53e,f] can be incorporated into the CSF scoring functions guiding the choices the synthetic-planning programs make during pathway design. On the other hand, combination of "intrinsic" and "extrinsic" measures (also see Ref. [53g]) can be useful during evaluation and ranking of complete pathways these programs ultimately identify. In particular, ranking and taking statistics over large numbers (cf. example in Section S18 and also Movie S4) of computer-generated synthetic plans might be more indicative of the synthetic difficulty ("synthesizability") than scoring just a single pathway which might or might not work in the laboratory. Ideally, such rankings should be able to differentiate the "synthesizability" not only of targets having markedly different masses, numbers of stereocenters or numbers of rings, but also those that are structurally similar. This is illustrated by the examples in Figure 22, in which a relatively simple measure of synthesizability (based on the number of steps in the synthesis and the structural complexity of the starting materials) gives significantly different "synthesizability" histograms (i.e., counts of computer-generated syntheses having different scores). One area where such analyses would be of immediate relevance is the screening of "virtual" lead molecules that are nowadays "created" in large quantities[16] by the computational chemistry divisions of virtually all pharmaceutical companies, but are often[54] unsynthesizable (or at least, hard to synthesize). Being able to filter out such molecules can translate into substantial monetary savings and, above all, into faster lead-to-drug times.

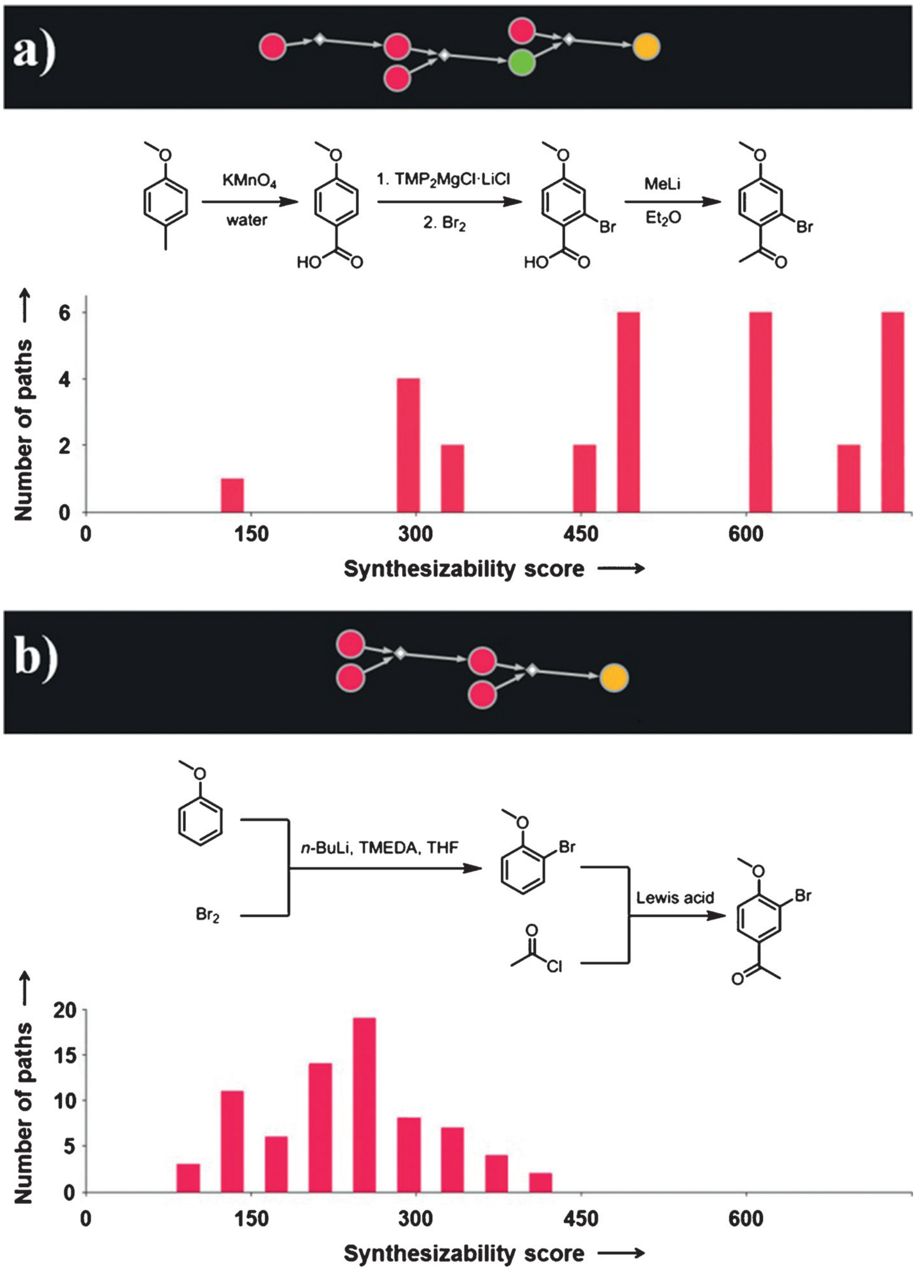


Figure 22. "Synthesizability" as estimated from the distribution of scores of computer-generated pathways. Even small differences in substitution patterns can translate into pronounced differences in synthesizability. Here, two bromomethoxyacetophenones are compared. For the example in (a) the best/lowest scoring pathway has the synthesizability score = 123; the average score for all pathways is 546.74 with standard deviation of 187.89. The compound in (b) is significantly easier to make, as reflected by the score of the best pathway = 83, average score = 231.15, and S.D. = 79.45. The computer also identifies more synthetic plans for this compound. All analyses were performed in Syntaurus with CSF = SMILES_LEN$^{3/2}$ + SMILES_LEN; RSF = 40 + 60·PROTECT + 50·CONFLICT^2 and tracing the searches to commercially available reagents with MW < 125.

