Organic Chemistry

DOI: 10.1002/anie.200600881

The Core and Most Useful Molecules in Organic Chemistry**

*Kyle J. M. Bishop, Rafal Klajn, and Bartosz A. Grzybowski**

On the most abstract level, the millions of known chemicals and reactions constituting organic chemistry can be represented as a complex network,[1] in which compounds correspond to nodes and reactions to directed connections between these nodes. We have recently shown[2] that such a network has a scale-free topology similar to that of the World Wide Web and that by analyzing its time evolution, it is possible to derive statistical laws that describe and also predict how and which types of molecules are/will be synthesized. The major limitation of this statistical approach is that it provides information only about the average properties of molecules and the overall network structure—at the same time, it does not tell us how specific chemicals are arranged or what roles they play in the network of chemistry. Herein, we address these important issues with the aim of identifying molecules

[*] K. J. M. Bishop, R. Klajn, Prof. Dr. B. A. Grzybowski
Department of Chemical and Biological Engineering and Northwestern Institute of Complexity
Northwestern University
2145 Sheridan Rd., Evanston, IL 60208 (USA)
Fax: (+1) 847-491-3728
E-mail: grzybor@northwestern.edu

[**] B.A.G. gratefully acknowledges financial support from the Camille and Henry Dreyfus New Faculty Awards Program, the NSF CAREER, and the 3M Awards. K.J.M.B. was supported by an NSF graduate fellowship.

Supporting information for this article is available on the WWW under http://www.angewandte.org or from the author.

central to and most useful in organic synthesis. By using mathematical tools from network theory and statistical physics, we demonstrate that 1) there exists a small set of strongly connected, chemically diverse substances (the "core") from which the majority of other known organic compounds (the "periphery") can be made in three or fewer synthetic steps, and that 2) this central structure is surrounded by small "islands" that do not connect either to the core or to the periphery. Furthermore, by defining the usefulness of a compound as proportional to the number of other compounds that can be made from it, we develop Monte Carlo (MC) search algorithms to identify small optimal sets of maximally useful chemicals. Aside from purely scientific interest, the knowledge of such "most useful" collections of compounds should be of practical value to specialty chemical companies, which could use it to optimize product selection and cater for the most diverse group of chemical customers.

Analysis was performed on data stored in the Beilstein Database (BD),[3] which is the largest—albeit not without omissions—repository of organic reactions reported in the literature from 1779 to the present. The data were pruned as described previously[2] to remove catalysts, solvents, substances that participated in no reactions, and reactions that lacked either reactants or products (that is, "half reactions"). Duplicate reactions, that is, reactions with identical reactants and products, were considered only once and were characterized by the date of the earliest publication. This selection procedure left 5.9 million substances and 6.5 million reactions.[4]

The final dataset was converted into a directed network by connecting reactants and products by directed edges (Figure 1a). We implemented and tested several wiring schemes to ascertain that the results of our analysis were not artifacts of the network construction, but captured the true architecture of organic chemistry. The most condensed representation was one in which only the heaviest reactant and the heaviest product of each reaction were retained and connected by a single directed edge. This so-called one-to-one representation is chemically the most intuitive, and it reflects the fact that small reagents usually play only an auxiliary role when transforming larger molecules (Figure 1a, top). At the same time, it excluded some real (albeit rare) connections in the network—for example, in cases when an auxiliary substance (such as a protective group) is more massive than the primary reactant, or when the lighter of two products is the actual synthetic target. These connections were accounted for in an all-inclusive (but clearly redundant) scheme, in which all reactants were connected to all products irrespective of mass (all-to-all; Figure 1a, bottom). Finally, intermediate schemes were also examined in which reactants and products within a given percentage of the mass of the heaviest reactant/product in a reaction were retained and connected (Figure 1a, middle). As all schemes gave qualitatively similar

