Computational Chemistry

DOI: 10.1002/ange.200502272

Architecture and Evolution of Organic Chemistry**

*Marcin Fialkowski, Kyle J. M. Bishop, Victor A. Chubukov, Christopher J. Campbell, and Bartosz A. Grzybowski**

For almost two centuries, chemists all over the world have applied their expertise and creativity[1–5] to the synthesis of new molecules. Since each individual chemist—or a collaborating group of chemists—tries to select unique synthetic targets[6–9] and come up with a maximally original and/or efficient method of making them, it might appear that the activities of such independent "agents" should be largely uncorrelated, and that no generalizations about the evolution of chemistry *en large* could be made. As we show here, however, there exist several statistical laws that describe how molecules are made and interconverted. We analyze organic synthesis at the level of an abstract network representation whereby molecules correspond to nodes characterized by molecular masses, and reactions to directed edges connecting these nodes (Figure 1a). We show, among others, that the connections between the nodes form a time-evolving scale free network[10–14] of structure similar to that of the world wide web (WWW),[12,15,16] and that masses of molecules in this network are governed by a single stochastic process.[17,18] Aside from fundamental interest, the trends we identify allow making predictions of potential economical impact for the chemical industry, for example, how many molecules will be synthesized in the future, molecules of which molecular

[*] Dr. M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell, Prof. Dr. B. A. Grzybowski
Department of Chemical and Biological Engineering and Northwestern Institute of Complexity
Northwestern University
2145 Sheridan Road, 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. K.J.M.B. and C.J.C. were supported in part by the NSF-IGERT program "Dynamics of Complex Systems in Science and Engineering" (DGE-9987577).

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

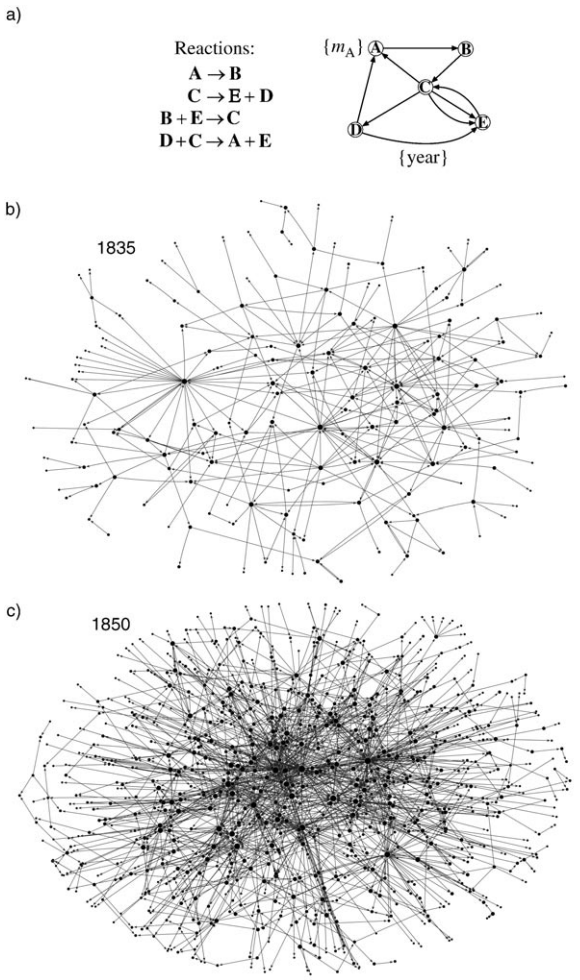

Figure 1. a) Illustration of the conversion of a collection of chemical reactions into a directed-graph (network) representation. Chemical compounds **A**–**E** correspond to the nodes of the network characterized by molecular masses, *m*. Directed edges are assigned for a given reaction by connecting all reactants to all products. The connectivity of each node is described by the number of incoming arrows, k_{in} (e.g., $k_{in}(C) = 2$), and the number of outgoing arrows, k_{out} (e.g., $k_{out}(C) = 4$). The edges are characterized by the earliest publication date of the corresponding reaction. Note that multiple edges—of the same or different directions—can connect two nodes if these nodes are involved in more than one chemical reaction (e.g., **C** and **E** above). b) Visual representation of the giant connected component (GCC)—that is, the largest subset of nodes that are connected to each other—of the organic-chemistry network in 1835. This network contains only 176 compounds; the size of a node corresponds to its total connectivity. Note the scale-free architecture in which most nodes connect to others through highly connected "hub" molecules. c) The GCC in 1850 contains 867 compounds. While the scale-free architecture remains unchanged, the complexity of the network has increased substantially within the elapsed 15 years. One can only imagine the network of organic chemistry in 2004, which contains approximately six million compounds! (The graphs were created using Pajek, a freely available software package for the visualization of large networks.)

masses are more likely to be made or used as substrates, how the network connectivities can be used to assess a molecule's industrial importance, and how the equation describing evolution of masses could help in designing fragment libraries for combinatorial chemistry.

