

## MOLECULES IN SILICO: POTENTIAL VERSUS KNOWN ORGANIC COMPOUNDS

Adalbert Kerber<sup>†</sup>, Reinhard Laue<sup>†</sup>,  
Markus Meringer<sup>‡,1</sup>, Christoph Rücker<sup>§</sup>

<sup>†</sup>Department of Mathematics, University of Bayreuth,  
95440 Bayreuth, Germany

<sup>‡</sup>Department of Medicinal Chemistry, Kiadis B.V., Zernikepark 6-8,  
9747 AN Groningen, The Netherlands

<sup>§</sup>Biocenter, University of Basel, Klingelbergstrasse 70,  
4056 Basel, Switzerland

(Received April 12, 2005)

**ABSTRACT.** For molecular weights up to 150, all molecular graphs corresponding to possible organic compounds made of C, H, N, O were generated using the structure generator MOLGEN. The numbers obtained were compared to the numbers of molecular graphs corresponding to actually known compounds as retrieved from the Beilstein file. The results suggest that the overwhelming majority of all organic compounds (even in this low molecular weight range) is unknown. Within the set of C<sub>6</sub>H<sub>6</sub> isomers, a very crude and a highly sophisticated energy content calculation perform amazingly similar in predicting a particular structure's existence as a known compound.

### 1. INTRODUCTION

Most chemical compounds, and in particular the molecules of all organic compounds, can be abstracted as molecular graphs. In fact, the primary means of communication among organic chemists, structural formulae, are molecular graphs. A molecular graph is a multigraph consisting of vertices representing atoms and edges representing covalent

---

<sup>1</sup>Corresponding author e-mail: m.meringer@kiadis.com

bonds. The vertices are colored by the symbols of chemical elements and optionally of atomic states [1], the edge multiplicities correspond to single, double, and triple bonds.

Although the number of known chemical compounds amounts to several millions, quite a lot of possible chemical compounds is still not known. The major aim of the present study was to quantify the gap between the number of existing organic compounds on the one hand, and of potential such compounds on the other.

## 2. METHODS

**2.1. Molecular graphs corresponding to potential organic compounds.** In order to keep the amount of data manageable, we restricted consideration to one-component compounds made of C, H, N, and O exclusively, to molecular weights up to 150 amu, to standard valences (4, 1, 3, 2 for C, H, N, O, respectively), and to the most abundant isotope of each element ( $^{12}\text{C}$ ,  $^1\text{H}$ ,  $^{14}\text{N}$ ,  $^{16}\text{O}$ ).

For each integer molecular weight  $m = 12, \dots, 150$  all valid non-ionic molecular formulae made of elements from the set  $\{\text{C}, \text{H}, \text{N}, \text{O}\}$  and containing at least one C atom were calculated under these restrictions. In doing so, several constraints have to be taken into account.

Let  $\beta(X)$  denote the number of atoms of element  $X$  in molecular formula  $\beta$ . From the elements' nominal masses we obtain the equation

$$12 \cdot \beta(\text{C}) + \beta(\text{H}) + 14 \cdot \beta(\text{N}) + 16 \cdot \beta(\text{O}) = m.$$

This diophantine equation can be solved by backtracking. However, not every elemental composition obtained thereby is the molecular formula of a molecular graph. For example,  $\text{CH}_2$ ,  $\text{C}_2\text{H}_8$ ,  $\text{C}_2\text{H}_7\text{O}$ , being solutions of this equation for  $m = 14, 32$ , and  $47$ , respectively, do not correspond to a molecular graph. In order to result in a molecular graph, a solution of the diophantine equation has to obey further restrictions in terms of the atoms' valences.

- (i)  $4 \cdot \beta(\text{C}) + \beta(\text{H}) + 3 \cdot \beta(\text{N}) + 2 \cdot \beta(\text{O}) \equiv 0 \pmod{2}$ ,
- (ii)  $4 \cdot \beta(\text{C}) + \beta(\text{H}) + 3 \cdot \beta(\text{N}) + 2 \cdot \beta(\text{O}) - 8 \geq 0$ ,
- (iii)  $2 \cdot \beta(\text{C}) - \beta(\text{H}) + \beta(\text{N}) + 2 \geq 0$ .

The left-hand side of equation (i) is the sum of all valences, which is required to be an even number in order to avoid dangling bonds. Inequation (ii) says that a carbon compound should contain at least four bonds. Condition (iii) requires a molecular graph to be connected (one component). The above examples  $\text{C}_2\text{H}_7\text{O}$ ,  $\text{CH}_2$ , and  $\text{C}_2\text{H}_8$  violate restrictions (i), (ii), and (iii), respectively. A derivation of these conditions is given in [2].