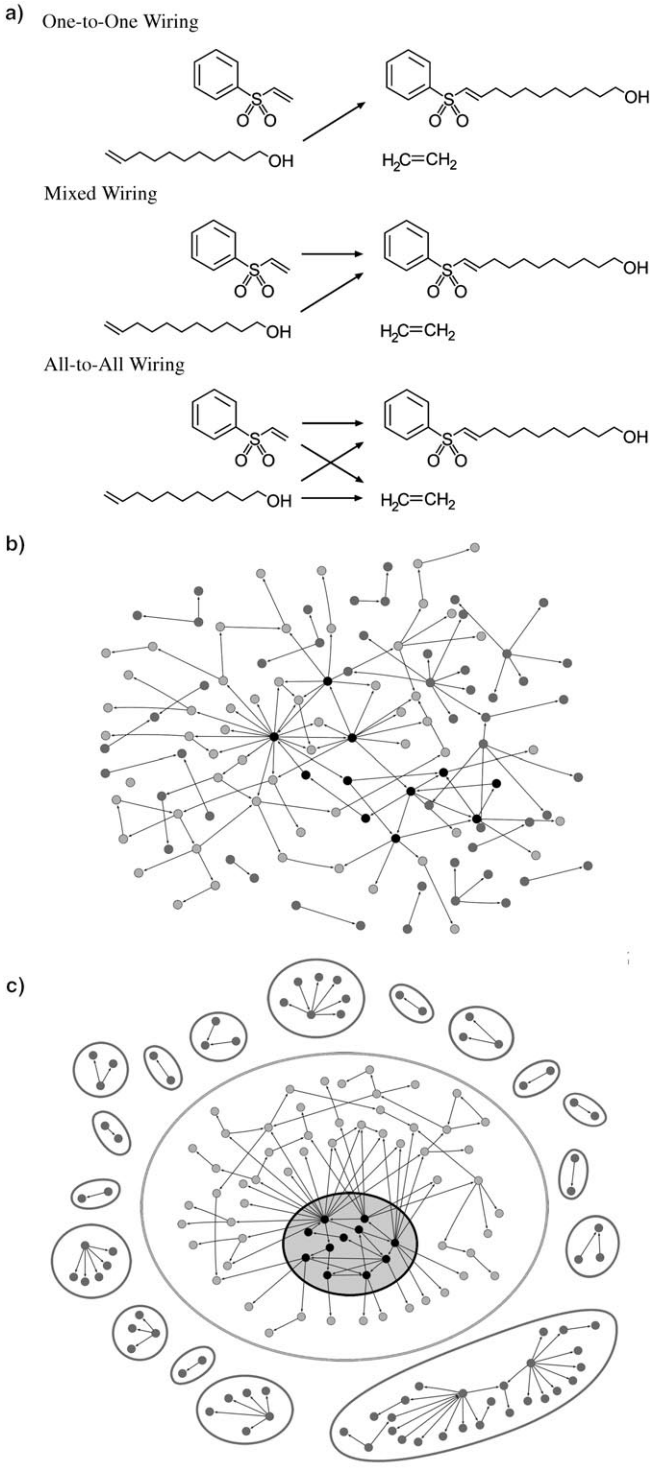

Figure 1. Wiring schemes and network structure. a) Illustration of the three possible wiring schemes as applied to a ruthenium-catalyzed olefin cross-metathesis reaction.[21] In this example, the "mixed-wiring" scheme includes products/reactants within 5% of the heaviest reactant/product; because the molecular weight of phenyl vinyl sulfone is within 2% of that of 10-undecen-1-ol, it is included. On the other hand, ethylene is 91% lighter than (*E*)-11-(phenylsulfonyl)-10-undecen-1-ol and is not wired. b) Schematic representation of the organic-chemistry network based on the reactions reported before 1840. Even with the nodes appropriately color-coded, it is hard to decipher their global roles in this simple network. c) Analysis based on the algorithms described in the text and in the Supporting Information allows identification of the network's major topological components: the core (black), the periphery (light gray), and the islands (dark gray). Although the current (year 2004) organic-chemistry network is several tens of thousands times larger, its qualitative features are similar.

results, we chose to base our further discussion on the simplest one-to-one network and use other representations only when they illustrate some specific points more clearly.