Analysis was performed on data stored in the Beilstein database (BD),[19] which is the largest repository of organic reactions, containing (up to April 2004) 9550398 chemical substances and 9293250 reactions in which these substances participate. In choosing BD, we adopted its well-established criterion for the classification of chemical substances as "organic" and its comprehensive coverage of the chemical literature dating back to 1779 (see the Supporting Information and reference [19] for details of BD). Although we stress that BD is not without omissions, it provides the single, most-complete description of organic chemistry and its evolution. Therefore, it seems reasonable to assume that statistical laws derived from data described therein are indeed representative of organic chemistry in general, and we present them as such.

In the translation of organic synthesis into a network of chemical connectivity, each node represented a chemical compound characterized by its molecular mass (99.7% of the compounds had mass data). Substances that participated in no reactions or acted only as catalysts or solvents were excluded from the network, thereby decreasing the number of "active" nodes to 5957807. Reactions between these chemicals were used to assign directed edges, each characterized by the year in which the reaction was published (99.7% of the reactions had date information). Duplicate reactions—that is, reactions with identical reactants and products—were considered only once and characterized by the date of the earliest reaction. Reactions that lacked either reactants or products (that is, "half reactions") were not included in the network. Stoichiometry and reaction yields were not considered, as they were reported for only a few percent of database entries. After parsing, the total number of reactions was decreased to 6539158. In the conversion of the set of chemical reactions into a directed network, all reactants were connected to all products by directed edges, as illustrated in Figure 1a (see Supporting Information).

The starting point of our analysis was 1850, and since then both the numbers of molecules and the numbers of chemical reactions have increased exponentially. The corresponding growth rates were constant within time periods up to the beginning of the twentieth century ($r_m = 0.083\ \text{year}^{-1}$ for molecules and $r_r = 0.087\ \text{year}^{-1}$ for reactions) and afterwards ($r_m = 0.044\ \text{year}^{-1}$ and $r_r = 0.038\ \text{year}^{-1}$; Figure 2a). At the same time, the average connectivity between molecules (Figure 2b)—defined as the number of edges divided by the number of nodes—initially increased, reached a maximum by about 1885, and then steadily decreased to the value of approximately 2 in 2004. It appears that the early days of chemistry were dominated by "wiring" existing molecules (presumably, to perfect/optimize known synthetic methodologies); when these methodologies matured, exploration of unknown structural space became a dominant activity.

To establish the topological characteristics of the expanding network of chemical reactions, we analyzed distributions of the numbers of connections outgoing from a given node, k_{out} (i.e., the number of times a given molecule was used as a reaction substrate), and the number of connections incoming to a given node, k_{in} (i.e., the number of times a molecule was obtained as a reaction product). As shown in Figure 2c, the distributions of both $p(k_{out})$ and $p(k_{in})$ decay algebraically

