Chemical Networks (3)

DOI: 10.1002/anie.201202210

Chemical Network Algorithms for the Risk Assessment and Management of Chemical Threats**

*Patrick E. Fuller, Chris M. Gothard, Nosheen A. Gothard, Alex Weckiewicz, and Bartosz A. Grzybowski**

Among the millions of known chemical substances, there exist nerve gases, toxic agents, explosives, and many other dangerous compounds. Once proprietary information held by a select few, the synthetic methods leading to these chemical weapons (CWs) have nowadays become readily available, and the internet is home to entire message boards devoted to the nuances of creating nerve gases, such as sarin, VX, and tabun. Although it is sometimes argued that such amateur activities are not worrisome (a large proportion of the poorly trained and ill-equipped terrorists die trying to synthesize nerve gasses), the 1995 attack on the Tokyo metro using sarin gas home-made on the preceding afternoon from commercially available reagents suggests a more cautionary attitude is required. Recognizing the potential threat, the regulatory governmental agencies put forth their best efforts to ban both hazardous chemicals and their immediate precursors. Yet, as we demonstrate herein using the network-theoretical approach, these efforts are only partially effective, and a range of extremely dangerous substances can still be easily made from unregulated and unmonitored reagents. The challenge of identifying and regulating all possible pathways leading to CWs is simply beyond human cognition; it is only through computational searches spanning the entire network of chemical knowledge that we can appropriately identify, rank according to the potential ill-intent (using the elements of game theory), and ultimately eliminate unregulated syntheses of CWs. Network analyses are an unprecedented and powerful method for a watchful management of chemical threats.

The entirety of synthetic-organic knowledge reported in literature comprises on the order of 10^7 reactions and a similar number of compounds. As we have shown in our previous publications,[1–4,5] this dataset can be translated into a giant, highly interconnected network (henceforth, simply the network; Figure 1), in which all the causal synthetic dependencies can be accounted for by using the so-called bipartite graph representation with two types of nodes: those corre-

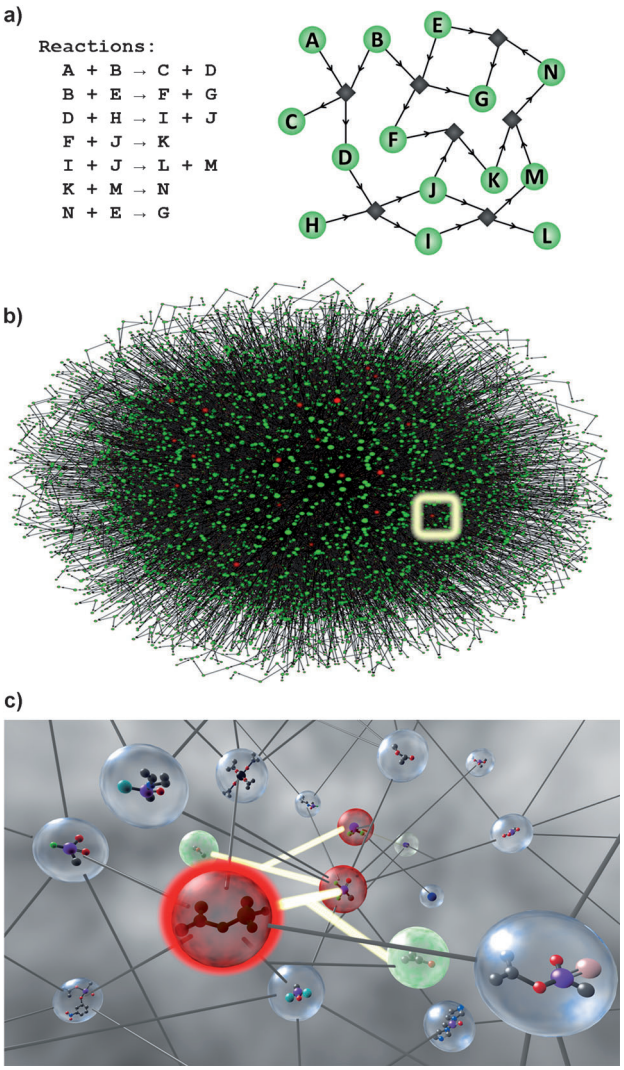

Figure 1. The network of organic chemistry. a) Illustration of the mapping from a list of chemical reactions to a directed, bipartite graph/network. Chemical substances (green circular nodes) and connected to one another by reactions (gray diamonds). This wiring plan captures all necessary synthetic relationships, including the fact that each reaction can have multiple reactants and/or products. b) A small fragment of the network of organic chemistry (here: some 10000 nodes). The entire network is about 7000 times larger than the fragment shown. Green nodes correspond to unregulated substances, whereas red nodes are regulated chemicals. c) An insider view into the network in the immediate vicinity of sarin (bright red node). Other red nodes correspond to regulated precursors; green nodes are the key unregulated substances from which sarin can be made in three steps (see Figure 2b). The connectivity of the environment of sarin is simplified for illustration purposes (for a realistic view of how the maze of connections within the network really looks, see the Supporting Information, Figure S1).