To identify the core of organic chemistry, we first searched for all strongly connected components (SCCs) present in the network. Mathematically, an SCC of a directed graph is defined as a maximal set of nodes (herein, molecules) such that there exists a path (that is, a sequence of reactions) between any two of these nodes.[5] In our case, an SCC corresponds to a tightly knit "cluster" of molecules all connected by synthetic pathways.[6] Searches performed according to a depth-first search (DFS) algorithm (see the Supporting Information for definitions and algorithmic details) identified all possible SCCs, whose size distribution obeys—with one exception—a power-law relation $f(x) \approx x^{-3.8}$, where $f(x)$ is the frequency of observation of an SCC of size x (Figure 2a). The exception itself is a very notable one and corresponds to an SCC that is thousands of times larger (206993 substances) than the next-largest cluster (79 substances). Although this so-called "giant" SCC contains only 3.5% of all organic substances, these substances are usually more highly connected, are involved in over 35% of known reactions, and give rise to more than 60% ($\approx$3.6 million) of known organic chemicals.[7] For these reasons, this collection of compounds can be considered the "core" of organic chemistry.

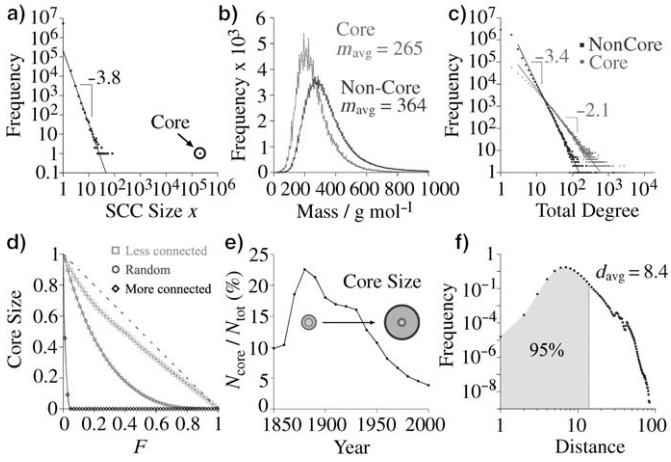

Figure 2. Properties of the core of the organic-chemistry network. The negative numbers give the slopes of the best-fit straight lines. a) Size distribution of the SCCs present in the network. The core corresponds to a giant SCC that is more than three orders of magnitude larger than the next-largest cluster. b) Normalized mass distributions for molecules found in the core and for those outside it. c) Distributions of the total degree (that is, the number of both incoming and outgoing connections) position of the bracketed text of molecules for core and noncore substances. d) Relative size of the core as a function of the fraction F of nodes (and their associated connections) removed. The rate of core "disintegration" depends on the types of nodes that are removed; for example, least-connected nodes (□), highest-connected nodes (◇), or nodes chosen at random (○). e) Evolution of the relative core size from 1850 to 2000. Over the past 125 years, the periphery has been growing larger at the expense of the core. f) Normalized frequency distribution for the synthetic distance (that is, the number of reaction steps) between any two nodes in the core. The average distance between any two molecules in the core is 8.4, and 95% of connections are fewer than 15 steps.

The core has several important characteristics. First, the molecules it contains are on average significantly lighter (average molecular weight, $MW_{avg} = 265\ g\,mol^{-1}$) than those outside it ($MW_{avg} = 364\ g\,mol^{-1}$; Figure 2b), an observation that is perhaps not surprising since, as already stated, the core molecules are precursors of more complex/massive[8] compounds. Second, the core is not simply a collection of highest-connected nodes in the network. Although it does contain the vast majority of such nodes, it also has many poorly connected chemicals (for example, 50% of the nodes have fewer than five connections; Figure 2c). The role of these chemicals is to keep the "traffic" throughout the core open, so that one can connect any two of its members. Third, all of the compounds in the core matter for its integrity and tight-knit structure, in the sense that the removal of any of them shrinks and fragments the core rapidly. Had some molecules been unimportant/redundant, their hypothetical removal would reduce the core size (defined by the number of substances it comprises, N_{core}) only linearly; in other words, if the fraction F of the nodes was removed, N_{core} would decrease by $\Delta N_{core} = N_{core} F$. In reality (Figure 2d), the core shrinks and fragments faster than linearly with F, even if only the lowest-connected nodes are removed.