4.2. The Missing versus Implausible Reaction Suggestions and Post-Processing of Synthetic Routes

In the early synthetic-planning programs, one of the major problems was an insufficient number of reaction rules. If such a program did not identify any plausible reactions, it was not necessarily that a viable synthesis did not exist but rather that

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

some relevant reaction rules/transforms were missing in the machine's knowledge base (e.g., as in LHASA[5] or SYNCHEM[8]). In contrast, modern programs such as ARChem,[41] IC$_{\text{SYNTH}}$,[42] or Syntaurus rely on complete or nearly complete reaction databases—where they go wrong is in predicting reactions which on paper might look acceptable but will not work in the laboratory. As we narrated earlier, the majority of these problems can be avoided at the level of properly coded reaction rules taking into account admissible substituents, stereochemistry, regiochemistry, protection group chemistry and reactivity conflicts, as well as electronic and most of steric effects, the latter at the scale of the reaction motif being coded. Perhaps the hardest challenge is to account for steric effects in complex molecules whereby the entire three-dimensional structure may dictate reactivity or its lack (Figure 23 and Ref. [55]). The study of the influence of steric effects on reaction outcomes dates back to Taft's work on the relative rates of the acid-catalyzed hydrolysis of esters.[56a] Since then, several models based on topological descriptors have been proposed,[53b,56b,c] of which the one developed by Cao and Liu[56d] has gained most popularity and has been included in the Marvin software from ChemAxon.[56e] In Cao and Liu's Topological Steric Effect Index (TSEI) each atom is assigned a number proportional to the steric crowding contributions of its neighbors weighted by the inverse cube of the topological distance (i.e., number of bonds) from the atom of interest. The main virtue of the model is that, unlike in many previous alkane-specific measures, it works for heteroatoms and can be calculated on sub-second scales even for very complex molecules. On the other hand, this and related approaches (e.g., Sello's steric descriptors[56f] accounting for the congestion of all reaction substrates and products) using topological rather than Cartesian distances do not truly account for the molecules' 3D shapes. One way around this problem might be to perform molecular mechanics conformational analyses, derive the probabilities of molecule's different conformations from energies (via Boltzmann relation, $p(E) \propto \exp(-E/kT)$, and then assign to all atoms conformation-averaged steric hindrance parameters $\langle S \rangle = \sum_i p_i S_i$ representative of the molecule's real shape.

Such calculations are, however, necessarily time-consuming—even if one conformational analysis takes on the order of 1 s, the evaluation of very large numbers (millions) of molecules considered during synthetic planning is clearly impractical. On the other hand, such analyses are reasonable for limited numbers (say, hundreds) of top-scoring complete pathways generated. This approach is reminiscent of computational drug design in which only the best-scoring candidates from large in silico screens are evaluated in detail. In synthetic design, such "post-processing" of the best pathways might extend beyond steric-hindrance analysis discussed above and also include molecular mechanics calculations to flag highly strained molecules, quantum mechanical calculations to examine electronic effects in detail, or even transition-state calculations to predict the outcomes of stereoselective reactions (e.g., using rapid molecular mechanics methods such as the ACE program developed by Moitessier and coworkers[56g,h]).

