Chemical Networks (2)

DOI: 10.1002/ange.201202209

Parallel Optimization of Synthetic Pathways within the Network of Organic Chemistry

Mikołaj Kowalik, Chris M. Gothard, Aaron M. Drews, Nosheen A. Gothard, Alex Weckiewicz, Patrick E. Fuller, Bartosz A. Grzybowski, and Kyle J. M. Bishop**

The entire chemical-synthetic knowledge created since the days of Lavoisier to the present can be represented[1–4] as a complex network (Figure 1 a) comprising millions of compounds and reactions. While it is simply beyond cognition of any individual human to understand and analyze all this collective chemical knowledge, modern computers have become powerful enough to perform suitable network analyses within reasonable timescales. In this context, a problem that is both fundamentally interesting and practically important is the identification of optimal synthetic pathways leading to desired, known molecules from commercially available substrates. In either manual searches or semi-automated search tools, such as Reaxys,[5a] this procedure is done by back-tracking the possible syntheses step-by-step. Such "manual" methods, however, give virtually no chance of finding an optimal pathway, as the number of possible syntheses to consider is very large (for example, ca. 10^{19} within five steps). Moreover, the problem becomes dramatically more complex when one aims to optimize the syntheses of multiple substances simultaneously when, for example, a company producing N products would strive to design synthetic pathways sharing many common substrates/intermediates and minimizing the overall synthetic cost (Figure 1 a). As we show herein, however, judicious combination of combinatorial optimization with network search algorithms allows the parallel optimization of tens to thousands of syntheses. The algorithms we describe traverse the network of organic chemistry (henceforth, NOC or simply the network) probing different synthetic paths according to the cost criterion as defined by a combination of labor cost and the cost of staring materials. In a specific case study, we show that our optimization can reduce the cost of an existing synthetic company (here, ProChimia Surfaces)[5b] by almost 50%. Overall, this communication is the first instance in which synthetic optimizations are based on the entire body of synthetic knowledge as stored in the NOC and combined with economical descriptors (that is, prices). While each of the individual reactions in the NOC is known, the network search algorithms create new chemical knowledge in the form of near optimal reaction sequences; notably, the syntheses that are optimal for making any molecule individually can be different from those optimizing the synthesis of this and other molecules simultaneously.

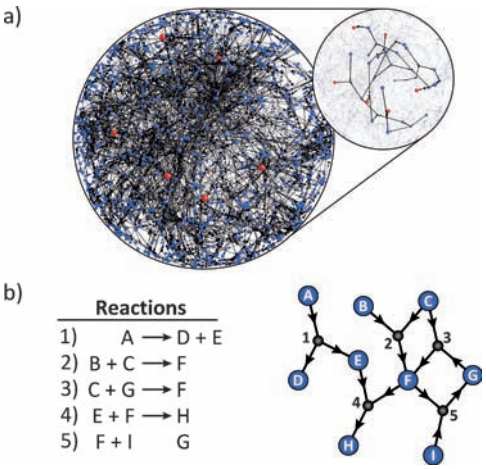

Figure 1. The network of organic chemistry and its bipartite wiring plan. a) Small fraction of the network (ca. 0.025%) centered on six target compounds (red). Computational methods described herein allow for the identification of near optimal synthesis plans (inset) despite the size and complexity of the network. b) Illustration of the mapping from a list of chemical reactions to a directed, bipartite network.

Our analyses are based on a network of about 7 million reactions and about 7 million substances derived as described in the first communication in this series[6] (also see Refs. [1,2]). While in our earlier analyses of NOC, the simple dot–arrow representation was typically sufficient, the analysis of specific syntheses involving multiple substrates and/or products requires the so-called bipartite-graph representation with two types of nodes: those corresponding to specific substances (blue dots in Figure 1 b), and those representing the reactions (black dots in Figure 1 b). This representation of the NOC captures the causal synthetic dependencies and accounts for the fact that a viable synthesis (see the Supporting Information, Section 2) cannot proceed without all of the necessary reactants, which must either be synthesized by another suitable reaction or purchased.