[*] P. Fuller, Dr. C. Gothard, N. Gothard, A. Weckiewicz, 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

[**] This work was supported by the Non-equilibrium Energy Research Center (NERC) at Northwestern, which is an Energy Frontier Research Center funded by the U.S. Department of Energy under Award Number DE-SC0000989.

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/anie.201202210.

sponding to specific substances (green colored dots in Figure 1 a), and those representing the reactions (gray colored diamonds in Figure 1 a). Over a thousand chemicals on this complex network are considered chemical threats; for example, there are 285 chemicals on the US Department of Homeland Security (DHS) Regulated Chemicals List, and the US Environmental Protection Agency (EPA) monitors some 985 substances. Furthermore, chemical weapons violations have spurred international collaboration, with organizations such as the United Nation's Australia Group (AG) working to globally modernize and homogenize regulated chemical substances (currently, AG monitors 48 of the most dangerous precursors). Apart from these "flat" lists, there are also multi-tiered lists, such as the 1993 CW Convention list, which was signed and sponsored by over 180 countries.

Together, these agencies/organizations strive to regulate not only CW target molecules, but also precursors that are indispensable in their syntheses. The key problem in designing the regulatory schemes is, however, that the numbers of pathways on the densely connected network can be astronomical. For example, within five synthetic steps of a typical substance, there are about 10^{19} possible syntheses in which this substance can be made.[5b] Clearly, attempts to scrutinize all these syntheses based on the knowledge of any individual chemist (or even a group of chemists) are destined to be incomplete. At the same time, the numbers to be considered are within reach of modern computers and network search algorithms. Importantly, once the synthetic pathways are enumerated, algorithms originating from game theory can be used 1) to predict what synthetic strategies the terrorists are likely to pursue to make CWs; and 2) to assess whether certain precursors or combinations of precursors are indicative of intent to make CWs.

We first scrutinized the effectiveness of the existing regulatory strategies. To this end, we implemented exhaustive breadth-first searches (BFS) propagating backwards on the network from the dangerous targets T to various depths d (that is, to within d synthetic steps from the target). Figure 2 shows examples of pathways these searches identified, tracing the syntheses back to commercially available and unregulated substances within $d<3$. Figure 2 a has a relatively simple method of making a nerve agent called tabun, which is a colorless, tasteless liquid (or vapor if heated) with a faint fruity odor. The key intermediate in this synthesis is dimethylaminophosphoryl dichloride, which is a regulated substance; however, an inexpensive precursor, bis(dimethylamino)phosphoryl chloride,[6] is not regulated and is commercially available in large quantities. Also, the regulated sodium cyanide precursor can be either made from unregulated calcium carbodiimide or can even be extracted from natural sources (for example, cassava root or apple seeds). Figure 2 b illustrates the synthesis of sarin (note: the algorithm finds even simpler syntheses; see Figure 4 a). Sixteen years after the Tokyo subway terrorist attack, this deadly substance can still be easily made from methylphosphonyl difluoride and isopropanol. Among these precursors, methylphosphonyl difluoride is a regulated substance but can be prepared using unregulated potassium fluoride, dichloronitrosoethane, and dichloro(methyl)phosphine. In the above two examples, the existing regulatory strategies are at least efficient in the very vicinity of the target molecules; this is unfortunately not the case for many other targets. Figure 2 c shows three possible syntheses that in one step produce mustard gas, a CW dating back to World War I, from unregulated precursors. While common synthetic routes exist for sulfur mustard (for example, the Depretz or the Meyer methods), such techniques involve the use of regulated precursors (ethylene and thiodiglycol, respectively). Our algorithm explicitly avoids regulated precursors, favoring easily obtainable substances.