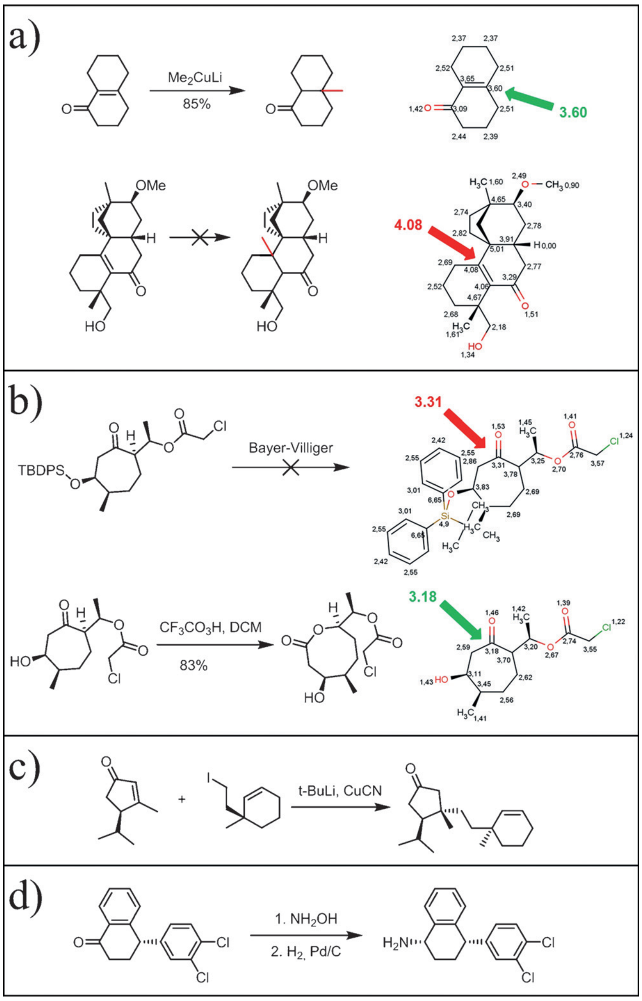


Figure 23. Examples of reactions whose outcomes are determined not by "local" reaction rules but by molecules' "global" 3D conformations. The numbers in the rightmost pictures are the values of the TSEI steric crowding index[56d,e] implemented in the Marvin software. a) Introduction of a quarternary methyl group via 1,4-addition of cyanocuprate might appear straightforward but in the example of Overman's synthesis of scopadulcic acid B,[55a] the desired product was not observed under any of the conditions tested. This complication was caused by the presence of bulky substituents close to the β-carbon of the enone. In this case, the high value of the TSEI index (4.08, red arrow) correctly reflects lack of reactivity at the position of interest. b) In Clardy's synthesis of (+)-octalactins A and B,[55b] the attempted Bayer–Villiger oxidation of a protected ketone was impossible due to presence of a bulky TBDPS group. After removal of this group, however, the desired lactone was obtained in good yield. In this case, the TBDPS-protected molecule does not react even though the value of the TSEI index at the carbonyl carbon is much lower (3.31) than in example (a). c) Diastereoselective addition of an organocuprate to an unsaturated ketone (an asymmetric induction popular in stereocontrolled synthesis and used here in the total synthesis of guanacastepene A)[55c] is, due to steric hindrance, favored from the less substituted side of the enone. In this case, topological indices based on distances measured in terms of bonds from the reaction center (like TSEI) are not sensitive to steric induction and thus not applicable. d) In Aggarwal's synthesis of Zoloft,[55d] reductive amination carried out in the last step gives high diastereomeric excess (1,4-*syn*:*anti* 96.5:3.5 isomer ratio)—this effect is due to the presence of a distant substituent which prohibits the approach of the catalyst from the more crowded side of a molecule. Again, topological indices such as TSEI cannot help us in determining reaction outcome.

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

One other aspect of the pathway post-processing that deserves a mention is an important problem of protection group management. In the earlier synthetic examples (cf. Figures 14 and 20 a,b), we have seen that computer can determine the protections at individual steps. On the other hand, a program planning in the "backwards"/retro direction does not know what chemistry will be chosen "next," and whether the protecting groups needed for yet-untaken steps will have survived during the "current" step we already considered. The only way around this problem is to first construct the entire retrosynthetic pathway and then retrace it "forward," from the substrates to the product—this time, however, with complete knowledge which protective groups are present in each step and whether they will or will not survive in subsequent steps. Although there are currently no computational tools for determining optimal management of protection groups over complete synthetic pathways, we believe the problem can be addressed by optimization algorithms combining individual protection/deprotection rules at each step (e.g., as already incorporated in Syntaurus) with pathway-wide minimization of the total number of protecion/deprotection steps required. We are actively working on developing such algorithms and hope to report the results in the near future.

4.3. Predicting Reaction Conditions

In addition to predicting whether a given reaction will or will not proceed, it would also be desirable for the computer to suggest suitable reaction conditions (solvent, catalyst, etc.). Such a capability is largely missing in the existing software, and the user is either referred to the literature on "similar" reactions (ARChem,[41] $\mathrm{IC}_{\mathrm{SYNTH}}$[42]) or to key publications describing a given reaction type (Syntaurus). It has only been recently that progress has been made by the Varnek's group based on their Condensed Graph of Reaction (CGR) formalism.[57a] In CGR, the set of all reactants and products is encoded by a single graph such that the reaction is represented as a pseudo-molecule for which molecular descriptors can be generated and further used in different chemoinformatics tasks (e.g., reaction similarity searching[57b] or the development of predictive models for kinetic parameters of $\mathrm{S_N2}$ reaction[57c]). In their latest publication, Varnek and co-workers showed that reaction-wide, CGR descriptors can be combined with machine learning methods (Support Vector Machines, Naive Bayes, and Random Forests) to successfully predict the conditions required for Michael addition reactions.

If extended to other reaction classes, methods like Varnek's could help interface synthesis-planning software with automated synthesis systems[58a–h] which have recently gained substantial popularity. These systems are often based on flow-chemistry[58b–h] hardware (e.g., flow chemistry approach is a stringent requirement in the recently announced DARPA's Make-It program)[58i]—in such cases, being able to predict whether reactions will succeed or fail in a specific solvent compatible with the automated flow protocols is of capital importance.

4.4. Predicting New Reaction Types/Mechanisms

Finally, with the synthetic planning using known methodologies finally maturing, we envision the next grand challenge of computer-assisted synthesis to be the in silico discovery of new reaction types and mechanisms. We know it is, in principle, possible since already in the 1990s Ivar Ugi demonstrated a computational discovery of some novel reactions using his bond-electron matrices.[31b] Today, software packages like RMG[59a,b] or EXGAS[59c] combine prediction of reaction rates using transition-state theory with quantum chemical calculations of activation barriers to generate feasible multistep reaction mechanisms for combustion processes. In an elegant recent paper,[60a] Aspuru-Guzik and co-workers showed how a combination of heuristic chemical rules (deriving from "arrow pushing")[60b,c] with rigorous quantum mechanical calculations and network theory can reproduce most intermediates and products involved in the complex mechanism of the well-known but still incompletely understood formose reaction.[61] These are promising early examples and the marriage of high-end theory with mechanistic organic chemistry can become truly impactful, especially if the QM calculation software is made user-friendly and thus accessible to practicing chemists.

5. Coda

In summary, we believe that there is plenty of exciting chemical discovery with modern computers. It is no longer 1960s when the computers were simply inadequate to the problems of organic chemistry—today, machines can already predict viable syntheses leading to quite complex targets and, with further development of computational methods, they can only become better. The time is ripe to integrate computational approaches with the everyday practice of organic synthesis and to initiate (or perhaps, recommence) the dialogue with our computer science colleagues. Our own experience tells us it is a fruitful discourse.