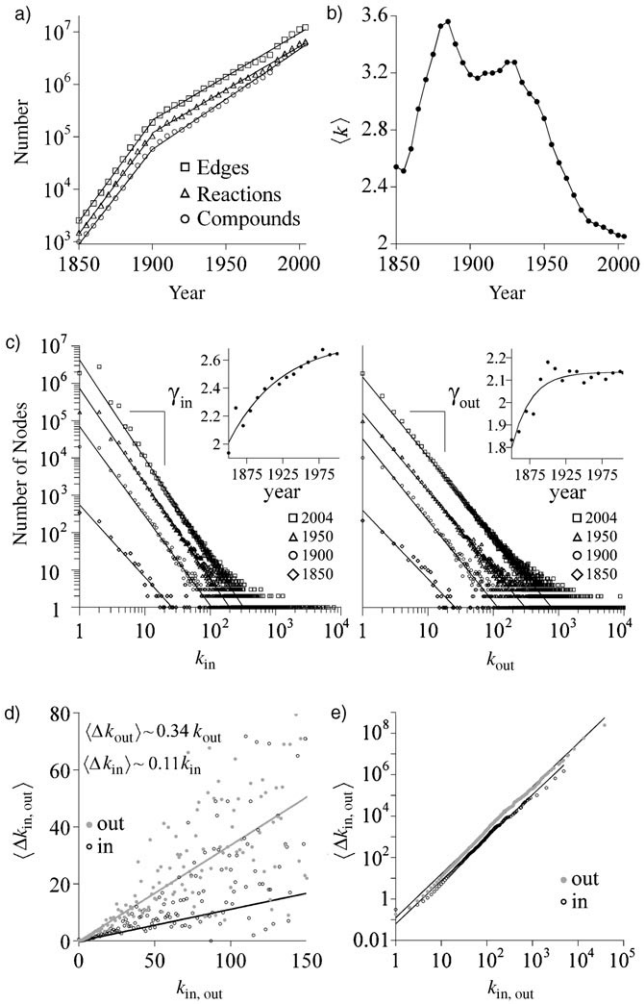

Figure 2. a) The number of chemical compounds, reactions, and graph edges as a function of time. Although all three grow exponentially with time, the rate of growth changes near the turn of the twentieth century. b) Average connectivity (degree)—defined as the number of edges divided by the number of nodes—as a function of time. In the nineteenth century, chemistry focused on "rewiring" existing compounds, whereas in the twentieth century, it has been dominated by the synthesis of new compounds. c) In- and out-degree distributions of molecules that comprise the chemistry network obey a power law $p(k) = k^{-\gamma}$ characteristic of scale-free networks. From 1850 to 2004, the power-law exponents γ_{in} and γ_{out} (inserts) have been steadily growing and approach constant values of approximately 2.7 and 2.1, respectively. d) Evolution of the chemistry network is governed by a "two-way" (i.e., both "in" and "out") mechanism of preferential attachment. The plots show the average increase in connectivities of one million random nodes from 1990 to 2000. The "strength" of preferential attachment (measured by the slope of the linear fit, $d\langle\Delta k\rangle/dk$) is approximately three times greater for out than for in connectivities. This helps to explain why the power-law exponent γ_{out} is smaller than γ_{in} and indicates that highly connected substrates are more likely to be found than highly connected products. e) The corresponding cumulative distribution for preferential attachment defined as $\kappa(k) = \sum_{k_i=0}^{k} \langle\Delta k_i\rangle$. The distribution follows a power law $\kappa(k) = k^{\sigma}$, where $\sigma \approx 2.1$ for both the in and out distributions, which implies that the preferential attachment distribution plotted in (e) is indeed linear ($\langle\Delta k\rangle \propto k^{\sigma-1} \approx k$).

with the number of connections, $p(k) \sim k^{-\gamma}$, and the γ exponents for both in and out distributions increase with time (measured in years) approximately as $\gamma_{in}(t) = 2.67[1-\exp(-(t-1780)/52.4)]$ and $\gamma_{out}(t) = 2.14[1-\exp(-(t-1780)/36.2)]$. Interestingly, the values these exponents asymptotically approach ($\gamma_{in} = 2.67$ and $\gamma_{out} = 2.14$) are similar to those that characterize the directed network of the WWW (2.71 and 2.1, respectively), thus suggesting that the network of chemistry and the WWW have similar topologies (although the latter is more highly connected; $\langle k\rangle = 7.5$ for the WWW in 2000 vs $\langle k\rangle = 2.1$ for chemistry in the same year).[16]

The observed power-law dependencies indicate that organic reactions—like the WWW,[12,15,16] the internet,[20] metabolic networks,[21] and even societies[22,23]—form a scale-free network,[10] whose architecture is distinguished by the presence of highly connected "hubs." These hub molecules of organic chemistry are directly analogous to those found in the scale-free network of the airline system in which they facilitate transportation from one poorly connected airport to another. Likewise, the synthesis of one molecule in organic chemistry from another by a series of chemical transformations will likely utilize one of these versatile hub compounds as an intermediate (as we shall see later, this also has an added economic advantage, since hub compounds are significantly less expensive than poorly connected ones).

The presence of a scale-free topology also provides evidence as to the mechanism by which the network evolves over time. Specifically, the connectivities of organic molecules evolve according to a two-way mechanism of preferential attachment (i.e., both in and out connections), which stipulates that well-connected substances are more likely to participate in new reactions than poorly connected compounds. We verified this mechanism directly, by analyzing how the connectivities of one million randomly chosen nodes changed over time. We found that the average increase in both in or out connectivities over a given time period was proportional to their values at the beginning of this period ($\langle\Delta k_{in,out}\rangle \propto k_{in,out}$; see Figure 2d,e). This result means that 1) the more times a molecule has been used as a synthetic substrate (i.e., large k_{out} value), the higher the chances that it will be used again in the future; 2) conversely, the higher its k_{in} value, the more likely that chemists will try to make it by a new reaction. Interestingly, this mechanism also implies that the evolution of a molecule's usefulness—measured by its participation in new reactions—increases exponentially over time. Therefore, highly connected compounds create exponential "explosions," thus leading to the formation of preferred substrates and target molecules (the "hubs") in organic chemistry.

The fact that the exponent characterizing outgoing connections, γ_{out}, is smaller than that characterizing incoming ones, γ_{in}, indicates (see Figure 2c) that the most connected hubs of the chemistry network have on average more outgoing than incoming connections—the most-connected molecules are usually used as synthetic substrates. At the same time, as chemistry develops, the degree of correlation between in and out connectivities of the molecules increases from $R(k_{out},k_{in}) = 0.327$ in 1850 to 0.571 in 2004 (see Support-

ing Information). The rationale for this trend is that the more useful a compound is as a synthetic substrate, the more ways are designed to prepare it (presumably, in an effort to maximize yields and minimize reaction costs); conversely, the more synthetic recipes exist for making a compound, the more available it becomes and can be used for further syntheses.

Information about the molecular masses (henceforth, simply "masses" or m) of the network's nodes (i.e., molecules) provides an additional source of information about its evolution. While mass is the simplest of the possible molecular descriptors, it was chosen because it was readily available from the database and because it correlated with several other scalar descriptors (e.g., molecular volume[24,25] or structural complexity factor[26,27] of organic molecules).

Figure 3a,b shows the frequency distribution of masses that were used as substrates ($g_{out}(m,t)$; left) and products ($g_{in}(m,t)$; right) in reactions reported between 1850 and 2004. The most commonly used substrates are those near $m = 150\,g\,mol^{-1}$ and the most common products near $m = 250\,g\,mol^{-1}$. Importantly, the shapes of both distributions and the locations of their maxima do not change with time but only shift upwards according to $g_{in,out}(m,t) = \theta_{in,out}(t)\, g^{*}_{in,out}(m)$, where $g^{*}_{in,out}(m)$ are "master" distributions (here, in 1850) and $\theta_{in,out}(t) = N_0^{in,out} \exp((t-t_0)/\tau_{in,out})$ are the propagating/scaling functions ($t_0 = 1850$, $N_0^{in} = 5910.1$, $N_0^{out} = 4869.0$, $\tau_{in} = 19.34$, $\tau_{out} = 19.07$)). These results show that the masses of the most popular substrates and products have not significantly changed for 150 years; moreover, these masses *will remain* the most popular in the future, as they correspond to the most rapidly connecting nodes of the network.

Abundances of molecules depend on their masses. Figure 3c shows the $M(m,t{=}2004)$ mass distribution of all nodes in the network. Aside from an obvious overall trend (few very small and very large molecules), the curve exhibits pronounced "high-frequency" oscillations. The Fourier spectrum of the distribution (Figure 3c, insert) shows a dominant, sharp peak at 2 and a broader peak centered at 14–15. The former indicates that there are significantly more molecules (indeed, by $\approx 48\,\%$) of even than of odd masses; the latter corresponds to the masses of the most common chemical "building blocks" (e.g., CH_2, CH_3, OH, NH_2, etc.) from which molecules are made.

More detailed, time-dependent analysis of $M(m,t)$ distributions (Figure 3d) reveals that chemists create molecules according to a single stochastic process. When plotted on a log–log scale (Figure 3e), the $M(m,t)$ curves for $t = 1850$–2004: 1) are log-normal for the majority ($> 90\,\%$) of masses around the peak; 2) decay algebraically, $M(m,t) \sim m^{-\beta}$ with the exponent $\beta \approx 4$, for $m > 800$; and 3) exhibit a characteristic "shoulder" for $m < 50$. Since 1850, the average mass increased from 200 to 350 (2004)

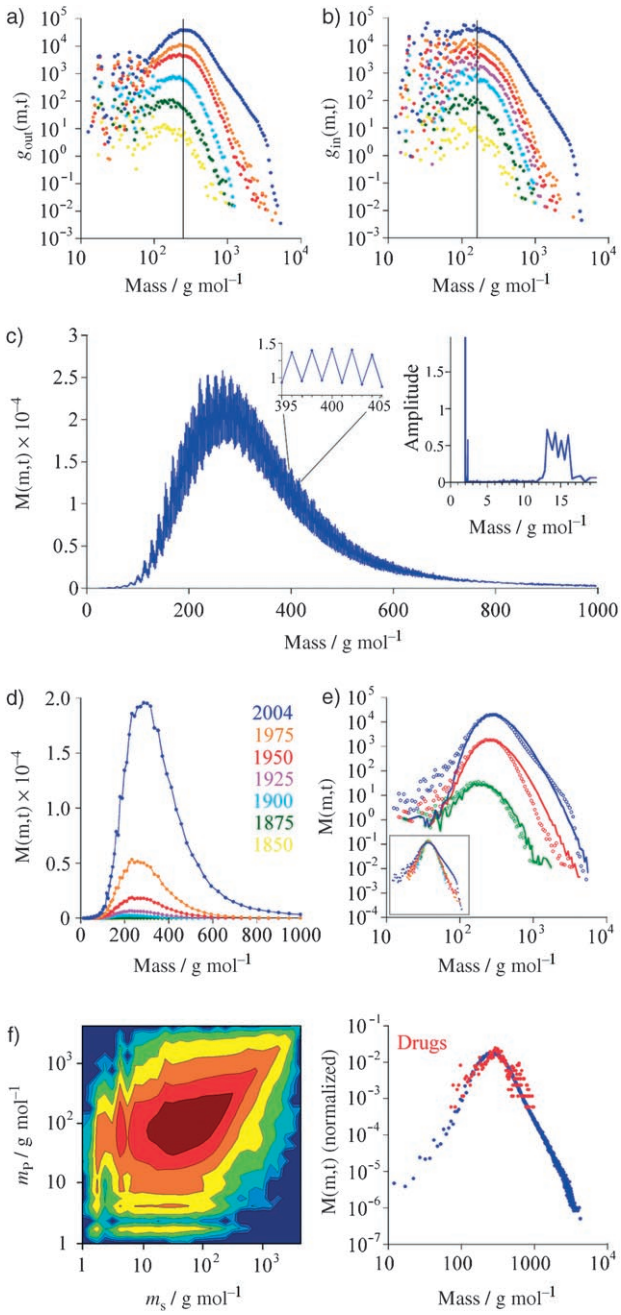

Figure 3. Frequency distributions of masses that were used as a) substrates ($g_{out}(m,t)$) and b) products ($g_{in}(m,t)$) in reactions reported in 25-year intervals between 1850 and 2004. Specifically, the substrate- and product-mass distributions were found by calculating the statistics of all masses weighted by k_{out} and k_{in}, respectively. c) Distribution of masses, $M(m,t)$, of all molecules in the chemistry network as of 2004. The zoom-in illustrates periodic variation in the abundances of molecules with even and odd masses. The insert is a Fourier spectrum of the $M(m,t)$ distribution. d) Evolution of mass distributions with time. All $M(m,t)$ curves are plotted against masses binned into $m = 2$ intervals; this binning removes high-frequency oscillations and simplifies calculations but does not, as we verified, affect the generality of the results. e) The lines in this log–log plot correspond to the theoretical mass distributions for 1875 (green), 1950 (red), and 2004 (blue) propagated from the initial distribution in 1850 according to the master equation; these simulated curves agree with the Beilstein data (points). The characteristic shape of the distributions indicates that molecules are created by a stochastic Kesten process. Insert: Mass distributions overlap when rescaled[20] with respect to the location of the peak and the total number of molecules. f) Correlation between masses of substrates m_S and products m_P based on all reactions stored in the database. g) Log–log plot of normalized mass distributions of all compounds stored in Beilstein (blue) and 1382 most important drugs (red). The two distributions are nearly identical with the goodness-of-fit $R^2 = 0.779$ (for which the Beilstein data is taken as the reference distribution).

according to $\langle m(t)\rangle = m_{fin}[1-\exp((t-1780)/\tau)]$, where $m_{fin} = 356.55\ g\,mol^{-1}$ and $\tau = 94.7$, and, as mentioned previously, the total number of masses (i.e., of molecules) increased exponentially with time. Importantly, when the $M(m,t)$ distributions for different times were rescaled,[28] they collapsed onto a single master curve (Figure 3e, insert).

The shapes and scaling properties of the mass distributions indicate that they evolve according to a Kesten-type mechanism,[17,18,29] in which new masses are created according to the multiplicative-additive master equation $m_{t+1} = \lambda_t m_t + \alpha_t$, where index t has a meaning of time. Briefly, at each update (i.e., time step), every molecule from a currently existing distribution $M(m,t)$ is used with equal, constant probability p_0 to generate a new molecule whose mass is prescribed by the master equation. The value of p_0 is determined by the "experimental" growth rate of the total number of molecules—$p_0 = \ln(N(2004)/N(1850))/n$, where n is the number of distribution updates simulated between 1850 and 2004—and sets the relationship between t and the real time. The number of updates, T, that correspond to one year is $T = 154/n$. After the update, both the "old" and the newly created masses are retained in a distribution $M(m,t+1)$, thus accounting for the increase in the number of molecules with time.[30] When stochastic (i.e., random) variables λ and α were sampled from uniform distributions (whose bounds—$\lambda_{min} = 0.9$, $\lambda_{max} = 1.091$ and $\alpha_{min} = 0$, $\alpha_{max} = 5.0$—were determined by fitting the mass distribution in year 2004), the master equation propagated experimental mass distributions faithfully; at every instant of time, these distributions were given by $M(m,t) = M(m,0)/(1+p_0)^t + p_0 \sum_{i=0}^{t-1} (1+p_0)^{i-t}\,\delta(m,i)$, where $M(m,0)$ is the initial mass distribution and $\delta(m,i)$ denotes the distribution of masses derived from $M(m,i-1)$ in the ith update.

Although the identified stochastic process reflects the fundamental statistical dependence between the masses of the substrates m_s and the products m_p in the network $m_p = a m_s + b$ (Figure 3f),[31] its practical significance goes well beyond this simple correlation. Because molecular masses evolve according to the master equation from the very inception of modern chemistry—irrespective of any prevalent "trends of the day" and constant progress in synthetic methodologies—this equation can be extrapolated into the future. In other words, it can be used to *predict* what masses (and with what probabilities) will be obtained in any reaction or set of reactions to be carried out.

Two important generalizations follow: First, the master equation indicates that very large and very small molecules are hard to make.[32] When large substrates are used, the multiplicative term dominates to give a product of a smaller mass; conversely, for small substrates, the additive term increases the mass. Beyond this common-sense result, the master equation provides quantitative information in the form of probabilities, which give an estimate of *how difficult* it is to make a molecule of a given mass. We stress that these observations should be understood only in a statistical sense and by no means exclude the possibility of the preparation of any particular small/large molecules. Second, the knowledge of the stochastic process that governs the generation of masses can be used to predict the outcomes of parallel (combinatorial) syntheses. Making a reasonable assumption that such syntheses are based on standard chemical transformations (i.e., likely described in BD), it is expected that the masses of the initial collections of fragments should evolve according to the master equation. This property could be used in designing fragment libraries[33,34] that would evolve into a desirable, final distribution of masses after a specified (and, hopefully, minimal) number of synthetic steps. An example of such an analysis for three potential libraries is illustrated in Figure 4a and further discussed in the Supporting Information.

One of the most prominent areas in which such desirable mass distributions are sought is in drug design—where it is widely believed that masses of druggable substances differ statistically from those of random ones.[35–37] Our analysis, however, reveals (Figure 3g) that this belief is unfounded and that most important drugs[38] have a mass distribution virtually identical to that of all chemicals stored in Beilstein. Whether this is a result of drugs being discovered by random sampling from the pool of existing substances or of synthetic limitations imposed by the master equation (see below) eventually leading to a log-normal distribution is an interesting, albeit speculative, question.

Unlike molecular mass, the network connectivity of a chemical is a valuable descriptor of its pharmacological or economic importance. As we have already suggested, it is reasonable to expect that for the most useful compounds, the number of ways in which they can be used should correlate with the number of ways in which they can be prepared. Indeed, for the 1382 most important drugs[38] and the top 300 most important industrial chemicals,[39] the correlations between k_{in} and k_{out} values are not only higher ($R \approx 0.6$) than for randomly chosen compounds, but are also historically persistent, as illustrated in Figure 4a, which shows time "trajectories" of these important substances in the $\ln(k_{in})$ versus $\ln(k_{out})$ plane. Their trajectories are long, highly focused, and linear—in sharp contradistinction to short and curvilinear paths traced by randomly chosen compounds. In other words, by looking at a historical evolution of the connectivity of a substance, one can make inferences about its industrial significance. In practice, this significance is often measured in monetary terms, and the prices of chemicals should be dictated by their uniqueness (here, measured by k_{in}, the known ways of preparing a chemical) and by the existing demand (measured by k_{out}, the number of reactions/chemists that use them). As expected, the prices per mole of the top 300 chemicals correlate well (Figure 4b) with these economic indicators and obey a power law of the form $\text{Cost} \propto k_{in,out}^{-\nu}$; the price of a chemical rapidly decreases with both k_{in} and k_{out}.

Of course, this and other trends we identified are only the first attempt to explore the richness of information contained in the history of chemistry. Although additional theoretical analyses can be envisioned with the Beilstein repository (e.g., network's centrality, diameter, shortest paths, aspects of molecular chirality and other structural descriptors, etc.), the unraveling of more economy-oriented trends will likely require the use of databases that report reaction stoichiometries, yields, and costs. Also, it would be interesting to see what

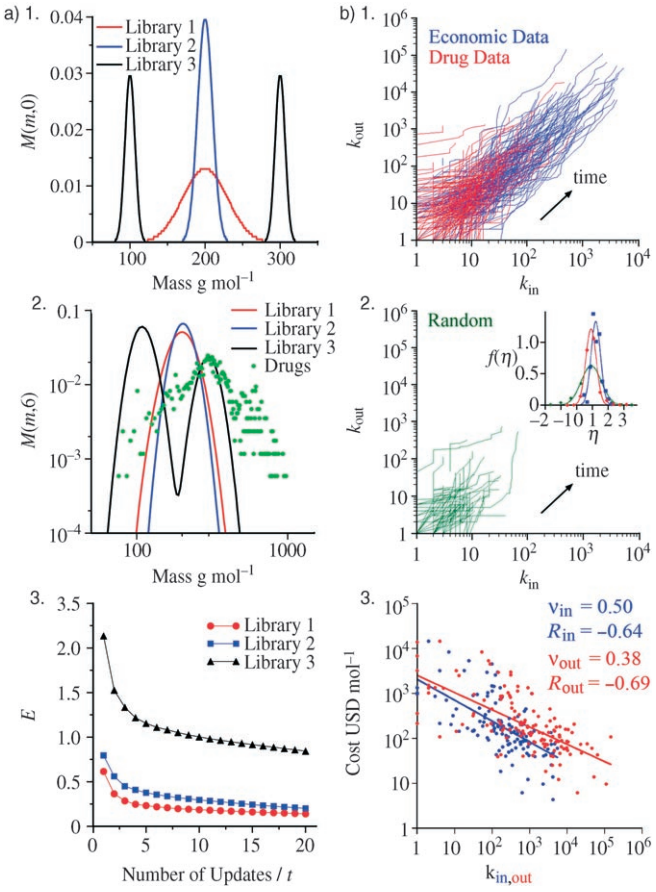

Figure 4. a) 1. Evolution of three initial "fragment libraries" in the realistic range of $m = 100$–300 simulated by the master equation. Libraries 1 and 2 represent normal distributions of masses centered at $m = 200\ g\,mol^{-1}$, and characterized by dispersions $\sigma_M = 30$ and $\sigma_M = 10\ g\,mol^{-1}$, respectively. The third library is a sum of two normal distributions of masses centered at $m = 100$ and $300\ g\,mol^{-1}$ and characterized by the same dispersion $\sigma_M = 7\ g\,mol^{-1}$. 2. Distributions obtained from the three initial "fragment libraries" after 6 updates. The target distribution of masses of 1382 most important drugs is also shown. 3. The average error coefficient, E, which measures similarity between the distributions ($E = 0$ if they are identical) obtained from Libraries 1, 2, and 3 and the target distribution of drugs plotted as a function of synthetic generations (i.e., the number of updates). The broadly distributed Library 1 converges most rapidly. b) Time trajectories of important (1382 drugs and the top 300 economically relevant chemicals and 2. 2000 randomly chosen molecules (right) plotted in the $\ln k_{in}$–$\ln k_{out}$ plane. The trajectories of the important chemicals are long and linear (i.e., $k_{out} \propto (k_{in})^{\eta}$ where $\eta \approx 1$), whereas those of "random" ones are short and of varying curvatures (with η between approximately -1.5 and 3). The insert shows probability distributions of the trajectory exponents η; the curves are best Gaussian fits to the data. Standard deviation of η is higher (0.64) for randomly chosen compounds than for economically relevant chemicals (0.30) and drugs (0.33). 3. Cost (per kmol) as a function of in and out connectivities of the top 300 economically relevant compounds. The power-law dependence indicates that more highly connected compounds are significantly less expensive. USD = US dollar.

trends could be established for particular branches of chemistry (e.g., the rapidly growing carbohydrate and lipid research, fluorine chemistry, etc.) and whether these trends would differ from the "mainstream" trends discussed herein.

On a fundamental level, our results provide yet another demonstration of how seemingly independent/uncorrelated activities of individual agents[40–44] fall into a larger scheme and obey—or, as proponents of emergence[11] would argue, "give rise to"—unifying statistical rules.[45]

Received: June 29, 2005
Revised: August 31, 2005
Please note: Minor changes have been made to this manuscript since its publication in *Angewandte Chemie* Early View. The Editor.

Keywords: combinatorial chemistry · master equations · networks · organic reactions · stochastic modeling

[1] K. C. Nicolaou, E. J. Sorensen, *Classics in Total Synthesis: Targets Strategies, Methods*, Weinheim, New York, **1996**.
[2] N. Anand, J. S. Bindra, S. Randanathan, *Art in Organic Synthesis*, Holden-Day, San Francisco, **1970**.
[3] E. J. Corey, A. K. Long, S. D. Rubenstein, *Science* **1985**, *228*, 408.
[4] S. L. Schreiber, *Science* **2000**, *287*, 1964.
[5] K. C. Nicolaou, D. Vourloumis, N. Winssinger, P. S. Baran, *Angew. Chem.* **2000**, *112*, 46; *Angew. Chem. Int. Ed.* **2000**, *39*, 44.
[6] M. D. Burke, S. L. Schreiber, *Angew. Chem.* **2004**, *116*, 48; *Angew. Chem. Int. Ed.* **2004**, *43*, 46.
[7] E. J. Corey, X.-M. Cheng, *The Logic of Chemical Synthesis*, Wiley, New York, **1989**.
[8] M. R. Spaller, M. T. Burger, M. Fardis, P. A. Bartlett, *Curr. Opin. Chem. Biol.* **1997**, *1*, 47.
[9] C. Lipinski, A. Hopkins, *Nature* **2004**, *432*, 855.
[10] R. Albert, A. L. Barabasi, *Rev. Mod. Phys.* **2002**, 74, 47.
[11] A. L. Barabasi, R. Albert, *Science* **1999**, *286*, 509.
[12] L. A. N. Amaral, J. M. Ottino, *Eur. Phys. J. B* **2004**, *38*, 147.
[13] L. A. N. Amaral, A. Scala, M. Barthelemy, H. E. Stanley, *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 11149.
[14] C. M. Song, S. Havlin, H. A. Makse, *Nature* **2005**, *433*, 392.
[15] R. Albert, H. Jeong, A. L. Barabasi, *Nature* **1999**, *401*, 130.
[16] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, *Comput. Networking* **2000**, *33*, 309.
[17] D. Sornette, R. Cont, *J. Phys. I* **1997**, *7*, 431.
[18] L. Dehaan, R. L. Karandikar, *Stoch. Process. Their Appl.* **1989**, *32*, 225.
[19] MDL Crossfire Beilstein database; we stress that although Beilstein has data on numerous man-made polymers, it is not a comprehensive repository of biological polymers (proteins, DNA).
[20] M. Faloutsos, P. Faloutsos, C. Faloutsos, *Comput. Commun. Rev.* **1999**, *29*, 251.
[21] H. Jeong, B. Tombor, R. Albert, Z. N. Oltval, A. L. Barabasi, *Nature* **2000**, *407*, 651.
[22] S. Redner, *Eur. Phys. J. B* **1998**, *4*, 131.
[23] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, Y. Aberg, *Nature* **2001**, *411*, 907.
[24] C. G. Screttas, G. A. Heropoulos, *THEOCHEM* **1994**, *309*, 149.
[25] C. Wandrey, A. Bartkowiak, D. Hunkeler, *Langmuir* **1999**, *15*, 4062.
[26] J. B. Hendrickson, P. Huang, A. G. Toczko, *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 63; we wrote a computer program based on the Hendrickson algorithm, which was used to calculate complexity factors of 1000 randomly chosen organic molecules from the BD. We found that the values of the factors correlated with molecular masses with the correlation coefficient $R = 0.72$. The executable of the program along with instructions can be downloaded free of charge from our web page (http://www.dysa.northwestern.edu).

[27] T. K. Allu, T. I. Oprea, *J. Chem. Info. Model.* **2005**, *45*, 1237.
[28] The rescaling procedure consisted of: 1) multiplying the distribution $M(m,t)$ by the ratio $N(t{=}2004)/N(t{=}1850)$, where $N(t)$ stands for the total number of molecules at time t, and 2) translation by $\Delta M = M_{max}(2004) - M_{max}(t)$, where M_{max} corresponds to the distribution maximum.
[29] In its original version, the Kesten process describes the transformation of an arbitrary distribution of variable x according to a multiplicative-additive equation $x_j = \lambda_j x_j + \alpha_j$, where j stands for the iteration number and λ_j and α_j are independent, positive stochastic variables. In the limit of large j, this procedure generates a log-normal distribution $w_K(x)$ with a characteristic "algebraic" tail $w_K(x) \sim x^{-(1+\mu)}$ in which the exponent is determined by the condition $\langle \lambda^\mu \rangle = 1$.
[30] This is unlike the pure Kesten process, in which the old distribution is discarded after each iteration.
[31] The values of the effective multiplicative ($a = 0.67$) and additive ($b = 179.94$) coefficients are commensurate with those derived from the stochastic master equation, $a = \langle \prod_{i=1}^{i=T} \lambda_i \rangle = 0.91$ and $b = \langle \sum_{j=1}^{j=T} \alpha_j \prod_{i=j+i}^{T} \lambda_i \rangle = 47.7$.
[32] These observations can also be understood in terms of combinatorial considerations and reflect the fact that 1) on one hand, there are relatively few combinations of atoms that give rise to structures of low molecular mass. It can be readily estimated that the possible numbers of molecules of mass m increase roughly exponentially with m. 2) On the other hand, when two large molecules are "stitched" together, there are many possible combinations (orientations) of doing so and achieving a selective, high-yield synthesis is problematic.
[33] D. C. Rees, M. Congreve, C. W. Murray, R. Carr, *Nat. Rev. Drug Discovery* **2004**, *3*, 660.
[34] E. J. Martin, R. E. Critchlow, *J. Comb. Chem.* **1999**, *1*, 32.
[35] A. K. Ghose, V. N. Viswanadhan, J. J. Wendoloski, *J. Comb. Chem.* **1999**, *1*, 55.
[36] T. I. Oprea, *J. Comput.-Aided Mater. Des.* **2000**, *14*, 251; therein a comparison is made between the mass distributions of the Available Chemical Directory (ADC) and MACCS II Drug Data Report (MDDR) (we note that the ADC is not necessarily representative of "random" substances).
[37] J. Xu, J. Stevenson, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177.
[38] Medical Essential Drug Database http://lsb.sbs.auckland.ac.nz/classic/medbase.
[39] Chemical Prices, *Chemical Market Reporter*, March 7, **2005** http://www.chemicalmarketreporter.com.
[40] S. Camazine, *Behav. Ecol. Sociobiol.* **1991**, *28*, 61.
[41] S. Camazine, J.-L. Deneubourg, N. R. Franks, J. Sneyd, G. Theraulaz, E. Bonabeau, *Self-Organization in Biological Systems*, Princeton University Press, Princeton, **2001**.
[42] D. J. Watts, S. H. Strogatz, *Nature* **1998**, *393*, 440.
[43] R. V. Sole, J. M. Montoya, *Proc. R. Soc. London Ser. B* **2001**, *268*, 2039.
[44] A. Sih, G. Englund, D. Wooster, *Trends Ecol. Evol.* **1998**, *13*, 350.
[45] Details of an example for combinatorial chemistry, a discussion of the network wiring scheme, and details of the Beilstein Database are given in the Supporting Information.