Perhaps the most illustrative and indeed quite worrisome result is shown in Figure 3. Here, the question we asked is

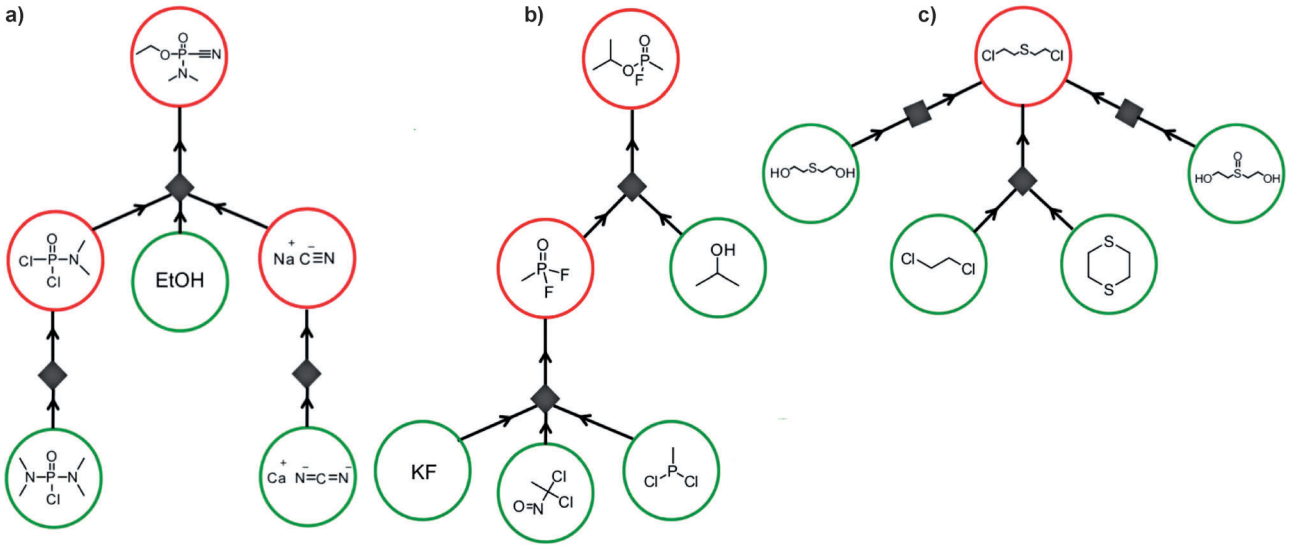

Figure 2. Network searches to identify unregulated synthetic routes to CWs: a) tabun, b) sarin, c) mustard gas. Regulated precursors are circled in red, and unregulated chemicals are circled in green. The pathways shown are only select examples, and the algorithm finds many more short routes to these and other CWs.

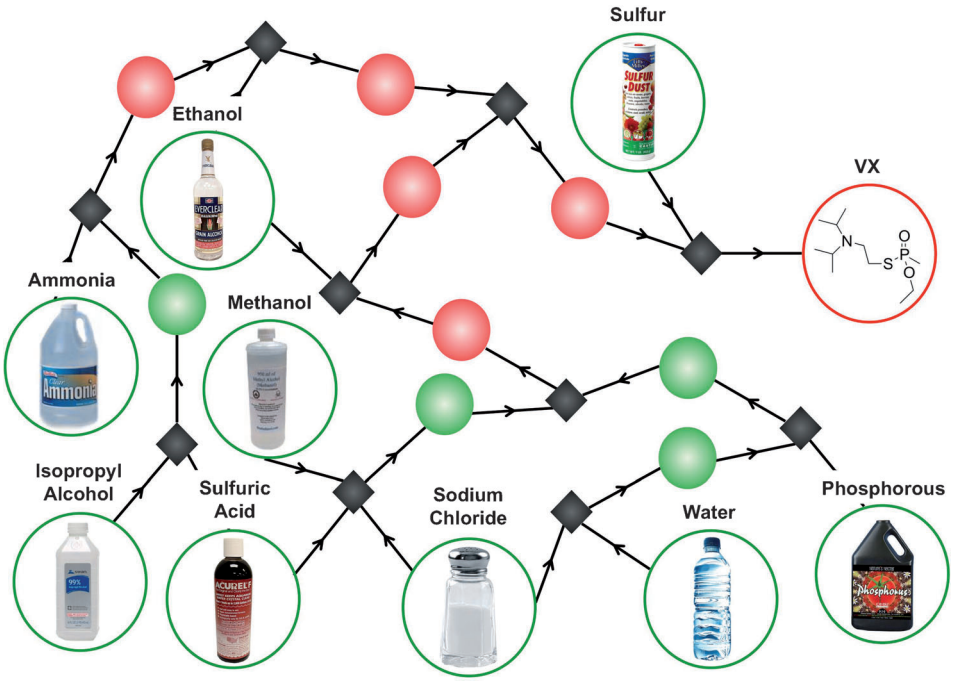

Figure 3. The synthesis of VX starting from household precursors. The details of the syntheses are not provided but are simple enough for a beginner chemist. Red nodes are regulated substances, and green nodes are unregulated substances; gray diamonds are reaction operations.

whether it would be possible to identify a relatively short synthetic route to an extremely toxic VX nerve agent not only from unregulated substances obtained from licensed chemical suppliers (for example, see Ref. [7]), but also from substances that can be freely bought in a local supermarket. The answer is, unfortunately, in the affirmative. The synthesis that our algorithm identified (with the nature of intermediates or specific reactions removed for security reasons) demonstrates that VX can indeed be made from household products as simple as clarifiers, garden fertilizers, and kitchen salt (Figure 3).

The picture that emerges from the BFS analyses above is that even in the very vicinity of the dangerous targets, the regulatory schemes are not bulletproof, and simple unregulated synthetic pathways exist that can be carried out even by amateur chemists. This conclusion does not, however, provide any valuable hints as to more efficient regulatory schemes. Furthermore, we observe that the brute-force approach of regulating more substances would not be effective, not least because it would put additional burden (and the associated bureaucratic cost) on the chemical practice both in academia and industry. Instead, a conceptually different regulatory strategy should be based on a realization that certain unregulated pathways and substances are more likely to be used with terrorist intent.

In this spirit, one of the approaches we developed draws upon the algorithms used to predict, or at least estimate, the

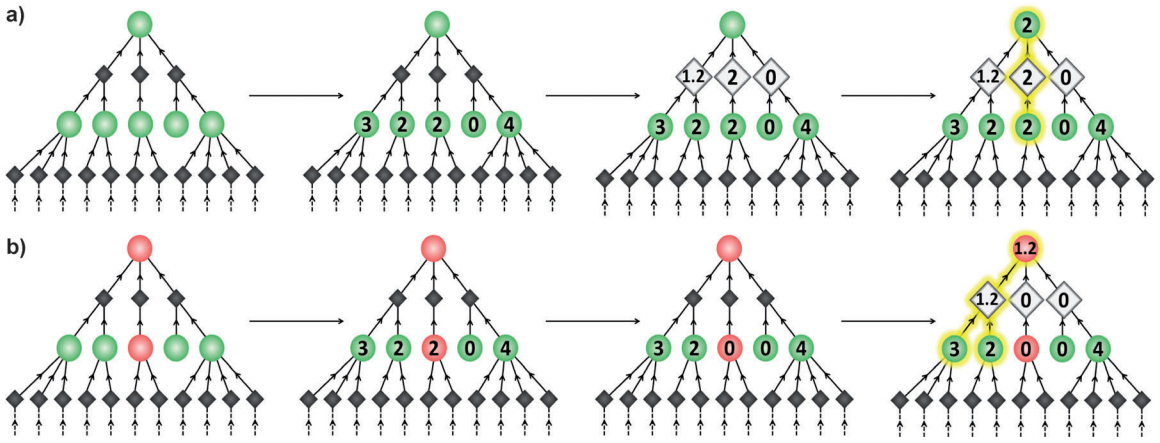

Figure 4. Scoring and algorithmic procedure: a) Explanation of the inverse sum algorithm (over only one reaction step in this case) for synthetic pathways containing only unregulated chemicals. Each chemical *n* steps away from the chemical of interest is scored based on its in-degree centrality, or the number of reactions pointing to a chemical node as a product. These scores are passed upwards on the graph to the parent reaction nodes. If a reaction has multiple reactants, its score is determined by an inverse sum of the scores of each reactant, thus ranking the pathways according to availability of substrates and synthetic complexity. The maximum of these scores is passed to the chemical product node. If this node is the chemical of interest, as it is in this example, the algorithm terminates and the final score is saved. If not, the previous step is repeated recursively. In general, the algorithm is propagated up the graph until the chemical of interest is reached. The resultant pathway, highlighted in yellow, is the highest-scoring unregulated pathway to the target chemical. b) The same example, but with a regulated target and precursors (shown as red circles). Regulated precursor scores are set to a pre-specified value; here this value is chosen as zero, overwriting their in-degree centrality and eliminating pathways involving regulated chemical. Of course, other values can be chosen by the algorithm user to weigh various synthetic possibilities.

outcomes of strategic games such as chess or bridge, where opponents are assumed to make choices maximizing their chances of success. In the context of the present work, synthesis of CWs can be construed as a game of strategy between the regulatory agencies and the potential terrorists: the former strive to ban possible routes, while the latter seek pathways that 1) evade regulated substances and 1) employ only the available/popular and least suspicious chemicals. From a network analysis point of view, a convenient and chemically sound measure of availability and synthetic popularity[1–4] is the so-called in-degree centrality C, which is the number of reactions in which a given substance can be made (that is, the number of reaction "arrows" pointing into a given substance node on the network).