Acknowledgements

We gratefully acknowledge support from the Symfonia Award (UMO-2014/12/W/ST5/00592) from the Polish National Science Center (NCN). B.A.G. also gratefully acknowledges personal support from the Institute for Basic Science Korea, Project Code IBS-R020-D1. The development of automatic Syntaurus was supported by B.A.G.'s private funds. P.D. and M.S. acknowledge support from GSI, Inc. We thank Prof. Kyle J. M. Bishop (Penn State), Dr. Mikołaj Kowalik, Dr. Jarosław Granda and Piotr Janiuk for their help with algorithm development and useful discussions and also Prof. Bartłomiej Furman (Warsaw), Dr. Nicolas Armanino (Munich), and Prof. Dariusz Witt (Gdańsk) for their careful scrutiny of syntheses designed by Syntaurus. B.A.G. is thankful to all his former students and postdocs who contributed to the development of Chematica.

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

How to cite: *Angew. Chem. Int. Ed.* **2016**, *55*, 5904 – 5937
Angew. Chem. **2016**, *128*, 6004 – 6040

[1] P. Judson, *Knowledge-based Expert Systems in Chemistry: Not Counting on Computers*, RSC, Cambridge, **2009**.

[2] a) http://www.elsevier.com/online-tools/reaxys; b) https://scifinder.cas.org; c) http://www.chemspider.com; d) http://www.infochem.de/products/databases/index.shtml.

[3] a) M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2005**, *44*, 7263 – 7269; *Angew. Chem.* **2005**, *117*, 7429 – 7435; b) K. J. M. Bishop, R. Klajn, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2006**, *45*, 5348 – 5354; *Angew. Chem.* **2006**, *118*, 5474 – 5480; c) B. A. Grzybowski, K. J. M. Bishop, B. Kowalczyk, C. E. Wilmer, *Nat. Chem.* **2009**, *1*, 31 – 36.

[4] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31 – 36.

[5] a) E. J. Corey, W. J. Howe, R. D. Cramer, *J. Am. Chem. Soc.* **1972**, *94*, 421 – 430; b) http://cheminf.cmbi.ru.nl/cheminf/olp/history.shtml.

[6] D. A. Evans, *Angew. Chem. Int. Ed.* **2014**, *53*, 11140 – 11146; *Angew. Chem.* **2014**, *126*, 11320 – 11325.

[7] P. Y. Johnson, I. Bernstein, J. Crary, M. Evans, T. Wang, *Designing an Expert System for Organic Synthesis in Expert Systems Application in Chemistry* (Eds.: B. A. Holme, H. Pierce), ACS Symposium Series, Am. Chem. Soc. Washington, **1989**.

[8] a) H. L. Gelernter, A. F. Sanders, D. L. Larsen, K. K. Agarwal, R. H. Boivie, G. A. Spritzer, J. E. Searleman, *Science* **1977**, *197*, 1041 – 1049; b) D. Krebsbach, H. Gelernter, S. M. Sieburth, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 595 – 604.

[9] a) S. Hanessian, J. Franco, B. Larouche, *Pure Appl. Chem.* **1990**, *62*, 1887 – 1910; b) S. Hanessian, *Curr. Opin. Drug Discovery Dev.* **2005**, *8*, 798 – 819.

[10] S. Steinerberger, *Int. J. Game Theory*, **2015**, *44*, 761 – 767.

[11] L. V. Allis, *Searching for Solutions in Games and Artificial Intelligence*, Ph.D. thesis, University of Limburg, Maastricht, **1994**, p. 171.

[12] R. E. Korf, *Proc. AAAI'97/IAAI'97*, **1997**, pp. 700 – 705.

[13] Estimate based on applying Syntaurus' complete knowledge base of reactions to molecules from 99 paths selected from http://chemistrybydesign.oia.arizona.edu/; the paths chosen were 2 to 19 steps long with 9-step average.

[14] http://www.cube20.org/.

[15] http://www.eisai.com/news/news201133.html.

[16] a) J. L. Reymond, *Acc. Chem. Res.* **2015**, *48*, 722 – 730; b) Q. Hu, Z. Peng, J. Kostrowicki, A. Kuki, *Methods Mol. Biol.* **2011**, *685*, 253 – 276.

[17] The original 1957 report to the U.S.S.R. Academy of Sciences was later revised, expended and published in English as a) G. E. Vléduts, V. K. Finn, *Inf. Storage Retr.* **1963**, *1*, 101 – 116. Another publication from the same year is b) G. E. Vléduts, *Inf. Storage Retr.* **1963**, *1*, 117 – 146.

[18] https://vimeo.com/63343963.

[19] http://www.cas.org/etrain/scifinder/sciplanner.html.

[20] a) D. J. Watts, S. H. Strogatz, *Nature* **1998**, *393*, 440 – 442; b) M. Girvan, M. E. J. Newman, *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7821 – 7826.

[21] a) A. L. Barabási, Z. N. Oltvai, *Nat. Rev. Genet.* **2004**, *5*, 101 – U15; b) E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, A. L. Barabasi, *Science* **2002**, *297*, 1551 – 1555.

[22] C. Chaouiya, *Briefings Bioinf.* **2007**, *8*, 210 – 219.

[23] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, B. O. Palsson, *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 1777 – 1782.

[24] a) R. Albert, A. L. Barabasi, *Rev. Mod. Phys.* **2002**, *74*, 47 – 97; b) R. Albert, H. Jeong, A. L. Barabasi, *Nature* **1999**, *401*, 130 – 131; c) A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, *Comput. Netw.* **2000**, *33*, 309 – 320; d) M. Faloutsos, P. Faloutsos, C. Faloutsos, *Comput. Commun. Rev.* **1999**, *29*, 251 – 262; e) H. Jeong, B. Tombor, R. Albert, Z. N. Oltval, A. L. Barabasi, *Nature* **2000**, *407*, 651 – 654; F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, Y. Aberg, *Nature* **2001**, *411*, 907 – 908.