Although the sizes of both the core and the entire network (N_{tot}) have been growing exponentially with time, the relative core size (N_{core}/N_{tot}) has evolved nonmonotonically (Figure 2e). Until the turn of the twentieth century it slowly increased until, in 1880, it reached 22% of the then known chemical (structural) space; subsequently, it has been steadily decreasing to the present 3.5%. It appears that once chemists have learned how to make and wire the key core molecules, the emphasis has been increasingly on the exploration of the new structural space.

For such an exploration from the core into the periphery to be efficient, the former should contain a diverse set of building blocks—and, indeed, it does. The diversity can be studied either by inspecting the molecules directly (see the Supporting Information) or by topological measures. The latter approach is based on the observation that if two molecules can be connected by a short synthetic pathway (for example, one or two steps) they are likely to be structurally similar (but see reference [9]). If the distance is long, the resemblance between the final and the initial molecules should be, on average, minimal. By performing a set of breadth-first searches (BFSs) starting at each node in the core, we found that the average synthetic distance between core molecules is nine steps and that 95% of all paths are fewer than 15 steps (Figure 2f). While this number might not look high, it is three times larger than the average number of steps needed to make any of the millions of compounds reachable from the core (see below).

We conclude our discussion of the core's properties by noting that it contains the majority of the most important industrial chemicals—of the top 200,[10] over 70% (145) are found in the core. To summarize, the core is compact, wired, yet chemically diverse, and contains important chemicals from which the majority of known organic substances can be made.

The region in the network outside the core can be subdivided into a large periphery, which contains molecules

that can be synthesized from core substrates, and smaller isolated islands not reachable from the core. Although quantitative details of the subdivision depend on the network-wiring scheme, it can be estimated that the periphery contains at least 3.6 million chemicals, and the islands not more than 1.2 million compounds. A lower bound on the number of island compounds is 270000, as estimated using the all-to-all wiring scheme, which allows identification of islands that are not reachable even by ineffective or "unchemical" reactions, and are thus truly synthetically isolated.

The average distance from the core to any molecule in the periphery is only three steps, and 95% of the peripheral substances lie within seven steps from the core (Figure 3a,b). As one moves away from the core, the average mass and/or complexity[8] of molecules increase linearly with distance (measured in synthetic steps) before leveling off to just over 700 g mol^{-1} after 15 steps (that is, after reaching over 99% of the periphery; Figure 3c). The average "in" and "out" connectivities of compounds in the periphery are both close to unity regardless of their distance from the core. At the same time, there are some molecules that have significant "out" branching, while most molecules (86%) have only one incoming connection (Figure 3d). This property implies that the periphery has a treelike structure (illustrated in Figure 3a). We note that the trees emanating from the various regions of the core do overlap (Figure 3e), thus creating a nontrivial problem of identifying optimal sets of core molecules (see below).

Unconnected to the core/periphery are the molecules in the islands. The islands are generally small (fewer than four molecules on average), and their most connected "hubs" usually correspond to either complex natural products or

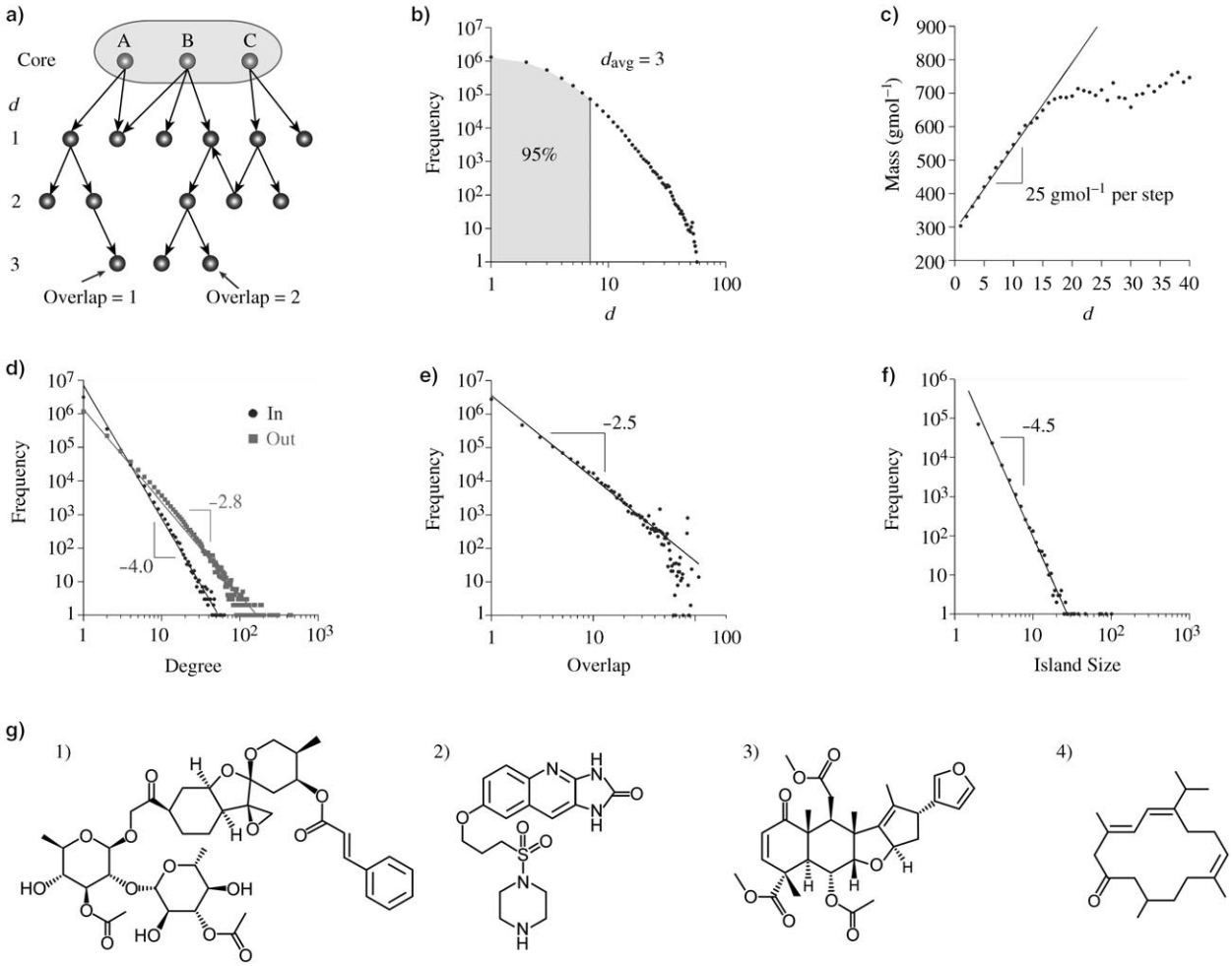

Figure 3. Properties of the network's periphery and islands. a) Schematic illustration of the core (neglecting internal connections) and periphery. The distance *d* from the core represents the shortest synthetic pathway from any node in the core to a given node in the periphery. The "overlap" of a molecule in the periphery is defined as the number of substrates in the core from which the given molecule can be synthesized. b) Number of node in the periphery as a function of their distance from the core. c) Average molecular mass of compounds in the periphery as a function of their distance from the core. d) In and out degree distributions for all molecules in the periphery. In degree is the number of reactions in which a molecule is a product; out degree gives the number of reactions in which a molecule is a substrate. e) Distribution of synthetic overlap for all molecules in the periphery. f) Size distribution of islands found in the all-to-all wired network. The average island includes fewer than three molecules. g) Examples of highly connected island molecules, whose total syntheses have either been reported—1) phyllanthoside[11,12], 2) a phosphodiesterase inhibitor[11,12]—or not described in the literature—3) nimbin[13,14], 4) a diterpenoid from the Caribbean gorgonian *E. calycu-lata*.[16].

specialized substances (for example, non-natural isotopes). We wish to emphasize that the fact that these compounds are found in isolated parts of the network does not necessarily mean that they *cannot* be made from the core/periphery substances, but rather that they have *not yet* been connected to them. This could be either 1) because pertinent reactions have not yet been tried and/or deposited in the database, or 2) because achieving a connection to the core is difficult enough to have thwarted previous attempts.

The first scenario is scientifically trivial and reflects the imperfections of the database. Illustrative examples include phyllanthoside, (Figure 3g 1; the 11th most connected hub of the island molecules) and a phosphodiesterase inhibitor (Figure 3g 2; the 81st most connected hub of the island molecules), whose total syntheses from core substances have been reported[11,12] but are not included in the BD. The second scenario is more interesting, as it suggests that an island compound is a challenging synthetic target. For instance, while both nimbin[13,14] (a complex terpenoid from *Melia azadirachta* and an ecdysone 20-monooxygenase inhibitor;[15] Figure 3g 3) and a diterpenoid from the Caribbean gorgonian *Eunicea calyculata*[16] (Figure 3g 4) have been subjects of several chemical studies, no total syntheses of these compounds have been reported.[17]

Aside from satisfying purely scientific curiosity, knowledge of the network's topological structure can be of practical value. Consider a company that manufactures specialty chemicals. While it probably could not supply all the 206 993 core compounds, it might be interested in optimizing its limited product line (say, comprising a few hundred compounds) in such a way that the compounds it sells would allow the making of a maximal number of other chemicals. We have developed a simulated annealing MC algorithm that allows such "most useful" sets of molecules to be found.

Briefly, an initial set of desired size M is first chosen from among the most connected core molecules. We note that such a trial solution based on "local" connectivity is not an optimal one, as the trees propagating from the chosen compounds into the periphery often overlap (see Figure 3e). The optimization procedure is based on the Metropolis algorithm frequently used in statistical physics[18,19] for finding global energetic minima of ensembles with very many degrees of freedom. In our case, the energy is replaced by the "usefulness" function defined as $U=(N_{max}-N(M))/N_{max}$, where $N(M)$ is the number of compounds reached from a given set of size M, and N_{max} is the total number of substances in the periphery. With this formulation, U ranges from zero to one and, importantly, is minimal for the optimal set of core substances. For a given set, U is calculated by performing a DFS from each of its nodes into the periphery (see the Supporting Information for details).

At each MC move, a molecule from the current set is exchanged at random, and the values of U for the old and the new sets are compared. If the move is favorable (that is, $\Delta U = U_{new}-U_{old}<0$), it is accepted unconditionally. If, on the other hand, $\Delta U>0$, it is accepted with conditional Boltzmann probability $\exp(-\Delta U/T)$, where T is a control parameter analogous to the "temperature" of the system. During the search, parameter T is decreased exponentially according to $T=T_I\exp(-\log(T_I/T_F)(s/S))$, where T_I and T_F are the initial and final "temperatures", respectively, s is the number of moves taken thus far, and S is the total number of moves in the MC run (typically, millions to tens of millions). This so-called simulated annealing procedure[19] allows the exploration of maximally diverse chemical space in the early stages of the search (when T is high), while preventing the "escape" from the optimal solution ultimately found at low values of T.

Simulations based on this algorithm allowed the identification of most useful sets of various sizes (Figure 4a). The

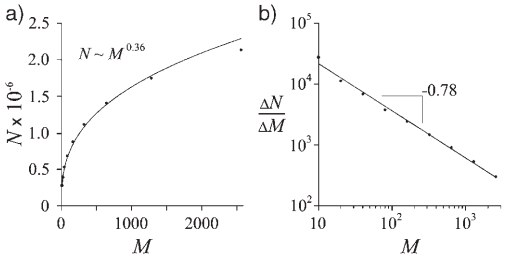

Figure 4. Optimal sets of most useful molecules. a) The number of nodes N reachable from an optimal set of core nodes as a function of the set size M. b) Closely connected to the relationship in (a), the gain rate—defined as the number of new nodes reached (ΔN) per increase in the set size (ΔM)—decreases according to a power law with increasing set size M. This relationship could help chemical companies to find an optimal balance between the costs (likely scales with M) and revenues (likely scales with N) associated with their product line.

