

The 'wired' universe of organic chemistry

Bartosz A. Grzybowski^{1,2*}, Kyle J. M. Bishop¹, Bartłomiej Kowalczyk¹ and Christopher E. Wilmer¹

The millions of reactions performed and compounds synthesized by organic chemists over the past two centuries connect to form a network larger than the metabolic networks of higher organisms and rivaling the complexity of the World Wide Web. Despite its apparent randomness, the network of chemistry has a well-defined, modular architecture. The network evolves in time according to trends that have not changed since the inception of the discipline, and thus project into chemistry's future. Analysis of organic chemistry using the tools of network theory enables the identification of most 'central' organic molecules, and for the prediction of which and how many molecules will be made in the future. Statistical analyses based on network connectivity are useful in optimizing parallel syntheses, in estimating chemical reactivity, and more.

The synthesis of organic compounds^{1–3} is one of the most important and creative pursuits in modern science, requiring not only technical expertise, but also imagination, intuition and individual judgement. Sometimes, as in the title of Nicolaou's classic review³, chemical synthesis is equated with art, which, by definition, reflects individual imagination and often defies convention, statistics and order. Yet, even if each of us is a truly free-spirited synthetic Bohemian pursuing our respective callings, the universe of chemistry we are collectively creating — one comprising millions upon millions of known reactions and compounds — is surprisingly well-ordered, and its evolution obeys trends that have not changed since the pioneering times of Lavoisier^{4,5}. Here, we analyse this universe using the tools of network theory and topology, and suggest some areas where such a 'global' look at chemistry might offer new insights and lead to industrial applications in parallel optimization of multiple syntheses, security-oriented identification of key precursors to regulated/dangerous substances, and physical organic prediction of chemical reactivity.

In most of these initial demonstrations, we work with a simplified representation of chemistry, whereby the molecules are point-like nodes and the reactions are the arrows connecting them. Of course, there is much more structural information one should be able to derive from the network by considering the specific functional groups, structural motifs and reaction types. Although the analysis at this level of detail is much more computationally demanding and is only in its infancy, it is within the reach of modern computational resources. With the help of computers, we could — for the first time in history — look at chemistry in its entirety and analyse globally how different groups 'travel' together, and which are mutually exclusive in different reactions, which bond types can/cannot be made given the presence of other functionalities, and more. In this spirit, the ultimate goal of the network-based approach should be to quantify the collective knowledge of all those chemists who contributed to the 'wiring' of the existing chemical universe, and to use this knowledge to guide syntheses of new chemicals. This is the grand vision; below are some first, modest attempts to realize it.

On the most abstract level, the millions of known chemicals and reactions constituting organic chemistry can be represented as a complex network⁶, in which compounds correspond to nodes, linked together by reactions through directed connections (Fig. 1a). In our analyses, this network is constructed from data stored in the Crossfire

Beilstein Database (BD, Elsevier Informations Systems). Although BD is not free of omissions (for example, it reports only select types of polymer and is not a comprehensive repository of proteins, DNA, or many important non-covalent organic architectures), it is the largest available collection of organic reactions reported in literature from 1779 to the present. Pruning⁴ this database to remove catalysts, solvents, substances that participate in no reactions, duplicate reactions, and reactions that lack either reactants or products (that is, 'half reactions') leaves us with a universe of known organic chemistry comprising some 6.5 million substances and about 7.0 million reactions connecting them.

Beginning with the entries from the first years of the 1800s, both the numbers of molecules and the numbers of chemical reactions have been increasing exponentially to create a network whose complexity exceeds that of metabolic networks⁷ and rivals that of the World Wide Web^{8–10}. Despite its apparent randomness (Fig. 1b), this network has a well-defined, modular architecture and three distinct regions: the core, the periphery and the islands (Fig. 1c and Fig. 2).

The core is the subset of chemistry defined such that any two of its members can be connected by a synthetic path. The core molecules are structurally diverse, relatively small ($MW_{\text{avg}} = 265 \text{ g mol}^{-1}$ versus $MW_{\text{avg}} = 364 \text{ g mol}^{-1}$ for molecules outside of it), and include many useful synthetic building blocks and important industrial chemicals (of the top 200 (ref. 11), over 70% are found therein). Although they constitute only about 4% of all organic compounds, the core molecules are involved in over 35% of known reactions, and give rise to more than 78% (~5 million) of the known organic universe. Remarkably, an optimized set of as few as 300 core molecules (including chemicals for various functionalization schemes, heterobifunctional reagents, protective-group-introducing agents, important natural products, biological molecules and more; see the list in the Supplementary Information) leads to over 1.5 million other compounds — these 'most-useful' compounds should probably be considered for inclusion in the product line of any fine-chemicals company wishing to cater to the widest chemical clientele possible.