[25] a) M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Weckiewicz, P. E. Fuller, B. A. Grzybowski, K. J. M. Bishop, *Angew. Chem. Int. Ed.* **2012**, *51*, 7928 – 7932; *Angew. Chem.* **2012**, *124*, 8052 – 8056; b) C. M. Gothard, S. Soh, N. A. Gothard, B. Kowalczyk, Y. H. Wei, B. Baytekin, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2012**, *51*, 7922 – 7927; *Angew. Chem.* **2012**, *124*, 8046 – 8051; c) P. E. Fuller, C. M. Gothard, N. A. Gothard, A. Weckiewicz, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2012**, *51*, 7933 – 7937; *Angew. Chem.* **2012**, *124*, 8057 – 8061.

[26] a) L. C. Lee, *IRE Trans. Electron. Comput.* **1961**, *EC10*, 346 – 365; b) S. Skiena, *The Algorithm Design Manual*, Springer, London, **2008**, p. 480.

[27] a) S. Even, *Graph Algorithms*, 2nd ed., Cambridge Univ. Press, Cambridge, UK, **2011**, pp. 46 – 48; b) K. Mehlhorn, P. Sanders, *Algorithms and Data Structures: The Basic Toolbox*, Springer, Berlin, Heidelberg, **2008**.

[28] a) J. J. Masters, S. J. Danishefsky, J. T. Link, L. B. Snyder, W. B. Young, *Angew. Chem. Int. Ed.* **1995**, *34*, 1723 – 1726; *Angew. Chem.* **1995**, *107*, 1886 – 1888; b) S. J. Danishefsky, J. J. Masters, W. B. Young, J. T. Link, L. B. Snyder, T. V. Magee, D. K. Jung, R. C. A. Isaacs, W. G. Bornmann, C. A. Alaimo, C. A. Coburn, M. J. Di Grandi, *J. Am. Chem. Soc.* **1996**, *118*, 2843 – 2859; c) P. Wieland, K. Miescher, *Helv. Chim. Acta* **1950**, *33*, 2215 – 2228; d) M. L. Miller, P. S. Ray, *Synth. Commun.* **1997**, *27*, 3991 – 3996; e) M. Colin, D. Guenard, F. Gueritte-Voegelein, P. Potier (Rhone-Poulenc Sante), US4924012, **1990**; f) B. Ganem, R. R. Franke, *J. Org. Chem.* **2007**, *72*, 3981 – 3987.

[29] P. T. Anastas, M. M. Kirchoff, *Acc. Chem. Res.* **2002**, *35*, 686 – 694.

[30] a) R. S. Vardanyan, V. J. Hruby, *Synthesis of Essential Drugs*, Elsevier, Amsterdam, **2006**, p. 46; b) D. Farge, V. Marne, M. N. Messer, E. Moutonnier, C. Moutonnier (Rhône – Poulenc S. A.), U.S. Pat. 3641127, **1972**; c) W. Li, J. Li, Z.-K. Wan, J. Wu, W. Massefski, *Org. Lett.* **2007**, *9*, 4607 – 4610; d) A. R. Hajipour, A. E. Ruoho, *Org. Prep. Proced. Int.* **2002**, *34*, 647 – 651; e) A. A. Jalil, N. Kurono, M. Tokuda, *Synthesis* **2002**, *18*, 2681 – 2686; f) M. Allegretti, R. Bertini, M. C. Cesta, C. Bizzarri, R. D. Bitondo, V. D. Ciocco, E. Galliera, V. Berdini, A. Topai, G. Zampella, V. Russo, N. D. Bello, G. Nano, L. Nicolini, M. Locati, P. Fantucci, S. Florio, F. Colotta, *J. Med. Chem.* **2005**, *48*, 4312 – 4331; g) K. Mahesh, WO2013168001, **2013**; h) L. Baiocchi, M. Giannangeli, M. Bonanomi, G. Picconi, P. Ridolfi, *Gazz. Chim. Ital.* **1985**, *115*, 199 – 216; i) R. Ugo, P. Nardi, R. Psaro, D. Roberto, *Gazz. Chim. Ital.* **1992**, *122*, 511 – 514.

[31] a) R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, J. Lederberg, *Artif. Intell.* **1993**, *61*, 209 – 261; b) I. Ugi, J. Bauer, K. Bley, A. Dengler, A. Dietz, E. Fontain, B. Gruber, R. Herges, M. Knauer, K. Reitsam, N. Stein, *Angew. Chem. Int. Ed. Engl.* **1993**, *32*, 201 – 227; *Angew. Chem.* **1993**, *105*, 210 – 239.

[32] E. J. Corey, *Pure Appl. Chem.* **1967**, *14*, 19 – 37.

[33] a) E. J. Corey, X. M. Cheng, *The Logic of Organic Synthesis*, Wiley, New York, **1989**; b) S. Warren, P. Wyatt, *Organic Synthesis: The Disconnection Approach*, Wiley, Chichester, **2008**; c) K. C. Nicolaou, D. Vourloumis, N. Winssinger, P. S. Baran, *Angew. Chem. Int. Ed.* **2000**, *39*, 44 – 122; *Angew. Chem.* **2000**, *112*, 46 – 126; d) M. H. Todd, *Chem. Soc. Rev.* **2005**, *34*, 247 – 266.

[34] E. J. Corey, W. T. Wipke, *Science* **1969**, *166*, 178 – 192.

[35] a) W. T. Wipke, W. J. Howe, Computer-Assisted Organic Synthesis ACS Centennial Meeting, New York, **1976**; b) W. T. Wipke, G. I. Ouchi, S. Krishnan, *Artif. Intell.* **1978**, *11*, 173 – 193.