number of substances (N) that can be made from M core molecules increases as $N(M)\approx M^{0.36}$. Irrespective of their size, the sets contained diverse and widely applicable chemicals. For example, the set of size $M=300$ (our estimate of a typical offer of a small chemical company) contained common chemicals for various functionalization schemes, heterobifunctional reagents, protective-group-introducing agents, important natural products, biologicals, and more. The specific categories are summarized in Table 1 and the accompanying structures are included in the Supporting Information. We suggest that these most useful compounds should certainly be considered for inclusion in the product line of any fine-chemicals company.[20] In addition, we note that the list reveals certain interesting facts about the functionalities that organic chemists prefer to use in their syntheses.

For instance, although incorporation of a benzoyl moiety can be accomplished through a variety of functional groups (for example, aldehyde, acyl chloride, carboxy, nitrile, anhydride), aromatic aldehydes are the most abundant group of the top-300 core molecules (21 hits) used for this task. This finding can be rationalized based on the fact that benzaldehydes are reactive yet stable. In contrast, the reactivity of benzoic acids, benzoic anhydrides, and nitriles (3, 3, and 0 hits, respectively) is considerably lower, while benzoyl chlorides (8 hits) are too reactive and thus harder to handle. A similar, common-chemical-sense example is that of alkyl halides, which are represented in the core by bromides (9 hits;

Table 1: Categorization of the optimal set of 300 core substrates.

Compounds[a]	Count	List no.
simple aromatic compounds (32%)		
aromatic aldehydes (+ *3 in category* **)	18	1–18
anilines (+ *1 in* **)	12	19–30
benzoyl chlorides	8	31–38
benzoic acids and their derivatives (+ *1 in* **)	7	39–45
phenols (+ *2 in* **)	6	46–51
α-haloacetophenones	6	52–57
benzyl halides	5	58–62
sulfonyl chlorides	5	63–67
unsubstituted aromatic compounds	4	68–71
chloronitrobenzenes	4	72–75
acetophenones	4	76–79
other reactive di- or polysubstituted benzenes**	4	80–83
other monosubstituted benzenes	13	84–96
simple aliphatic compounds (27%)		
small, bifunctional molecules	13	97–109
simple reagents useful in organic synthesis	12	110–121
very simple small molecules	11	122–132
n-bromoalkanes	9	133–141
aliphatic and heterocyclic amines	7	142–148
α,ω-dibromoalkanes	6	149–154
acyl chlorides	6	155–160
anhydrides	6	161–166
alkyl iodides	5	167–171
other bromides	3	172–174
orthoesters	2	175–176
specialized reagents (23%)		
1,3-dioxoanions or their precursors	16	177–192
protective-group introduction: silyl (8), trityl (3)	11	193–203
Wittig reagents or their precursors	10	204–213
Grignard reagents	6	214–219
α-halo/pseudohaloesters	6	220–225
chlorocarbonates	5	226–230
Michael acceptors	4	231–234
phosphonic chlorides, chlorophosphines	4	235–238
quinones	3	239–241
reagents for peptide synthesis	3	242–244
natural compounds and their building blocks (7%)		
carbohydrates and protected carbohydrates	9	245–253
protected amino acids	6	254–259
DNA building blocks	2	260–261
others (antibiotics, alkaloids, steroids, neurotransmitters)	4	262–265
simple inorganic compounds (2%)	5	266–270
others (10%)	30	271–300

[a] The two stars indicate the category other reactive di- or polysubstituted benzenes.

reactive yet stable) and iodides (5 hits; too reactive), whereas there are 0 hits for chlorides (weakly reactive). Finally, we observe that although nearly all of the "top" compounds are achiral (with the notable exception of natural products), many of them are prochiral (for example, benzaldehydes, 1,3-dicarbonyl compounds) and well-suited for the introduction of chirality through enantioselective addition/substitution reactions.

In summary, we have identified and analyzed major substructures within the network of organic chemical reactions, and shown that the analysis of this network can aid the selection of sets of most useful synthetic substrates. This study and the one preceding it probably exhaust the capabilities of our purely topological approach; the uncovering of more interesting facts about the structure of chemistry will require additional (and downloadable) data, such as detailed structural information, reaction stoichiometries, and yields. With such information at hand, one could look into particular subfields of organic chemistry, select the most efficient pathways, and maybe even draw conclusions about reaction thermodynamics by comparing/studying various cycles/loops in the network. Such analyses, however, will have to rely on databases that report more details about the substances and reactions one wishes to study. We hope that with the rapid progress of information technology such databases will be available in the near future.

Received: March 7, 2006
Revised: May 11, 2006
Published online: July 11, 2006

Keywords: history of science · network theory · organic chemistry · synthetic methods

[1] R. Albert, A. L. Barabasi, *Rev. Mod. Phys.* **2002**, *74*, 47.
[2] M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell, B. A. Grzybowski, *Angew. Chem.* **2005**, *117*, 7429; *Angew. Chem. Int. Ed.* **2005**, *44*, 7263.
[3] MDL Crossfire Beilstein Database (see the Supporting Information for details).
[4] Stoichiometry and reaction yields were not considered as they were reported for only a few percent of database entries.
[5] A. V. Aho, J. E. Hopcroft, J. D. Ullman, *Data Structures and Algorithms*, Addison-Wesley, Reading, MA, **1983**.

[6] Note the importance of directed connections; an undirected network would have no chemical meaning, and the core (that is, the giant SCC) could not be defined.
[7] With the most permissive all-to-all scheme, the percentage is 80%; with the 20-% mixed scheme it is 70%.
[8] Molecular "complexity" scales with molecular mass; see: T. K. Allu, T. I. Oprea, *J. Chem. Inf. Model.* **2005**, *45*, 1237.
[9] We stress that this is true only in a statistical sense, and that ingenious counterexamples can be found where only one synthetic step changes the molecular skeleton extensively (see the elegant tandem and domino reactions shown in the Supporting Information).
[10] March 7, 2005 chemical prices, Chemical Market Reporter, http://www.chemicalmarketreporter.com.
[11] N. A. Meanwell, P. Hewawasam, J. A. Thomas, J. J. K. Wright, J. W. Russell, M. Gamberdella, H. J. Goldenberg, S. M. Seiler, G. B. Zavoico, *J. Med. Chem.* **1993**, *36*, 3251.
[12] A. B. Smith, M. Fukui, H. A. Vaccaro, J. R. Empfield, *J. Am. Chem. Soc.* **1991**, *113*, 2071.
[13] V. Kabaleeswaran, S. S. Rajan, G. Gopalakrishnan, G. Suresh, T. R. Govindachari, *J. Chem. Crystallogr.* **1997**, *27*, 731.
[14] N. S. Narasimhan, *Chem. Ind.* **1957**, 661.
[15] M. J. Mitchell, S. L. Smith, S. Johnson, E. D. Morgan, *Arch. Insect Biochem. Physiol.* **1997**, *35*, 199.
[16] J. H. Shin, W. Fenical, *J. Org. Chem.* **1991**, *56*, 1227.
[17] This fact was verified by cross-checking other literature sources including ChemAbstract and SciFinder.
[18] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller, *J. Chem. Phys.* **1953**, *21*, 1087.
[19] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, *Science* **1983**, *220*, 671.
[20] Inclusion in the "optimal" set is based on the usefulness alone and does not take into account the chemical diversity among the chosen molecules; it is thus possible that some similar chemicals are included (see the Supporting Information). We suggest that for practical/industrial applications, the optimal sets of a few hundred molecules obtained from the usefulness analysis might be further scrutinized for chemical diversity. Alternatively, the usefulness function itself could be modified to simultaneously maximize both the usefulness and the diversity. Such a procedure, however, would be significantly more time-consuming as it would require the analysis of structural details of all molecules in the network.
[21] A. Michrowska, M. Bieniek, M. Kim, R. Klajn, K. Grela, *Tetrahedron* **2003**, *59*, 4525.