Also, as our network searches are intended to compare the actual costs of syntheses, we have linked the NOC to a test

[*] Dr. M. Kowalik, A. M. Drews, Prof. K. J. M. Bishop
Department of Chemical Engineering
The Pennsylvania State University, Fenske Laboratory
University Park, PA 16802 (USA)
E-mail: kjmbishop@engr.psu.edu
Homepage: https://sites.google.com/site/kjmbishop/
Dr. C. Gothard, N. Gothard, A. Weckiewicz, P. E. Fuller, Prof. B. A. Grzybowski
Department of Chemistry, Northwestern University
2145 Sheridan Rd., Evanston, IL 60208 (USA)
E-mail: grzybor@northwestern.edu
Homepage: http://dysa.northwestern.edu

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201202209.

database of about 20000 common, commercially available chemicals from Sigma Aldrich (other supplier databases can also be linked to the software). Since the majority of these chemicals are available in different quantities, we derived the scaling relationship between their cost c_{sub} (in US$) and amount x (in grams for solids, milliliters for liquids). For the majority of the Aldrich compounds, this relationship is well-approximated by a power law $c_{sub} = \alpha x^{\beta}$, where α and β are free parameters, and the distribution of β exponents has a sharp peak at 0.75 such that $c_{sub} \approx x^{0.75}$ (Supporting Information, Section 1). This scaling relation reflects economies of scale, whereby the cost per amount (c/x) decreases with increasing scale. Moreover, knowledge of the most probable exponent β allows the identification of α, which characterizes the cost of 1 g (or 1 mL) of a given substance and can be used to fairly compare the costs of different compounds.

With these preliminaries, we turn our attention to the synthetic pathways within the network. First, to estimate the complexity of the optimization problem we are about to address, we developed a recursive depth-first search (DFS) algorithm that counts the numbers of possible synthesis plans of a prescribed depth d (that is, the maximum number of synthetic steps from the target; Figure 2a). The results summarized in Figure 2b indicate that the number of substances and reactions relevant to the synthesis of a given target increases exponentially with depth, while the number

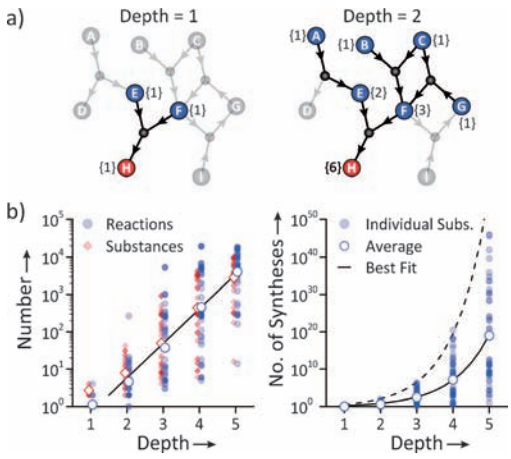

Figure 2. a) Counting possible syntheses at two different depths d from the target product (node denoted H). At depth one, there is only one possible synthesis from E and F. At depth two, there are six possible syntheses. F can be purchased and used as a starting substrate, synthesized from substrates B and C, or 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. b) Based on network searches in the vicinity of 51 different target substances (Supporting Information, Section 4), the number of individual reactions (blue) and substances (red) relevant to the synthesis of each target increases exponentially, as ca. $(8.5)^d$, with increasing distance from the target (left). The number of possible syntheses (that is, plans combining individual reactions) grows even faster; here, as $(1.4)^{(2.7)^d}$, as illustrated by the solid black curve (right). 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, and the dashed line is an upper bound for the estimates.

of possible syntheses increases even faster, as $(1.4)^{(2.7)^d}$ (Supporting Information, Section 2.2 for algorithmic details). Within five synthetic steps of a given target, the number of syntheses to consider reaches about 10^{19} on average. While this number might seem large in itself, it is but a small fraction of the possibilities associated with the synthesis of multiple targets, which for $d = 3$ and, say, 10 targets is about 300^{10} (or ca. 10^{24}); for 50 products there are about 10^{950} possible syntheses within five synthetic steps!

The above estimates suggest that: 1) for a single target the numbers of syntheses are still within computational limits, and searches can be performed by deterministic methods (following the branches of trees propagating on the network away from a substance of interest); 2) in contrast, the myriad possibilities involved in the parallel optimization of multiple targets cannot be enumerated exhaustively by even the fastest available computers; consequently, stochastic search methods are needed to bias the searches toward near optimal solutions.

Both the deterministic and the stochastic algorithms we develop strive to minimize the total cost of synthesizing one or more desired target molecules starting from commercially available substrates. In general, the synthesis cost may include numerous contributions such as starting materials, solvents, overhead costs, and labor costs. Here, as a proof of concept, we adopt a simplified model of total synthetic cost as a sum of reaction costs and substrate costs, $c_{tot} = c^0_{rxn} N_{rxn} + \Sigma c_{sub}(i)$. Each of N_{rxn} reactions in an overall synthetic path is assumed to contribute a fixed cost c^0_{rxn} and includes implicitly the costs of labor, overhead, and separations processes. Each substrate i obtained from a chemical supplier costs $c_{sub}(i)$. We make three further comments regarding this cost function: 1) Through the $c_{sub} \approx x^{0.75}$ relation, it can account for different scales of the syntheses; 2) at the same time, it does not reflect reaction yields as they are, unfortunately, not provided for the majority (ca. 90%) of reactions reported in the databases from which the NOC is derived; 3) by changing the reaction cost c^0_{rxn}, it allows for adjusting for different economic realities; for example, situations where labor costs are higher than substrate costs in developed countries versus the opposite in developing countries.

We first consider the optimization of syntheses leading to one specified target molecule. In this case, possible syntheses are examined using a recursive algorithm that back-propagates on the network starting from the target. At the first backward step, the algorithm examines all reactions leading to the target and calculates the minimum cost (given by the cost function discussed above) associated with each of them. This calculation, in turn, depends on the minimum costs of the associated reactants that may be purchased or synthesized. In this way, the cost calculation continues recursively, moving backward from the target until a critical search depth is reached (for algorithm details, see the Supporting Information, Section 2.3). Provided each branch of the synthesis is independent of the others (good approximation for individual targets, not for multiple targets), this algorithm rapidly identifies the synthetic plan which minimizes the cost criterion.

Figure 3 illustrates the results of searches for several targets of chemical interest. One noteworthy feature of our

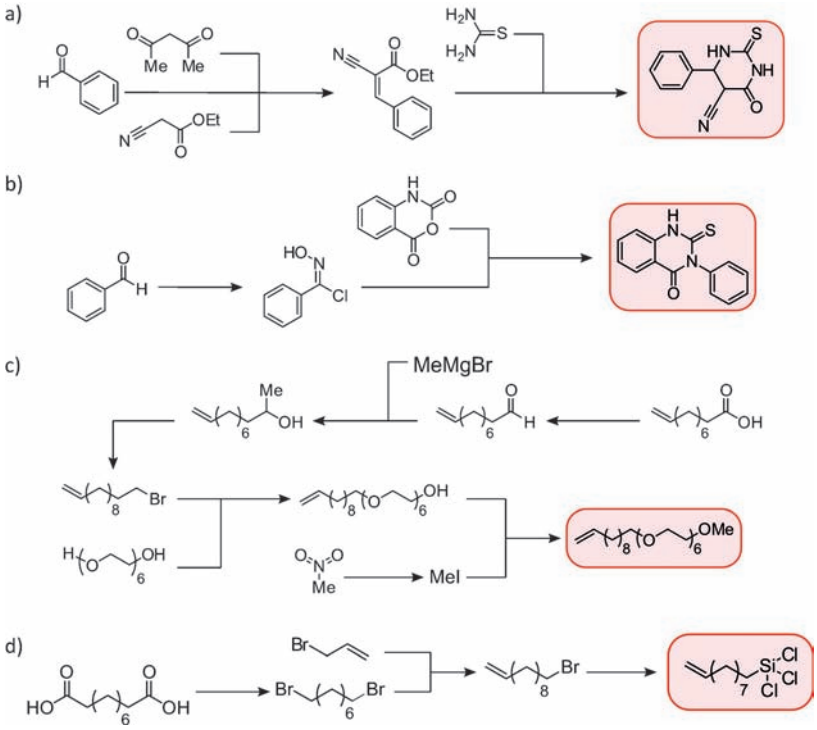

Figure 3. Algorithmically-identified optimal syntheses (with $c^0_{rxn} = 0.84$) of a) a pyrimidine derivative (an active pharmaceutical ingredient); b) a dihydroquinazoline derivative (a key intermediate in the synthesis of analgesic, anti-inflammatory, and anticonvulsant quinazolines); c) an oligo(ethylene glycol) derivative (a popular precursor to molecules constituting bioresistant monolayers); and d) a long-chain alkenyl trichlorosilane (used for functionalization of glass and PDMS surfaces). For each target, the optimal synthesis was selected from an enormous number of possible routes, namely, 10^{46}, 10^{64}, 10^{24}, and 10^{8} possibilities for targets (a)–(d), respectively.