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

[36] a) J. B. Hendrickson, A. G. Toczko, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 137–145; b) J. B. Hendrickson, *J. Am. Chem. Soc.* **1977**, *99*, 5439–5450; c) J. B. Hendrickson, P. Huang, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 145–151; d) J. B. Hendrickson, *Angew. Chem. Int. Ed. Engl.* **1990**, *29*, 1286–1295; *Angew. Chem.* **1990**, *102*, 1328–1338.

[37] Other types of retrosynthetic programs that we were able to identify, though not expound on in the main text, include: Synthetic accessibility of organic compounds: a) SILVIA: http://www.molecular-networks.com/products/sylvia; b) F. Pennerath, G. Niel, P. Vismara, P. Jauffret, C. Laurenço, A. Napoli, *J. Chem. Inf. Model.* **2010**, *50*, 221–239; Systems based on prediction of reaction's products: c) J. H. Chen, P. Baldi, *J. Chem. Inf. Model.* **2009**, *49*, 2034–2043; d) BEPPE: G. Sello, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 713–717; e) ROBIA: I. M. Socorro, J. M. Goodman, *J. Chem. Inf. Model.* **2006**, *46*, 606–614; f) I. M. Socorro, K. Taylor, J. M. Goodman, *Org. Lett.* **2005**, *7*, 3541–3544; g) LILITH, bond polarity guided: L. Baumer, G. Sala, G. Sello, *J. Am. Chem. Soc.* **1991**, *113*, 2494–2500; h) SYNSUP, allowing for reagent/functional group interactions: A. Tanaka, H. Okamoto, M. Bersohn, *J. Chem. Inf. Model.* **2010**, *50*, 327–338; i) CROSS, allowing for robust functionalisation of side chains and rescaffolding: A. Evers, G. Hessler, L. H. Wang, S. Werrel, P. Monecke, H. Matter, *J. Med. Chem.* **2013**, *56*, 4656–4670; j) ELN-mining software: C. D. Christ, M. Zentgraf, J. M. Kriegl, *J. Chem. Inf. Model.* **2012**, *52*, 1745–1756.

[38] a) J. Bauer, R. Herges, E. Fontain, I. Ugi, *Chimia* **1985**, *39*, 43–53; b) *Cheminformatics Developments: History, Reviews and Current Research* (Ed.: J. H. Noordik), IOS Press, Amsterdam, **2004**.

[39] http://www.cogsys.wiai.uni-bamberg.de/effalip/installguide.html.

[40] a) "The Prediction of Chemical Reactions": J. Gasteiger in *Chemoinformatics. A Textbook* (Eds.: J. Gasteiger, T. Engel), Springer, Heidelberg, **1990**, pp. 542–567; b) R. Höllering, J. Gasteiger, L. Steinhauer, K. Schultz, A. Herwig, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 482–494.

[41] a) O. Ravitz, *Drug Discovery Today Technol.* **2013**, *10*, e443–e449; b) J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade, H. Y. Ando, *J. Chem. Inf. Model.* **2009**, *49*, 593–602.

[42] a) H. Kraut, J. Eiblmaier, G. Grethe, P. Loew, H. Matuszczyk, H. Saller, *J. Chem. Inf. Model.* **2013**, *53*, 2884–2895; b) A. Bøgevig, H. J. Federsel, F. Huerta, M. G. Hutchings, H. Kraut, T. Langer, P. Loew, C. Oppawsky, T. Rein, H. Saller, *Org. Process Res. Dev.* **2015**, *19*, 357–368.

[43] K. O. Geddes, S. R. Czapor, G. Labahn, *Algorithms for Computer Algebra*, Springer US, **1992**.

[44] A. Cadeddu, E. K. Wylie, J. Jurczak, M. Wampler-Doty, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2014**, *53*, 8108–8112; *Angew. Chem.* **2014**, *126*, 8246–8250.

[45] a) SMARTS theory manual, Daylight Chemical Information Systems Inc., Aliso Viejo, CA 92656, USA. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, accessed 23rd June 2015; b) http://www.rdkit.org.

[46] F. A. Van-Catledge, *J. Org. Chem.* **1980**, *45*, 4801–4802.

[47] a) F. D. Da Silva Araújo, L. C. De Lima Fávaro, W. L. Araújo, F. L. De Oliveira, R. Aparicio, A. J. Marsaioli, *Eur. J. Org. Chem.* **2012**, *27*, 5225–5230; b) P. Ellerbrock, N. Armanino, D. Trauner, *Angew. Chem. Int. Ed.* **2014**, *53*, 13414–13418; *Angew. Chem.* **2014**, *126*, 13632–13636; c) S. Sang, J. D. Lambert, S. Tian, J. Hong, Z. Hou, J. H. Ryu, R. E. Stark, R. T. Rosen, M. T. Huang, C. S. Yang, et al., *Bioorg. Med. Chem.* **2004**, *12*, 459–467; d) C. B. Rao, D. C. Rao, D. C. Babu, Y. Venkateswarlu, *Eur. J. Org. Chem.* **2010**, *2010*, 2855–2859; e) G. Casiraghi, L. Battistini, C. Curti, G. Rassu, F. Zanardi, *Chem. Rev.* **2011**, *111*, 3076–3154.

[48] Yield estimations are based on thermodynamic calculations combined with multidimensional optimization. This model is described in F. S. Emami, A. Vahid, W. K. Wylie, S. Szymkuc, P. Dittwald, K. Molga, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2015**, *54*, 10797–10801; *Angew. Chem.* **2015**, *127*, 10947–10951.