Using this measure, the synthetic pathways leading to CWs can be ranked by an algorithm conceptually similar to the minimax procedure[9] used to rank chess strategies. First, an exhaustive search to depth d is performed to enumerate all synthetic pathways leading to a target substance of interest. This procedure gives synthetic trees, such as those illustrated in Figure 4. Starting at the bottom of the tree, each node is assigned the in-degree centrality value C_i (for example 3, 2, 2, 0, 4 in Figure 4a). If multiple substrates are involved in a specific reaction, their scores are combined using the inverse sum, $C_{\mathrm{rxn},d}^{-1}=\Sigma C_i^{-1}$, and the value of $C_{\mathrm{rxn},d}$ is passed to the reaction node "above". The inverse-summation biases the reaction score toward the least available/most specific chemical required for the reaction. For example, in the rightmost reaction of the synthetic tree in Figure 4, one substrate has the in-degree score $C=0$ and the other $C=4$. The overall ranking $(C_{\mathrm{rxn},d}=(\Sigma C_i^{-1})^{-1}=(0^{-1}+4^{-1})^{-1}=0)$ is, however, determined by the least synthetically popular and thus more suspicious chemical; that is, one that a terrorist who wishes to remain undiscovered is not likely to buy through official vendors. As a result, this reaction has a low score and represents a poor synthetic strategy from the terrorist's point of view. Also, all other things being equal, the inverse summation gives higher scores to reactions that involve less substrates and are therefore less chemically complex (for example, for a reaction requiring one substrate with score $C=2$, the reaction rank is $C_{\mathrm{rxn},d}=2$; if two substrates, each with $C=2$, are required, the reaction score is lower: $C_{\mathrm{rxn},d}=1$). Finally, if the reaction involves a regulated starting material, that is, one that potential terrorists are least likely to obtain, this chemical node is given a penalizing score (for example zero in Figure 4b), which then translates into the low score of the reaction (note: if a regulated substance is an intermediate in a multistep reaction starting from unregulated substances, it is treated as a temporary target at each iteration and assigned a non-zero score through recursion of the inverse-sum algorithm).

After all reactions at a certain level or depth d of the tree are scored, the best terrorist strategy characterized by the maximal value of $C_{\mathrm{rxn},d}$ is chosen and this value is passed to the chemical product node immediately above (that is, at depth $d-1$). The procedure is then repeated until the reaction nodes immediately below to the CW target are reached. Each of these nodes will have a numerical score assigned (for example 1.2, 2, 0 in Figure 4a), and the one with the highest value will represent the synthetic strategy the terrorists are most likely to pursue (highlighted in yellow in Figure 4).

This procedure was tested to rank the synthetic strategies leading to several CWs. The results are illustrated in Figure 5, which shows some top-scoring routes to sarin and mustard gas. While the identities of the individual substances are removed from these graphs for security reasons, the rankings are chemically sound. For example, the highest-ranking route to sarin (score 407) is chosen because the popular isopropyl intermediate with $C=2871$ connections would be easier to obtain by terrorists than the tetraisopropyl silica ester ($C=465$) intermediate in the second-highest scoring pathway. Similarly for the synthesis of mustard gas, reaction of ethylene glycol ($C=93$) with sulfuric acid ($C=359$) provides a more easily obtainable thioether than 2-chloroethanesulfenyl chloride ($C=48$).

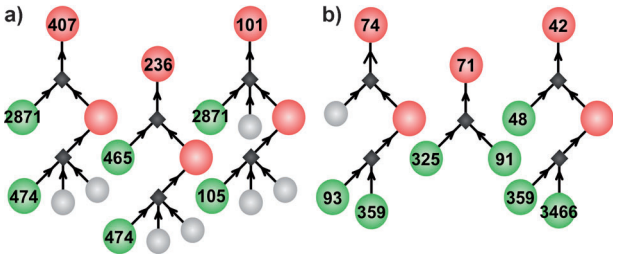

Figure 5. Top-scoring algorithmic routes to chemical weapons: Three high-scoring routes for the synthesis of a) sarin gas and b) sulfur mustard from unregulated precursors (shown in green). The algorithm ranks reactants by their in-degree centrality (synthetic popularity), which reflects the ease of acquisition. Elemental reactants (for example phosphorous) have no centrality measure and are marked as gray nodes. Other algorithms can be also be devised to emphasize ease of synthesis, financial cost, or some combination thereof.

The ranking scheme above replaces monitoring of individual compounds; instead, entire chemical pathways are considered. The key advantage here is that even though every substance a potential terrorist would buy might by itself appear harmless, the combination of several such substances is characteristic of a synthetic route leading to a dangerous target and is thus indicative of terrorist intent. To put it bluntly, no intelligent terrorist will strive to legally obtain immediate, regulated precursors to VX, sarin, tabun, or other CWs. Instead, terrorist groups will likely purchase common chemical precursors that raise no suspicion. The way to detect such activity is then to monitor chemical purchases for characteristic combinations of chemicals. Of course, such strategy should be implemented with as many chemical suppliers as possible such that the synthetic rankings are not evaded by, for example, individuals purchasing different reagents from different companies. Another precaution that should be taken is that monitoring of chemical sales be continuous in time (in this way, the algorithm could flag simultaneous purchases of suspect chemical combinations from different suppliers, by the same or different clients). We are currently working with the U.S. Government agencies on implementing these (and more sophisticated) protocols in the form of an integrated software environment with advanced

data architectures used to maximize search speeds. We note that in addition to performing on-demand individual searches, our software can also memorize the entire rankings of chemical pathways and the suspicious combinations of chemicals; these are then simply stored into a static, readily accessible database. Furthermore, the software (Supporting Information, Figure S2) allows the interactive use of ranking schemes other than the inverse sums we focused on here. Of particular interest are algorithms that further narrow down the searches to the so-called minimal sets of substances that are indicative of CW syntheses.

Looking forward, an interesting extension of the present work is the inverse problem: instead of tracing reactions leading to given targets, the targets are inferred based on substrate chemicals. This problem is relevant to industrial espionage activities; for example, knowing which substrates a given company purchases, outward-going network searches from these substrates can provide suggestions as to what chemicals the company is ultimately producing. These and other algorithms will be described in more detail in the near future. In the meantime, the major conclusion from the current work is that network searches spanning the entire known synthetic chemical knowledge provide a thorough, fast, and cost-effective means of assessing regulatory efficacy. Similar to the internet, which is being constantly searched for such factors as potential threats and hate speech, the activities within the immense network of chemical syntheses should be scanned for potential chemical dangers embodied in publicly available synthetic procedures.

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

Keywords: chemical networks · chemical weapons · minimax algorithm · synthetic methods

[1] B. A. Grzybowski, K. J. M. Bishop, B. Kowalczyk, C. E. Wilmer, *Nat. Chem.* **2009**, *1*, 31 – 36.
[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] K. J. M. Bishop, R. Klajn, B. A. Grzybowski, *Angew. Chem.* **2006**, *118*, 5474 – 5480; *Angew. Chem. Int. Ed.* **2006**, *45*, 5348 – 5354.
[4] B. Kowalczyk, K. J. M. Bishop, S. K. Smoukov, B. A. Grzybowski, *J. Phys. Org. Chem.* **2009**, *22*, 897 – 902.
[5] a) 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; b) M. Kowalik, C. Gothard, A. M. Drews, N. Gothard, A. Weckiewicz, P. E. Fuller, B. A. Grzybowski, K. J. M. Bishop, *Angew. Chem.* **2012**, *124*, 8052; *Angew. Chem. Int. Ed.* **2012**, *51*, 7928.
[6] M. Pianka, B. D. Owen, *J. Appl. Chem.* **1955**, *5*, 525 – 526.
[7] QL, a precursor to the VX gas, can be easily prepared from methyl diethoxyphosphine by transesterification with diisopropylaminoethanol. Methyl diethoxyphosphine can be synthesized from methyl phosphonous dichloride by reacting with ethanol to form the diester. Finally, methyl phosphonous dichloride can be produced by the methylation of phosphorus trichloride.
[8] M. Sion, *Pacif. J. Math.* **1958**, *8*, 171 – 176.
[9] J. Von Neumann, *Math. Annalen.* **1928**, *100*, 295 – 320.