algorithm is that it efficiently identifies possible tandem processes and multicomponent reactions within optimal synthetic routes. For example, a pyrimidine derivative in Figure 3a is prepared by an aldol–Michael tandem reaction sequence of nitriles to form an unsaturated nitrile intermediate, which then reacts with thiourea to provide a desired hexahydropyrimidine product. Similarly, synthesis of dihydroquinazoline derivative in Figure 3b starts with an oximation–chlorination tandem sequence of commercially available benzaldehyde to form hydroxybenzimidoyl chloride. The resulting imidoyl chloride then reacts with thiourea to form isothiocyanate. Subsequent multicomponent tandem reaction of isothiocyante with an alkylamine and isatoic anhydride provides dihydroquinazoline as desired. Figure 3c and d depict synthetic pathways leading to, respectively, a widely applicable oligo(ethylene glycol) derivative and a long-chain alkenyl trichlorosilane. Both of these synthetic plans are reasonable and have actually been used by ProChima, which markets these molecules.

In the above examples, the reaction cost parameter was fixed at $c^0_{rxn} = 0.84$; however, the optimal syntheses identified algorithmically depend strongly on the details of the cost function, and especially c^0_{rxn}. In a particularly interesting and chemically elegant example (Figure 4a), the optimal synthesis of the dihydroquinazoline natural product changes dramatically, from a single reaction for $c^0_{rxn} = 10$ to a set of 8 different reactions for $c^0_{rxn} = 0.1$. In the longer synthesis, the algorithm finds a pathway that starts from the common benzaldehyde substrate, then splits into two branches which ultimately reconnect into the final product. This second approach avoids the need for additional, more costly starting substrates (namely, isatoic anhydride). It is worth noting that identifying this pathway by traditional, one-step-at-the-time searches would be extremely improbable: this is so because upon back-propagation of the searches from the target, the left and the right sub-trees in the synthetic plan diverge and the chance of finding a path that reconnects them three steps away (at the benzaldehyde substarate) are about $1{:}10^6$. Further synthetic examples are included in the Supporting Information, Section 5 (for example, that of the cholesterol-lowering drug Ezetimibe).

These reaction-cost versus reaction-length trends are quite general. To show this, we computed the optimal synthesis plans for each of the 51 compounds (Supporting Information, Section 4) chosen among the products of ProChimia surfaces for different reaction costs c^0_{rxn} (Figure 4b,c). On average, decreasing the reaction cost c^0_{rxn} causes 1) the size of the synthesis (that is, the number of reactions) to increase and 2) the substrate cost to decrease (Figure 4c). High reaction costs favor shorter syntheses; low reaction costs lead to longer syntheses that make use of cheaper substrates.

While effective for optimizing syntheses of individual molecules, the deterministic algorithm cannot analyze exhaustively the staggering number of possibilities involved in the simultaneous optimization of syntheses leading to multiple targets. To remedy this, we implemented a simulated annealing method[7] by which probabilistic searches among possible syntheses are increasingly biased towards those with the lowest cost. Briefly, the search is initialized with some randomly generated, mock synthesis plan. This plan is then altered using two types of Monte Carlo moves: 1) reaction insertion/removal or 2) substrate insertion/removal. Those moves are accepted or rejected probabilistically in accordance with the Metropolis criteron such that each viable synthesis plan j is visited with probability $p_j \approx e^{-\beta c_{tot}(j)}$, where β is an adjustable parameter analogous to inverse temperature in a physical system, $\beta \approx 1/T$. As β is slowly increased, only those plans with the lowest costs are explored with any significant probability (Supporting Information, Section 3 for details), and the search evolves towards a globally minimal cost (Figure 5a).

To assess the real-world performance of this network-wide optimization method, we conducted a case study on a selection of 51 products (Supporting Information, Section 4) sold by

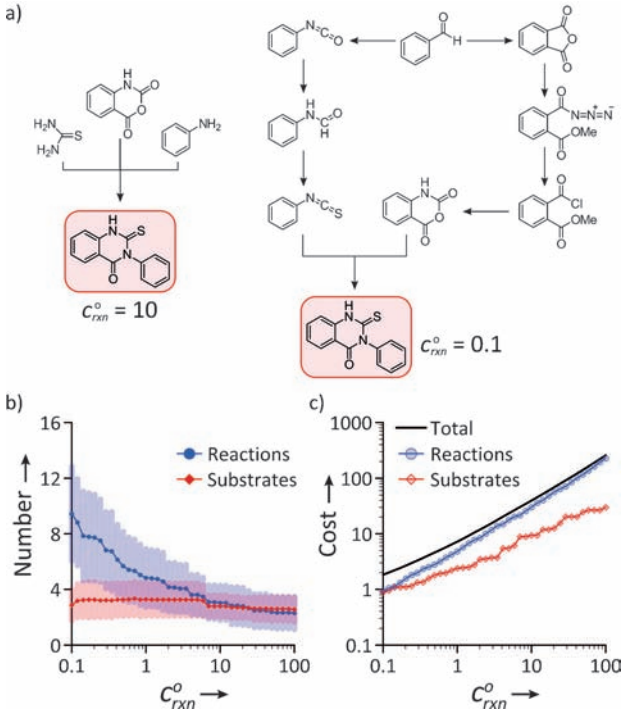

Figure 4. Varying the cost parameter leads to different optimal syntheses. a) Two different optimal syntheses of a dihydroquinazo-line natural-product derivative for $c^0_{rxn} = 10$ (left) and $c^0_{rxn} = 0.1$ (right) along with that shown in Figure 3 b for $c^0_{rxn} = 0.84$. b) For a set of 51 different target substances (Supporting Information, Section 3),the number of reactions (blue) and substrates (red) in the optimal synthesis decreases with increasing reaction costs c^0_{rxn}. c) As the reaction cost decreases and the synthetic pathways become longer, the algorithm traces the synthetic "trees" to cheaper substrates.

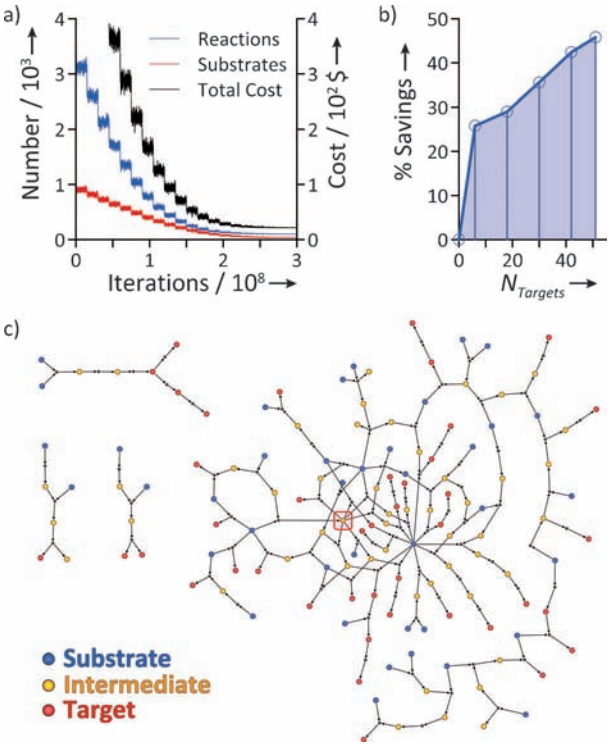

Figure 5. Global synthesis optimization by simulated annealing. a) During the annealing process, the total synthesis cost decreases steadily (but not monotonically) to an almost optimal value. b) Percent savings as a function of the number of target compounds (Supporting Information, Section 3). c) Network plan for the optimal synthesis plan for 51 ProChimia products for a reaction cost of $c^0_{rxn} = 10$. The total cost is $c_{tot} =$ US$ 21.5 per gram per target, with 90% coming from reaction costs and 10% from substrate costs (reflecting ProChimia's actual labor vs. substrate expenditures averaged from 2005–2012). Note that for most products, the algorithm finds a common synthetic tree in which key intermediates are shared in the synthesis of different products. The node enclosed by a red box is undecylenic bromide, which is one of the hub intermediates discussed in the main text.