[49] a) K. H. Lim, V. J. Raja, T. D. Bradshaw, S.-H. Lim, Y.-Y. Low, T.-S. Kam, *J. Nat. Prod.* **2015**, *78*, 1129–1138; b) X. Zhong, Y. Li, F. S. Han, *Chem. Eur. J.* **2012**, *18*, 9784–9788; c) H. Iio, M. Monden, K. Okada, T. Tokoroyama, *J. Chem. Soc. Chem. Commun.* **1987**, 358–359; d) D. Enders, T. Hundertmark, R. Lazny, *Synlett* **1998**, 721–722; e) A. L. Gutman, M. Etinger, G. Nisnevich, F. Polyak, *Tetrahedron: Asymmetry* **1998**, *9*, 4369–4379; f) M. A. Potopnyk, B. Lewandowski, S. Jarosz, *Tetrahedron: Asymmetry* **2012**, *23*, 1474–1479.

[50] a) Y. H. Lan, F.-R. Chang, Y. L. Yang, Y. C. Wu, *Chem. Pharm. Bull.* **2006**, *54*, 1040–1043; b) J. S. Yadav, N. Rami Reddy, V. Harikrishna, B. V. Subba Reddy, *Tetrahedron Lett.* **2009**, *50*, 1318–1320; c) M. Venkataiah, P. Somaiah, G. Reddipalli, N. W. Fadnavis, *Tetrahedron: Asymmetry* **2009**, *20*, 2230–2233; d) J. Li, H. Zheng, Y. Su, X. Xie, X. She, *Synlett* **2010**, *15*, 2283–2284; e) J. S. Yadav, R. Nageshwar Rao, R. Somaiah, V. Harikrishna, B. V. Subba Reddy, *Helv. Chim. Acta* **2010**, *93*, 1362–1368; f) G. S. Forman, R. P. Tooze, *J. Organomet. Chem.* **2005**, *690*, 5863–5866; g) N. Yoshikawa, N. Kumagai, S. Matsunaga, G. Moll, T. Ohshima, T. Suzuki, M. Shibasaki, *J. Am. Chem. Soc.* **2001**, *123*, 2466–2467; h) J. H. Xie, L. C. Guo, X. H. Yang, L. X. Wang, Q. L. Zhou, *Org. Lett.* **2012**, *14*, 4758–4761.

[51] a) J. Manville, C. Kriz, *Can. J. Chem.* **1977**, *55*, 2547–2553; b) R. G. Almquist, J. Crase, C. Jennings-White, R. F. Meyer, M. L. Hoefle, R. D. Smith, A. D. Essenburg, H. R. Kaplan, *J. Med. Chem.* **1982**, *25*, 1292–1299; c) F. Dehmel, H. G. Schmalz, *Org. Lett.* **2001**, *3*, 3579–3582; d) D. J. Chang, S. Lee, J. Jang, S. O. Kim, W. J. Kim, Y. G. Suh, *Bioorg. Med. Chem. Lett.* **2012**, *22*, 6750–6755.

[52] a) S. E. Drewes, B. M. Sehlapelo, M. M. Horn, R. Scott-Shaw, P. Sandor, *Phytochemistry* **1995**, *38*, 1427–1430; b) M. B. Boxer, H. Yamamoto, *J. Am. Chem. Soc.* **2007**, *129*, 2762–2763; c) P. R. Krishna, V. V. R. Reddy, *Tetrahedron Lett.* **2005**, *46*, 3905–3907; d) A. Martinez, K. Zumbansen, A. Dohring, M. van Gemmeren, B. List, *Synlett* **2014**, *25*, 932–934; e) K. Liu, G. Zhang, *Tetrahedron Lett.* **2015**, *56*, 243–246; f) B. List, P. Pojarliev, C. Castello, *Org. Lett.* **2001**, *3*, 573–575; g) M. A. Blanchette, M. S. Malamas, M. H. Nantz, J. C. Roberts, P. Somfai, D. C. Whritenour, S. Masamune, M. Kageyama, T. Tamura, *J. Org. Chem.* **1989**, *54*, 2817–2825; h) Y. Yamaoka, H. Yamamoto, *J. Am. Chem. Soc.* **2010**, *132*, 5354–5356; i) R. Mahrwald, *Modern Methods in Stereoselective Aldol Reactions*, Wiley-VCH, Weinheim, **2013**; In case of complications mentioned in the main text, we envision two potential solutions. The first one would involve separation of the obtained 1,5-*syn-* and 1,5-*anti-* diastereoisomers and conversion of undesired *anti-* into *syn-* derivative via invertive Mitsunobu methodology described in Refs. [52j–l]; j) S. F. Martin, J. A. Dodge, *Tetrahedron Lett.* **1991**, *32*, 3017–3020; k) T. Sammakia, J. S. Jacobs, *Tetrahedron Lett.* **1999**, *40*, 2685–2688; l) J-M. Vatèle, *Tetrahedron* **2007**, *63*, 10921–10929; In the second solution, the order of the two steps could be exchanged (i.e., acylation with acryoyl chloride carried out after aldol reaction with subsequent hydroxyl group protection/deprotection steps) forcing 1,5-*syn*-selectivity via "supersilyl" protecting group see Refs [52b,h]); m) K. M. Chen, G. E. Hardtmann, K. Prasad, O. Repič, M. J. Shapiro, *Tetrahedron Lett.* **1987**, *28*, 155–158.

[53] a) J. Li, M. D. Eastgate, *Org. Biomol. Chem.* **2015**, *13*, 7164–7176; b) M. Randić, *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615; c) M. Randić, *J. Am. Chem. Soc.* **1977**, *99*, 444–450; d) T. Gaich, P. S. Baran, *J. Org. Chem.* **2010**, *75*, 4657–4673; e) K. Boda, T. Seidel, J. Gasteiger, *J. Comput.-Aided Mol. Des.* **2007**, *21*, 311–

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

325; f) J. Gasteiger, *Nat. Chem.* **2015**, *7*, 619–620; g) P. Ertl, A. Schuffenhauer, *J. Cheminf.* **2009**, *1*, 8.