The region in the network outside of the core can be subdivided into a large 'periphery', containing molecules that can be synthesized from the core's substrates, and into smaller, isolated 'islands', not reachable from the core. The periphery is rather loosely wired (on average ~2.3 connections per molecule) but constitutes about 78% of chemistry, most of which is in close proximity to the core. Indeed,

¹Department of Chemical and Biological Engineering, and ²Department of Chemistry, Northwestern University, 2145 Sheridan Road, Evanston, Illinois 60208, USA. *e-mail: grzybor@northwestern.edu

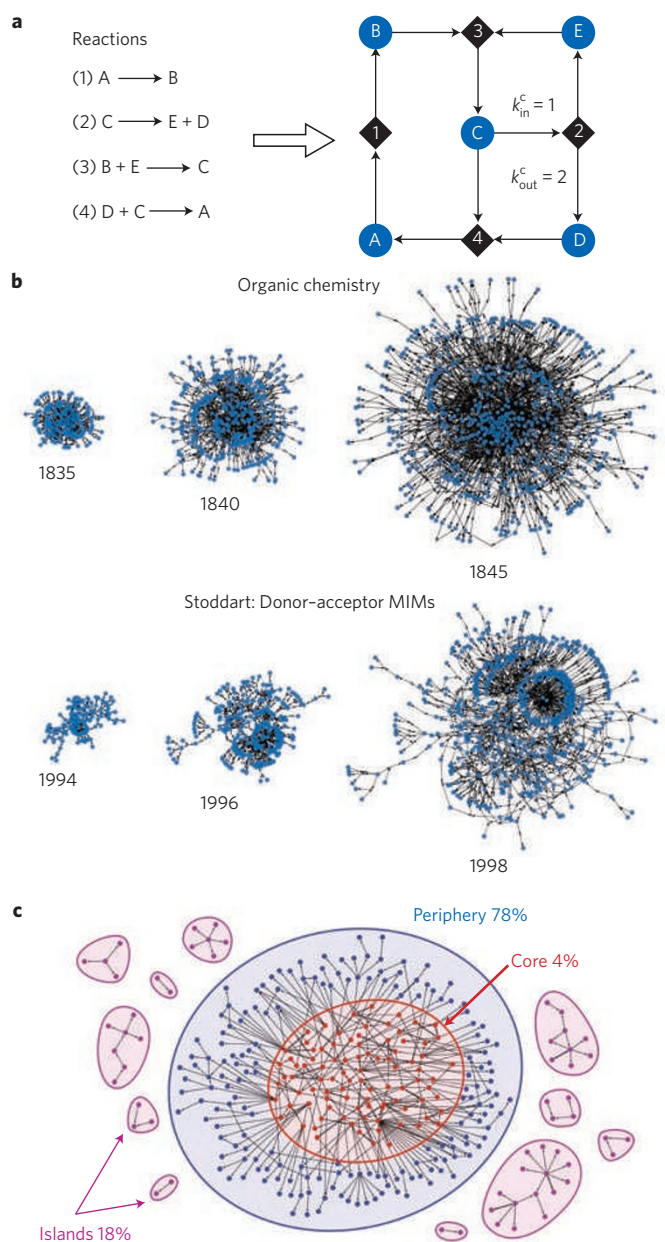


Figure 1 | Construction and architecture of the network of organic chemistry. **a**, Illustration of the conversion of a collection of chemical reactions into a directed-network (so-called bipartite) representation. Chemical compounds A–E correspond to ‘substance nodes’ (blue circles) within the network and are connected through ‘reaction nodes’ (black diamonds) via directed connections. The connectivity of each node is described by the number of incoming arrows, k_{in} (for example, $k_{in} = 1$ for compound C) and the number of outgoing arrows, k_{out} (for example, $k_{out} = 2$ for C). **b**, The network of chemistry (top) in the years 1835, 1840 and 1845: chemistry is growing exponentially and doubles in size every ~17 years. Below is a small subset of chemistry illustrating the growing contribution of an individual chemist — here, J. Fraser Stoddart and his work on mechanically interlocked molecules (MIMs)^{24,25}. Owing to chemistry’s exponential growth and scale-free architecture, the sub-network of Stoddart’s chemistry is self-similar to that of chemistry at large and comparable in size to the known chemical universe of 1845. **c**, Major topological components of the chemistry network: the core (red), the periphery (blue) and the islands (pink). The specific network shown here corresponds to the state of organic chemistry in the year 1840. Today, the network is approximately 10,000 times larger, but its key features are similar.

the average distance from the core to any molecule in the periphery is only three steps, with 95% of the peripheral substances lying within seven steps from the core. Moving away from the core, the average mass/complexity¹² of molecules increases linearly with distance (measured in synthetic steps; also see Fig. 2a) before levelling off to just over 700 g mol⁻¹ after 15 steps (that is, after reaching over 99% of the periphery).

Finally, unconnected to the core/periphery are the network’s ‘islands’, which are typically small (less than four molecules on average) but together constitute about 18% of the network. The most connected molecules in each island are usually either complex natural products or specialized substances (for example, non-natural isotopes). While some islands reflect imperfections of the database and its failure to report the existing syntheses connecting island molecules to the rest of chemistry, a sizeable fraction corresponds to substances that are difficult synthetic targets whose total syntheses have not yet been reported despite numerous attempts (see examples in Fig. 2b). When looking for a challenging synthetic target, one should consider searching the islands (the curious reader might want to examine the list of 100 island molecules we include in the Supplementary Information; the algorithms used to generate this list and to identify even more islands are described in detail in ref. 5).

Within the general framework above, the architecture of the network is further characterized by local connectivity measures — in particular, by the numbers of reaction arrows ‘emanating’ from each node, k_{out} (that is, the number of times a given molecule was used as a reaction substrate), and by the numbers of reaction arrows ‘pointing’ towards the nodes, k_{in} (that is, the number of times a molecule was used as reaction product; Fig. 1a). Counting these connectivities for all molecules, one can then plot the frequencies, $P(k_{out})$ and $P(k_{in})$ with which molecules of given k_{out} or k_{in} connectivity appear in the network. The end result of this operation is illustrated in Fig. 3a, which shows that these frequencies decay algebraically as $P(k_{out}) \propto k_{out}^{-\gamma_{out}}$ and $P(k_{in}) \propto k_{in}^{-\gamma_{in}}$.

Although this scaling might not seem very illuminating, it actually tells us that chemistry has the so-called scale-free structure⁶ similar to that of the World Wide Web^{8–10}, the Internet¹³, metabolic networks⁷ and even societies^{14,15}. This scale-free architecture is akin to a fractal in the sense that the structural/connectivity motifs characterizing the entire network repeat themselves in all of its sub-networks. Another distinguishing feature of being scale-free is the presence of highly connected ‘hub’ molecules directly analogous to the hubs of the airline system (Atlanta, Chicago, London, Frankfurt and so on) facilitating transportation from one poorly connected airport to another. Likewise, in organic chemistry, the synthesis of one molecule from another by a series of chemical transformations will probably use one or more of these versatile hub compounds as intermediates (as we shall see later, the use of hubs has an added economic advantage, as these compounds are significantly less expensive than poorly connected ones). Also, the fact that the scale-free structure is conserved as the network evolves in time indicates that it grows by the mechanism of preferential attachment, whereby highly connected substances are more likely to participate in new reactions than poorly connected compounds. The more times a molecule has been used as a synthetic substrate (that is, the larger its k_{out}) in the past, the higher the chances it will be used again in the future; similarly, the higher its k_{in} , the more likely it is that chemists will try to make it by a new reaction. Colloquially speaking, molecular ‘celebrities’ are becoming ever more popular (Fig. 3b).

Of course, molecules are more than connected ‘dots’ in a network and are characterized by detailed structures and an array of physicochemical properties. Although analysis of the millions of molecular structures requires computational power beyond our current reach (at least for our group’s resources), there are some simple molecular descriptors (mass, degree of unsaturation, number of stereocentres and so on) that can be analysed relatively

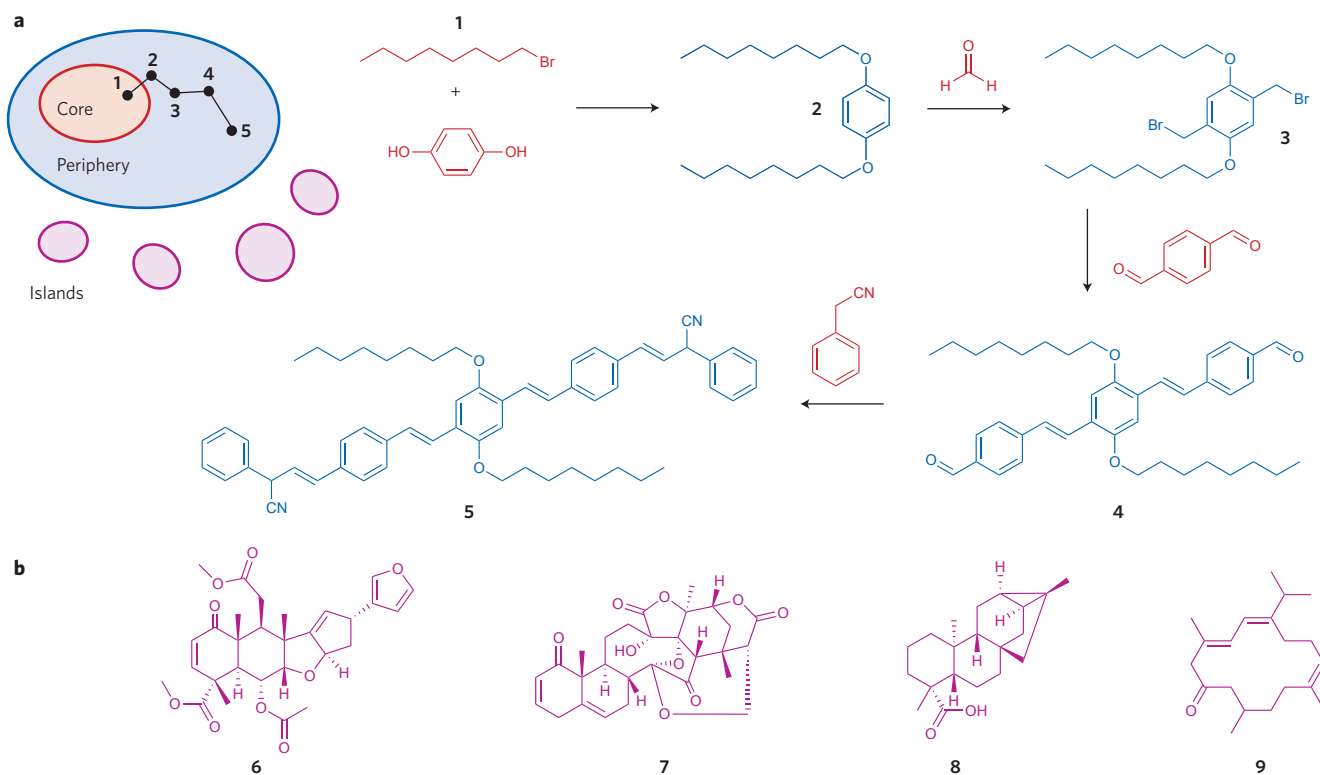


Figure 2 | Examples of synthetic pathways and synthetically challenging molecules identified by network analysis. **a**, A representative reaction pathway leading from the simple core substances (red) to peripheral molecules (blue) of increasing complexity. **b**, Examples of island molecules that have not yet been synthesized from simple precursors in the 'core' or 'periphery' and thus represent potential 'challenges' for future syntheses. The specific natural products shown are (**6**) nimbin, first isolated in 1942 from the seed oil of *Melia azadirachta*, (**7**) an antileukemic and antimycobacterial agent from *Physalis angulata*, (**8**) Ent-12,16-cycloauran-19-oic acid from *Helianthus annuus*, and (**9**) a diterpenoid from the Caribbean gorgonian *Eunicea calyculata*. Beilstein registry numbers for 100 more island molecules are included in the Supplementary Information.

easily. Analysis of molecular masses offers some interesting insights. For example, despite the enormous progress in the synthetic methodology since the times of Hofmann and Perkin, the most commonly used substrates and products remain those of molecular weights $MW_{\text{subst}} \approx 150 \text{ g mol}^{-1}$ and $MW_{\text{prod}} \approx 250 \text{ g mol}^{-1}$, respectively (Fig. 3c). Moreover, as these most popular substrates/products correspond to the most rapidly connecting nodes of the network⁴, the preferential attachment mechanism discussed previously predicts that these substances will remain the most popular ones in the future.

A related observation is that the shapes of the mass distributions in Fig. 3c do not change with time but only shift 'upwards'. To an astute mathematician, this self-similarity is an indication that the creation of new masses/molecules is based on some iterative ('self-repeating') growth process. Indeed, stochastic modelling (see ref. 4 for details) of how the molecular masses in the network of chemistry evolve allows one to back-track this process and, after some voluminous mathematics, reduce it to a surprisingly simple relationship between the masses of reaction substrates, m_s , and products, $m_p = am_s + b$, where a and b are stochastic (that is, drawn from an appropriate probability distribution) variables with mean values 0.67 and 180 g mol^{-1} , respectively. In interpreting this stochastic equation, it should always be remembered that it is statistical in nature — that is, it might not work for a specific reaction, but its accuracy improves with increasing numbers of reactions considered. Although such a statistical law might be of no value to an individual chemist working on a specific reaction, its long-term and/or large-scale predictability can benefit the chemical industry in cases such as combinatorial synthesis, where the mass-evolution equation can predict the distributions of masses in compound libraries from the masses

of substrates used to create these libraries (see the CombiChem example described in ref. 4).

Although there are many more interesting regularities in the network of chemistry (for example, even masses are 40% more likely to be made than odd masses; drugs have the same mass distribution as all known chemicals, industrially important molecules have different 'wiring' patterns than unimportant ones and so on), the key issue is whether these curiosities can be of any practical value to chemists. We believe that even at the present, simplified stage of analysis — that is, with point-like reactions and molecules connected by arrows — several useful observations can be made.

The first example deals with the optimization of multiple reactions and is addressed to industrial chemists. Suppose a speciality-chemicals company produces P products. A relevant question one might ask is then what set of substrates, S , and reaction pathways should the company use to minimize its overall production cost (Fig. 4a)? Mathematically, this question is equivalent to finding a set S that minimizes the cost function, C_{tot} , represented as a sum of the costs of reagents and all other 'labour' costs, $C_{\text{tot}} = C_{\text{subst}} + C_{\text{labour}}$. Denoting the price of substrate i as s_i , the total number of reactions as N_{rxn} , and the average cost of performing one reaction as α , we can then write the total cost function as $C_{\text{tot}} = \sum_i s_i + \alpha N_{\text{rxn}}$. The link between this general formulation and the architecture of the network is the correlation between the cost of a substrate and its local network connectivity. Analysis of specific substances reveals⁴ that synthetically popular substances are less expensive than poorly connected ones, with the cost per mole being proportional to the inverse square root of the molecule's network connectivity, $s_i \approx \beta/\sqrt{k}$, where β is a constant. Using this cost relation, stochastic search algorithms (based on simulated annealing Monte Carlo optimization¹⁶) can be back-propagated from

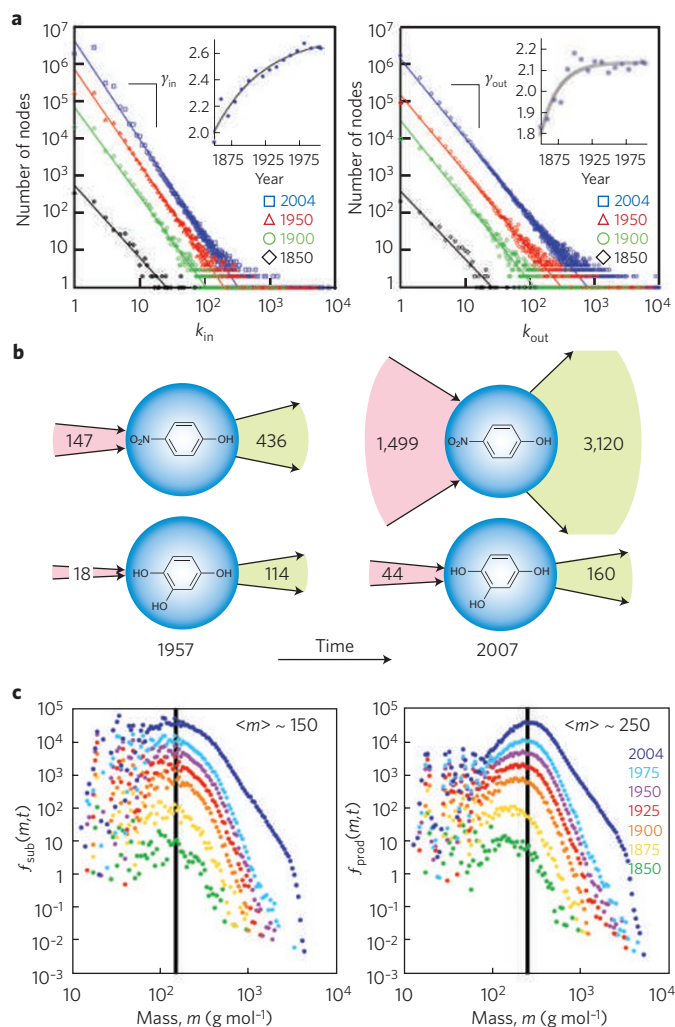


Figure 3 | The universe of organic chemistry is a scale-free network evolving according to the mechanism of preferential attachment. **a**, In- and out-connectivity distributions of molecules in the network of organic chemistry obey a power law $P(k) \sim k^{-\gamma}$ characteristic of scale-free networks. From 1850 to the present, the power-law exponents γ_{in} and γ_{out} (insets) have been steadily growing and approach constant values of approximately 2.7 and 2.1, respectively — nearly identical to those that characterize the World Wide Web (2.71 and 2.1, respectively). **b**, Schematic illustration of the preferential attachment mechanism responsible for chemistry's scale-free topology. 'Popular' molecules such as *p*-nitrophenol with $k_{out} + k_{in} = 583$ total connections in 1957 become even more popular in time, increasing to 4,619 total connections in 2007. Meanwhile, less popular molecules such as 1,2,4-trihydroxybenzene incorporate new connections less rapidly, growing from 132 to only 204 over the same period. **c**, Frequency distributions of masses of molecules that were used as substrates ($f_{sub}(m,t)$, left) and products ($f_{prod}(m,t)$, right) in reactions reported in 25-year intervals between 1850 and 2004. The masses of most popular substrates and most popular products correspond to the maxima of the curves and have not changed perceptibly since the times of Kekulé. Parts **a** and **c** reproduced with permission from ref. 4, © 2005 Wiley.

the products to find optimal substrates, which minimize the total production costs, C_{tot} . Although this approach does find optimal and economical solutions (as verified in practice for some small companies), it is a somewhat standard exercise in combinatorial optimization on a network. More intriguing are the universal trends that hold for different companies and for different labour costs, as characterized by the dimensionless parameter $\chi = \alpha/\beta$.

Figure 4b illustrates the optimization analysis of two chemical companies: ProChimia Poland, providing specialized surface-modification reagents for self-assembled monolayers (~150 products of low average network connectivity, approximately 4), and Toronto Research Chemicals (TRC), offering some 10,000 popular chemicals (average network connectivity 37). For ProChimia's relatively complex and poorly connected products, cheaper substrates may always be found by increasing the number of synthetic steps. Although using cheaper substrates requires more work to make the products, the optimized pathways are such that the relative overall labour versus substrates costs for the company remain constant (around 70% in labour and 30% in substrates) irrespective of the unitary labour cost χ (Fig. 4b, left). One of the authors of this article who is involved in ProChimia (B.A.G.) finds this reassuring as in the real balance sheet the labour-to-substrates ratio for the years 2001–2008 has remained between 63:37 and 72:28, suggesting that ProChimia manages to operate optimally despite significant changes in labour costs in Poland. For TRC, the division of costs is not expected to be so robust. For this company's relatively simple and well-connected products, it is no longer possible to find significantly cheaper substrates by simply expanding the 'net' farther from the products. Thus, as the relative cost of labour decreases, that of substrates increases and ultimately surpasses the former (Fig. 4b, right). This scaling should be characteristic of most large chemical companies, and published reports for Dow Chemical (we do not have access to TRC's financial reports) reveal that as much as 80% of their expenditure is in raw materials. Although the above approach is admittedly simplistic (for example, operating with a heuristic cost function and neglecting the differences in molar fluxes along various pathways), it illustrates how chemistry's scale-free architecture, governing the connectivity (and implicitly the costs) of chemical substances, can be relevant to the economics of the chemical industry.

The next example illustrates the usefulness of the network approach in identifying synthetic routes to controlled substances (for example, narcotic and psychotropic drugs) and chemical weapons. In theory, these substances and their precursors are tightly regulated. In practice, a simple breadth-first search within the network of chemistry rapidly discovers how subjective and incomplete the lists of such 'chemicals of interest' can be. For instance, although sarin gas (used in the Tokyo subway terrorist attack in 1995) and its immediate precursors are strictly regulated, it can still be synthesized — 13 years after the Tokyo tragedy — in only two reaction steps starting from commercially available and unregulated substrates (Fig. 4c, left). Part of the problem in regulating these and other substrates is that they might be highly popular and useful in many benign syntheses. On the other hand, even taking the strictest measures and regulating extremely popular precursors does not necessarily solve the problem, as illustrated by the synthesis of the anaesthetic drug and hallucinogen ('angel dust'), phencyclidine (PCP; Fig. 4c, right). Here, the US government regulates the immediate precursors and also piperidine, which is a popular synthon in the retrosynthetic analysis of many alkaloids, drugs, and other substances prone to abuse (for example, morphine, heroin, LSD). Unfortunately, this restriction not only puts extra burden on many well-wishing chemists (piperidine is also a substrate in ~16,200 'benign' reactions), but, even worse, it does not prevent the synthesis of PCP by a three-step reaction starting from unregulated and commercially available chemicals (see the leftmost highlighted route in Fig. 4c, right). A much more efficient regulatory strategy would be to monitor the combined inventories of chemical suppliers for the purchases of 'red-flag sets' of precursors that signal the intent to make dangerous substances. In the specific PCP example, the government should be alerted not if one buys piperidine alone, but when the person/company/organization acquires at least two out of three key substances (in Fig. 4c, circled by a dashed line). Although in this simple example an experienced

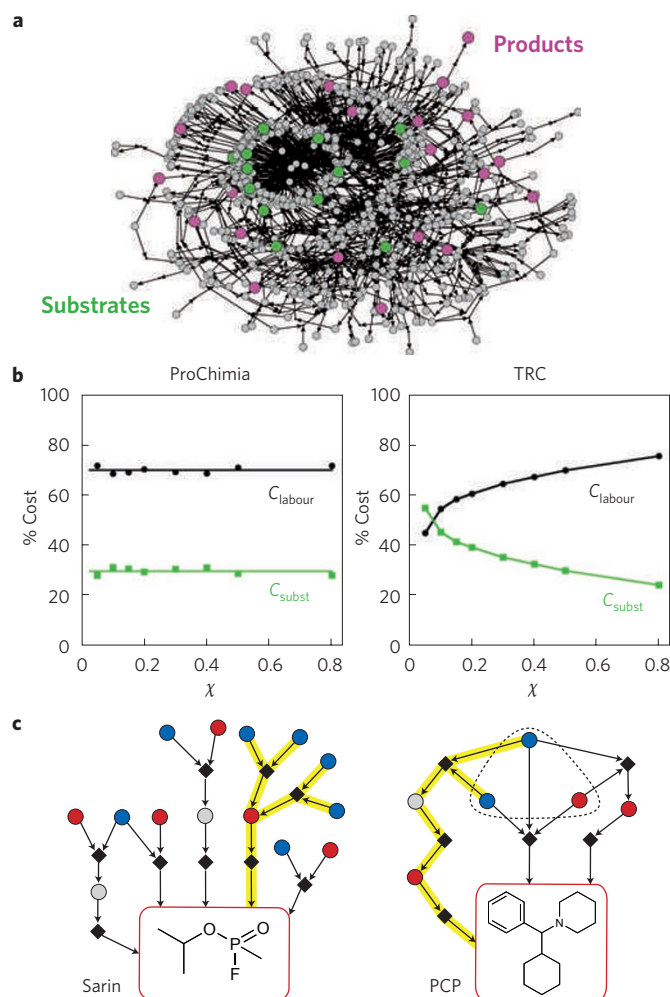


Figure 4 | Network-based optimization of multiple syntheses and monitoring of restricted substances. **a**, Schematic illustration of the network optimization problem: For a given set of products (pink nodes) find the set of substrates (green nodes) that minimize the overall production cost. **b**, The overall percentage cost due to labour and substrates for optimal synthetic pathways for ProChimia and TRC. The horizontal axis has the unitary ('per reaction') cost of labour, χ . For ProChimia — assuming its syntheses are optimized — the division of the overall costs does not depend on χ . For TRC, the relative substrate costs increase steadily with decreasing χ . For more descriptors of the optimal pathways, see Supplementary Fig. S2. **c**, Two examples of network analysis for monitoring dangerous substances. Compounds coloured red are restricted by the Department of Homeland Security and/or by the Drug Enforcement Agency (DEA). Unrestricted and commercially available chemicals are represented by blue markers. Compounds coloured grey are unrestricted but are not commercially available. Black diamonds indicate a reaction 'operation' (see Fig. 1a for the definition of this 'bipartite' notation). Left: A subset of the reaction network surrounding sarin. All syntheses starting from restricted compounds can be eliminated easily with current monitoring/restriction protocols. Unfortunately, sarin can be made readily in two steps from unrestricted and commercially available substances (in two different ways, reaction paths highlighted in yellow). Right: A subset of the reaction network leading to phencyclidine (PCP). Network analysis identifies three distinct routes to PCP. Two routes requiring the use of piperidine (a monitored substance corresponding to a red circle within the set encircled by the dashed line) can be eliminated using the current DEA screening methods. Still, PCP can be made in three steps from commercially available substances (pathway highlighted in yellow). To eliminate any possibility of synthesis, DEA should monitor the purchases of any two of the three encircled substances.

specialist at the Department of Homeland Security could come to the same conclusion without recourse to network analysis, the problem of set identification for the majority of dangerous substances is well beyond human cognition, and the use of parallel network-search algorithms is clearly a superior approach.

While these and other algorithms are already being developed for applications, the effort to interpret structural information contained in the network is only just beginning. Even at this early stage, some useful trends appear at the level of local network connectivity, where structural analyses relate synthetic popularity to molecular reactivity. To illustrate this concept, consider aromatic electrophilic substitutions of monosubstituted benzenes. Imagine that for a molecule $(C_6H_5)_X$ such substitutions in the *ortho* position have been reported (that is, found in the network) n_o times, those in the *meta*- position n_m times, and in the *para* position n_p times. Do these numbers reflect more than the popularity of specific reactions among chemists? They do. Normalizing the reactions counts by $n_o + n_m + n_p$ and plotting against frontier orbital populations¹⁷ reveals that the two measures of reactivity correlate strongly ($r = 0.9$) for different substituents X (see Supplementary Fig. S1). The same is also true for other reaction types and extends to cases for which relative reactivities of families of molecules need to be compared (for example, reaction counts correlate with Hammett constants, σ). In short, synthetic popularity reflects molecular reactivity, and straightforward counting of reaction 'arrows' in the network can lead to accurate reactivity measures. With hindsight, this result is not entirely unexpected if one notes that every reaction carried out is, in essence, an experiment in chemical reactivity. In this spirit, a successful reaction reported in the literature reflects the thermodynamic/kinetic feasibility of similar chemical transformations.

Finally, there are fascinating questions to answer if one dissects the molecules in the network's nodes into component structural motifs (rings, chains, functional groups, stereocentres) and analyses globally how they 'travel' on the network. Here, reactions may be considered as 'operators' that transform one collection of connected functional groups into another — an approach first introduced by the late Ivar Ugi in the 1970s and 1980s within the framework of his transformation matrices^{18,19}. Unfortunately, Ugi had precious little computational power to attack this problem on a global scale. Today, this power is within our grasp and we should soon be able to analyse chemistry exhaustively to find the cross-correlations between different motifs, determine which functionalities are mutually exclusive (for example, protecting groups), which characteristics distinguish 'creative' syntheses from typical ones, and more.

We close with some general remarks. First is that the trends derived from the known universe of organic chemistry might and should be expected. As organic chemistry uses the same few hundred reaction types to transform several hundred typical functional groups, the millions of syntheses performed and the millions of molecules made exhibit statistical regularities derived from the sequential application of relatively few chemical 'operators'. Provocatively said, the statistical laws governing chemistry at large — like death and taxation — are certain and unavoidable. At the same time, it is conceivable that these regularities might change provided that conceptually new types of chemistries are developed. There are already some candidates: rotaxanes^{20–22} and catenanes^{20,23} based on the so-called mechanical bond^{24,25} are one, and they are already creating their own sub-universe of chemistry (see Fig. 1b); non-covalent architectures are also there and need to be explored. Unfortunately, the current way in which non-covalent structures are represented in chemical databases prevents meaningful topological analysis of these new 'galaxies' within the chemical universe — consequently, new search algorithms able to recognize not only synthetic connections but also architectonic similarities need to be developed. From a more practical and optimistic perspective, the global view of chemistry advocated here may soon provide a valuable

tool for practising chemists. The ability to explore algorithmically the millions of potential synthetic routes along the network means that chemists may be able to select optimal pathways not only based on their own knowledge/intuition, but also on the collective knowledge of past generations of chemists. After all, this collective knowledge is what constitutes the 'wired' network of chemistry.

References

1. Tietze, L. F. & Beifuss, U. Sequential transformations in organic chemistry — a synthetic strategy with a future. *Angew. Chem. Int. Ed. Engl.* **32**, 131–163 (1993).
2. Corey, E. J. & Cheng, X.-M. *The Logic of Chemical Synthesis* (Wiley-Interscience, New York, 1995).
3. Nicolaou, K. C., Vourloumis, D., Winssinger, N. & Baran, P. S. The art and science of total synthesis at the dawn of the twenty-first century. *Angew. Chem. Int. Ed.* **39**, 44–122 (2000).
4. Fialkowski, M., Bishop, K. J. M., Chubukov, V. A., Campbell, C. J. & Grzybowski, B. A. Architecture and evolution of organic chemistry. *Angew. Chem. Int. Ed.* **44**, 7263–7269 (2005).
5. Bishop, K. J. M., Klajn, R. & Grzybowski, B. A. The core and most useful molecules in organic chemistry. *Angew. Chem. Int. Ed.* **45**, 5348–5354 (2006).
6. Albert, R. & Barabasi, A. L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).
7. Jeong, H., Tombor, B., Albert, R., Oltval, Z. N. & Barabasi, A. L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
8. Albert, R., Jeong, H. & Barabasi, A. L. Diameter of the World-Wide Web. *Nature* **401**, 130–131 (1999).
9. Broder, A. *et al.* Graph structure in the Web. *Comput. Netw.* **33**, 309–320 (2000).
10. Amaral, L. A. N. & Ottino, J. M. Complex networks — Augmenting the framework for the study of complex systems. *Eur. Phys. J. B* **38**, 147–162 (2004).
11. Chemical Market Reporter: Chemical prices, 7 March 2005 <<http://www.chemicalmarketreporter.com>>.
12. Allu, T. K. & Oprea, T. I. Rapid evaluation of synthetic and molecular complexity for in silico chemistry. *J. Chem. Inf. Model.* **45**, 1237–1243 (2005).
13. Faloutsos, M., Faloutsos, P. & Faloutsos, C. On power-law relationships of the internet topology. *Comput. Commun. Rev.* 251–262 (1999).
14. Redner, S. How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. J. B* **4**, 131–134 (1998).
15. Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E. & Aberg, Y. The web of human sexual contacts. *Nature* **411**, 907–908 (2001).
16. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
17. Fukui, K. & Fujimoto, H. *Frontier Orbitals and Reaction Paths: Selected Papers of Kenichi Fukui* (World Scientific, Singapore, 1997).
18. Ugi, I. *et al.* Computer-assistance in the design of syntheses and a new generation of computer-programs for the solution of chemical problems by molecular logic. *Pure Appl. Chem.* **60**, 1573–1586 (1988).
19. Ugi, I. *et al.* Computer-assisted solution of chemical problems — the historical development and the present state-of-the-art of a new discipline of chemistry. *Angew. Chem. Int. Ed. Engl.* **32**, 201–227 (1993).
20. Anelli, P. L. *et al.* Molecular Meccano 1. [2]rotaxanes and a [2]catenane made to order. *J. Am. Chem. Soc.* **114**, 193–218 (1992).
21. Bissell, R. A., Cordova, E., Kaifer, A. E. & Stoddart, J. F. A chemically and electrochemically switchable molecular shuttle. *Nature* **369**, 133–137 (1994).
22. Iijima, T. *et al.* Controllable donor-acceptor neutral [2]rotaxanes. *Chem.-Eur. J.* **10**, 6375–6392 (2004).
23. Asakawa, M. *et al.* A chemically and electrochemically switchable [2]catenane incorporating a tetrathiafulvalene unit. *Angew. Chem. Int. Ed.* **37**, 333–337 (1998).
24. Amabilino, D. B. & Stoddart, J. F. Interlocked and intertwined structures and superstructures. *Chem. Rev.* **95**, 2725–2828 (1995).
25. Stoddart, J. F. & Colquhoun, H. M. Big and little Meccano. *Tetrahedron* **64**, 8231–8263 (2008).

Additional information

Supplementary information accompanies this paper at www.nature.com/naturechemistry.