ProChimia Surfaces. ProChimia was chosen as it is owned by one of the authors (B.A.G.), and we thus had full access to the synthetic procedures it used before our optimization.

We first used the deterministic optimization algorithm to identify the optimal synthesis of each molecule individually; the combined cost of these syntheses is denoted c^0_{tot}. We then applied the global optimization procedure to generate a collective synthesis plan for a set of target compounds with a resulting cost $c_{tot} \le c^0_{tot}$. The value of the global optimization approach can then be measured by the fractional savings s, which is defined as $s \equiv (c^0_{tot} - c_{tot})/c^0_{tot}$. For example, for the entire set of 51 compounds and with $c^0_{rxn} = 10$, the individual synthesis cost is $c^0_{tot} =$ US$ 39.6 per gram per target. Global optimization of the collective synthesis results in a synthesis cost of $c^0_{tot} =$ US$ 21.5 per gram per target; that is, to a savings of more than 45%.

Furthermore, the savings increase with the number of targets, as there are more opportunities to exploit common reactions and intermediates (Figure 5 b).[8] To show this, we considered sets of 6, 18, 30, and 42 targets chosen at random from larger set of 51 (see the Supporting Information, Section 4, for explicit structure lists) and identified optimal synthesis plans for each set. As illustrated in Figure 5 b, the percentage savings increased from 25% to 45% as the number of targets increased from 6 to 42. Here, the target compounds are chemically similar and benefit from the collective optimization; however, in general, the degree of savings will depend on the particular targets of interest. For chemically diverse targets (characterized by large synthetic distances[1] on the network), their respective syntheses will remain more independent of one another, in which case the individual deterministic optimization algorithm is expected to perform well.

Figure 5 c provides an illustration of the optimal synthesis plan for the set of 51 ProChimia compounds. Many of these 51 substances, particularly those used for surface chemistry, share some key intermediates (hub compounds) used in their synthesis. For example, many syntheses go through undecylenic bromide ($Br(CH_2)_7CHCH_2$), because alkenes are useful handles to convert into other functional groups, such as thiols and silanes. On the other end of the chain, the haloalkyl functionality serves as an important precursor to such commonly used functional groups as azides, amines, amides, and sulfonates. Interestingly, prior to the optimization, ProChimia did not use this intermediate often but rather

purchased the more expensive C_{10} and C_{11} precursors to its products.

The general conclusion from these considerations is that parallel synthetic optimizations spanning the entire chemical knowledge can generate substantial savings for chemical manufacturers; typically, these savings are expected to scale with the number of products the company produces. In the future, the cost function governing the searches will incorporate chemical process information (reaction fluxes, flow rates, energy usage); these parameters are all easy to modify within our algorithms, which provide a powerful and unprecedented method for optimizing the performance of the chemical industry.

Received: March 20, 2012
Revised: May 18, 2012
Published online: July 13, 2012

Keywords: algorithms · chemical networks · optimization · synthetic methods

[1] K. J. M. Bishop, R. Klajn, B. A. Grzybowski, *Angew. Chem.* **2006**, *118*, 5474 – 5480; *Angew. Chem. Int. Ed.* **2006**, *45*, 5348 – 5354.
[2] M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell, B. A. Grzybowski, *Angew. Chem.* **2005**, *117*, 7429 – 7435; *Angew. Chem. Int. Ed.* **2005**, *44*, 7263 – 7269.
[3] B. A. Grzybowski, K. J. M. Bishop, B. Kowalczyk, C. E. Wilmer, *Nat. Chem.* **2009**, *1*, 31 – 36.
[4] B. Kowalczyk, K. J. M. Bishop, S. K. Smoukov, B. A. Grzybowski, *J. Phys. Org. Chem.* **2009**, *22*, 897 – 902.
[5] a) www.reaxys.com; b) www.prochimia.com.
[6] C. Gothard, S. Soh, N. Gothard, B. Kowalczyk, Yanhu Wei, B. Baytekin, B. A. Grzybowski, *Angew. Chem.* **2012**, *124*, 8046; *Angew. Chem. Int. Ed.* **2012**, *51*, 7922.
[7] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, *Science* **1983**, *220*, 671 – 680.
[8] Details of the intermediates are available upon request from one of the authors (B.A.G.).