[54] B. A. Grzybowski, A. V. Ishchenko, J. Shimada, E. I. Shakhnovich, *Acc. Chem. Res.* **2002**, *35*, 261–269.

[55] a) L. E. Overman, D. J. Ricca, V. D. Tran, *J. Am. Chem. Soc.* **1997**, *119*, 12031–12040; b) J. C. McWilliams, J. Clardy, *J. Am. Chem. Soc.* **1994**, *116*, 8378–8379; c) A. K. Miller, C. C. Hughes, J. J. Kennedy-Smith, S. N. Gradl, D. Trauner, *J. Am. Chem. Soc.* **2006**, *128*, 17057–17062; d) S. Roesner, J. M. Casatejada, T. G. Elford, R. P. Sonawane, V. K. Aggarwal, *Org. Lett.* **2011**, *13*, 5740–5743.

[56] a) R. W. Taft, *J. Am. Chem. Soc.* **1952**, *74*, 3120–3128; b) E. Estrada, E. Molina, L. I. Perdomo, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1015–1021; c) O. Ivanciuc, A. T. Balaban, *Croat. Chem. Acta* **1996**, *69*, 75–83; d) C. Cao, L. Liu, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 678–687; e) https://www.chemaxon.com/products/marvin/; f) G. Sello, *Tetrahedron* **1998**, *54*, 5731–5744; g) N. Weill, C. R. Corbell, J. W. De Schutter, N. Moitessier, *J. Comput. Chem.* **2011**, *32*, 2878–2889; h) C. R. Corbeil, S. Thielges, J. A. Schwartzentruber, N. Moitissier, *Angew. Chem. Int. Ed.* **2008**, *47*, 2635–2638; *Angew. Chem.* **2008**, *120*, 2675–2678.

[57] a) A. Varnek, D. Fourches, F. Hoonakker, V. P. Solov′ev, *J. Comput.-Aided Mol. Des.* **2005**, *19*, 693–703; b) A. de Luca, D. Horvath, G. Marcou, V. P. Solov′ev, A. Varnek, *J. Chem. Inf. Model.* **2012**, *52*, 2325–2338; c) T. I. Madzhidov, P. G. Polishchuk, R. I. Nugmanov, A. V. Bodrov, A. I. Lin, I. I. Baskin, I. S. Antipin, A. Varnek, *Russ. J. Org. Chem.* **2014**, *50*, 459–463; d) G. Marcou, J. Aires de Sousa, D. Latino, A. Deluca, D. Horvath, V. Rietsch, A. Varnek, *J. Chem. Inf. Model.* **2015**, *55*, 239–250.

[58] a) J. Li, S. G. Balmer, E. P. Gillis, S. Fuji, M. J. Schmidt, A. M. E. Palazzolo, J. W. Lehmann, G. F. Morehouse, M. D. Burke, *Science* **2015**, *347*, 1221–1226; b) D. Ghislieri, K. Gilmore, P. H. Seeberger, *Angew. Chem. Int. Ed.* **2015**, *54*, 678–682; *Angew. Chem.* **2015**, *127*, 688–692; c) K. S. Elvira, X. C. I. Solvas, R. C. R. Wootton, A. J. deMello, *Nat. Chem.* **2013**, *5*, 905–915; d) T. Kourti, *Anal. Bioanal. Chem.* **2006**, *384*, 1043–1048; e) R. L. Hartman, K. F. Jensen, *Lab Chip* **2009**, *9*, 2495–2507; f) S. V. Ley, D. E. Fitzpatrick, R. J. Ingham, R. M. Meyers, *Angew. Chem. Int. Ed.* **2015**, *54*, 3449–3464; *Angew. Chem.* **2015**, *127*, 3514–3530; g) S. V. Ley, D. E. Fitzpatrick, R. M. Meyers, R. J. Ingham, C. Battilocchio, R. J. Ingham, *Angew. Chem. Int. Ed.* **2015**, *54*, 10122–10136; *Angew. Chem.* **2015**, *127*, 10260–10275; h) B. Gutmann, D. Cantillo, C. O. Kappe, *Angew. Chem. Int. Ed.* **2015**, *54*, 6688–6728; *Angew. Chem.* **2015**, *127*, 6788–6832; i) Solicitation of the DARPA's Make-It Program can be downloaded from http://go.usa.gov/3Pzww.

[59] a) for documentation of Reaction Mechanism Generator, see http://rmg.mit.edu; b) M. R. Harper, K. M. Van Geem, S. P. Pyl, G. B. Marin, W. H. Green, *Combust. Flame* **2011**, *158*, 16–41; c) V. Warth, F. Battin-Leclerc, R. Fournet, P. A. Glaude, G. M. Côme, G. Scacchi, *Comput. Chem.* **2000**, *24*, 541–560.

[60] a) D. Rappoport, C. J. Galvin, D. Y. Zubarev, A. Aspuru-Guzik, *J. Chem. Theory Comput.* **2014**, *10*, 897–907; b) D. E. Levy, *Arrow-Pushing in Organic Chemistry. An Easy Approach to Understanding Reaction Mechanisms*, Wiley, Hoboken, **2008**; c) G. Knizia, J. E. M. N. Klein, *Angew. Chem. Int. Ed.* **2015**, *54*, 5518–5522; *Angew. Chem.* **2015**, *127*, 5609–5613.

[61] a) A. Boutlerow, *C. R. Acad. Sci.* **1861**, *53*, 145–147; b) T. Zweckmair, S. Böhmdorfer, A. Bogolitsyna, T. Rosenau, A. Potthast, S. Novalin, *J. Chromatogr. Sci.* **2014**, *52*, 169–175.

Received: July 3, 2015
Revised: September 14, 2015
Published online: April 8, 2016

© 2016 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim