Angewandte Chemie
Eine Zeitschrift der Gesellschaft Deutscher Chemiker

Supporting Information

69451 Weinheim, Germany

Architecture and Evolution of Organic Chemistry**

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

Evolution of molecular mass distributions: A combi-chem example.

We suggest that the master equation describing the evolution of molecular mass distributions can be a useful tool in combinatorial chemistry, especially in designing libraries of fragments, which after a specified (and, hopefully, minimal) number of synthetic steps, would evolve into a desirable, final mass distribution. As an illustrative example, we investigated how rapidly different initial collections of masses ("fragment libraries") approached the target distribution chosen here as the distribution of masses of 1382 most important drugs (but cf. the text for the discussion of this distribution's purported uniqueness). We simulated evolution of three different initial "libraries" in the realistic range of m = 100-300: (i) $M_1(m,0)$ composed of fragments of masses broadly distributed around m = 200 g/mol; (ii) $M_2(m,0)$ consisting of equal number of fragments but sharply distributed around m = 200 g/mol; and (iii) $M_3(m,0)$ with fragments distributed bimodally around m=100 g/mol and 300 g/mol (all distributions are normal Gaussians; Fig. 4a,i).

In each update ("synthetic generation"), the distribution of masses, $M(m,i)$ was transformed into $M(m,i+1)$ according to the formula:

$M(m,i+1) = (M(m,i) + p_0 d(m,i))/(1+p_0)$, where $d(m,i)$ denotes the distribution of masses derived from $M(m,i)$ in the i-th update, by the master equation (chemically, this means that molecules were reacted at random). Thus, after t updates, the mass

distribution $M(m,t)$ was given by:

$M(m,t) = M(m,0)/(1+p_0)^t + p_0 \sum_{i=0}^{t-1} (1+p_0)^{i-t} d(m,i)$, where $M(m,0)$ is the initial mass distribution.

The evolved distributions, $M_1(m,t)$, $M_2(m,t)$ and $M_3(m,t)$ are shown in Fig. 4a,ii. In order to measure similarity between these distributions and the target distribution of drugs, $D(m)$, we used the average error coefficient, *E*, which measures deviations of given variable, *f*, (here: $M_{1,2,3}(m,t)$) relative to some reference variable, *g* (here: $D(m)$). The value of *E* is given by $E = \frac{1}{N} \sum_{i=1}^{N} |f_i - g_i| / g_i$, where *N* denotes the number of the data points in distribution *g*. If the distributions *f* and *g* are identical, then *E* = 0. The error coefficient for all distributions is plotted as a function of the number of updates, *t*, in Fig. 4a, iii. It follows from the dependence *E(t)* that the initial distribution $M_1(m,0)$ gives faster convergence to $D(m)$ than $M_2(m,0)$ (i.e., a library more diverse in terms of masses is better); at the same time $M_3(m,t)$ distribution converges only slowly. In practical terms: if one would like to make maximally drug-like compounds (again, in terms of mass only; we cannot make any inferences about other relevant properties) in the smallest number of synthetic generations, one should begin with broad fragment library $M_1(m,0)$.

The Network Wiring Scheme

Conversion of a set of chemical reactions into a directed graph allows a certain degree of subjectivity. While in the main text we studied the network generated by connecting all substrates to all products (*All-to-All*), we also investigated other wiring

schemes, notably one created by connecting only the heaviest substrate to the heaviest product and producing only one directed edge per one chemical reaction. This *1-to-1* network represents a more "condensed" framework of organic chemistry and is a subgraph of the *All-to-All* construction. The *1-to-1* network is, in some cases, more "intuitive" (e.g., when it omits small functional/protective groups added onto a large molecule); in others, however, it might not capture all chemically-relevant information (e.g., when reagents are equally important and have similar molecular masses).

Fortunately (for the generality of results), these and other wiring schemes produce similar results in all aspects studied. For example, the *1-to-1* network is scale free, and its in- and out- connectivities obey the characteristic power law relations with exponents g_{in}^{1to1} = 2.70 and g_{out}^{1to1} = 2.30 in 2004 close to those of the *All-to-All* network (2.67 and 2.14, respectively). We note that the fact that the exponents of the *1-to-1* network are slightly larger than those of the *All-to-All* graph reflects the fact that the former construction is biased against lighter molecules. In the *All-to-All* network, such light molecules are also the most highly connected substrates and products; $g_{out}(m,t)$ and $g_{in}(m,t)$ have a maximums at m = 150 g/mol and m = 250 g/mol, respectively, as compared to the average mass, m_{avg} = 358 g/mol in 2004. Therefore, by selectively reducing the number of connections to/from light molecules (nodes), the *1-to-1* construction reduces the probability of highly connected nodes, thereby increasing the power law exponents. Furthermore, this effect is more pronounced for out-connectivities than for in-connectivities, because the most highly connected substrates are significantly lighter than the most highly connected products (cf. g-distributions in Fig. 2a-b).

The Beilstein Database

The following information was collected directly from Beilstein Data Field Reference Guide and from Heller's book entitled *The Beilstein System.*

1) Data Sources
    a. The Beilstein Handbook from the Basic Series to Supplement IV covering the literature from 1779 to 1959. For more than 1.1 million compounds the complete Handbook information is available.
    b. Primary literature data from 1960 to 1979: this data source contains ca. 3 million additional compounds. Specific data is available for melting point, boiling point, density, refractive index, optical rotatory power, isolation from natural products and chemical derivatives. All other physical and chemical properties are available as keywords together with corresponding references to the original literature. This data source provides the basis for the production of the Beilstein Handbook supplement V. This part of the file is being continuously updated to provide data for all data fields.
    c. Primary literature data >1979: in contrast to source 2. detailed information for all physical and chemical properties have been abstracted from the literature. All data fields contain references as well as data. The category "Pharmacological and Ecological Data" is also included. The yearly growth is in the range of ca. 250.000 structures with 220.000 reactions and ca. 40.000 citations.

2) Substance Criteria
    a. Substances are registered in the database if they contain, besides carbon, only the following elements out of the Periodic Table:
        Group I: H, Li, Na, K, Rb, Cs
        Group II: Mg, Ca, Sr, Ba
        Group III: B
        Group IV: C, Si
        Group V: N, P, As
        Group VI: O, S, Se, Te
        Group VII: F, Cl, Br, I

        These are designated as Beilstein substances. A further prerequisite is that an acceptable piece of data is given for the substance or that the substance is referred to in connection with another substance (components of mixtures or preparative methods).
    b. Iorganic substances (Gmelin substances) are in general not registered (see exceptions below). These are:
        i. Substances which contain, besides carbon, the elements which are not mentioned in the list above
        ii. Substances which do not contain carbon
        iii. Elements
    c. As an exception from the boundary Beilstein ⇔ Gmelin areas the following substances are not treated as Beilstein substances:

i. CO, CS, CO_2, CS_2, COS, C_3O_2, C_3S_2
ii. Carbonic acid and its thio analogs along with their salts with inorganic cations
iii. HCN, HOCN, HSCN and the corresponding iso-acids together with all metal salts and complexes of these acids
iv. Dicyanogene
v. Phosgene
vi. Metal carbides
vii. Metal salts of formic acid, acetic acid, and oxalic acid
viii. Fullerenes, which consist only of carbon
ix. Carboranes

3) Substance Types

Beilstein has extended the concept of substances to enable the abstraction of environmental, pharmacological and toxicological data. It now allows the acceptance of substances whose identity is not necessarily given by means of a structure. Substances will continue to be "chemical substances" in their organic chemical sense. Beilstein distinguishes between several types of substances:

a. Pure substances with a structural formula:
   A - Classical Beilstein compounds
b. Substances which may be described by means of names or information about the components:
   B - Biomolecules (biopolymers [carbohydrates, nucleic acids, proteins], enzymes, hormones, etc.)
   M - Mixtures (composed of components)
      1) composition completely given
      2) composition partially given
      3) composition completely not given
   P – Polymers
      1) monomers given
      2) monomers not given

4) Beilstein Reaction Database

The Beilstein reaction databases is the largest – both in time coverage and number of reactions – and the fastest growing reaction database in the world. Reaction data are collected directly from the primary literature.

Figure S1: As chemistry evolves, the in- and out- connectivities become more correlated; the correlation coefficient $R(k_{in}, k_{out})$ grows from 0.327 in 1850 to 0.571 in 2004. The contour plots have the numbers of molecules characterized by a particular set of (k_{in},k_{out}) values. The scale on the axes and the color scale are logarithmic.