For the molecular weight range up to 150 we obtained altogether 1405 valid molecular formulae. Since a closed formula for the number of structures corresponding to a molecular formula is not even known for the alkanes,  $\text{C}_n\text{H}_{2n+2}$ , for each valid molecular formula we had to

explicitly generate all possible structures. This was done using the generator MOLGEN 3.5 [3, 4]. This version of MOLGEN is based on orderly generation [5, 6], it is able to construct many thousands of non-isomorphic molecular graphs per second. For example, the generation of all  $C_8H_6N_2O$  constitutional isomers (109240025 structures) took 3.5 min on a 2.6 GHz Pentium 4 PC.

**2.2. Molecular graphs corresponding to existing organic compounds.** Within a certain molecular formula or molecular weight range, the existing organic compounds can be retrieved from a database such as the Beilstein file. Beilstein is meant to contain all known existing organic (i. e. carbon) compounds, either naturally occurring or synthesized. Some categories of carbon compounds are excluded from Beilstein, being considered as inorganic (e.g. organometallics), biochemical (higher peptides and proteins), or structurally not well-defined compounds (polymers).

The 2003 version of Beilstein examined here<sup>2</sup> has 8711107 compound entries. For each valid molecular formula a set of one-component non-ionic compounds was retrieved, and isotopically labeled compounds, radicals, and compounds containing non-standard atom valences were removed. Using a canonizer described earlier [7], we further removed duplicate structures and stereoisomers, the latter being individual compounds corresponding to the same molecular graph.

Thus, instead of compounds we actually counted molecular graphs realized in at least one compound. This approximation was mandatory since MOLGEN, as well, works on the level of molecular graphs, not on the level of compounds including stereoisomers.

### 3. RESULTS AND DISCUSSION

Table 1 shows, for molecular weights  $m$  up to 70, all molecular formulae  $\beta$  obtained as described above (the complete version of Table 1 for  $m$  up to 150 is printed in [8]). Note that there are molecular formulae for most integer molecular weights, the few exceptions being 17–25, 33–35, 37, and 49. Note further that, as a rule, for odd molecular weights there are fewer molecular formulae than for neighboring even molecular weights. This fact seems to result from the requirement (following from equation (i)) that an odd  $m$  molecular formula contains an uneven number of N atoms, whereas for an even  $m$  molecular formula the number of N atoms is even including zero.

In column *MG* of Table 1 the number of molecular graphs (structures) generated by MOLGEN for each molecular formula is given. Column *BS* indicates the number of corresponding structures existing in the form of at least one real compound, retrieved from Beilstein as

<sup>2</sup>Beilstein database BS0302PR with MDL CrossFire Commander server software, version 6.0, MDL Information Systems GmbH

<i>m</i>	$\beta$	<i>MG</i>	<i>BS</i>	<i>MS</i>
16	CH <sub>4</sub>	1	1	1
26	C <sub>2</sub> H <sub>2</sub>	1	1	1
27	CHN	1	1	1
28	C <sub>2</sub> H <sub>4</sub>	1	1	1
29	CH <sub>3</sub> N	1	1	0
30	CH <sub>2</sub> O	1	1	1
	C <sub>2</sub> H <sub>6</sub>	1	1	1
31	CH <sub>5</sub> N	1	1	1
32	CH <sub>4</sub> O	1	1	1
36	C <sub>3</sub>	1	0	0
38	C <sub>3</sub> H <sub>2</sub>	2	1	0
39	C <sub>2</sub> HN	2	0	0
40	CN <sub>2</sub>	1	0	0
	C <sub>2</sub> O	1	0	0
	C <sub>3</sub> H <sub>4</sub>	3	3	3
41	C <sub>2</sub> H <sub>3</sub> N	5	5	1
42	CH <sub>2</sub> N <sub>2</sub>	4	4	0
	C <sub>2</sub> H <sub>2</sub> O	3	3	1
	C <sub>3</sub> H <sub>6</sub>	2	2	2
43	CHNO	3	3	0
	C <sub>2</sub> H <sub>5</sub> N	4	4	1
44	CO <sub>2</sub>	1	1	1
	CH <sub>4</sub> N <sub>2</sub>	4	4	0
	C <sub>2</sub> H <sub>4</sub> O	3	3	2
	C <sub>3</sub> H <sub>8</sub>	1	1	1
45	CH <sub>3</sub> NO	5	5	2
	C <sub>2</sub> H <sub>7</sub> N	2	2	2
46	CH <sub>2</sub> O <sub>2</sub>	2	2	1
	CH <sub>6</sub> N <sub>2</sub>	2	2	1
	C <sub>2</sub> H <sub>6</sub> O	2	2	2
47	CH <sub>5</sub> NO	3	3	1
48	CH <sub>4</sub> O <sub>2</sub>	2	2	0
	C <sub>4</sub>	3	0	0
50	C <sub>4</sub> H <sub>2</sub>	7	1	1
51	C <sub>3</sub> HN	7	1	1
52	C <sub>2</sub> N <sub>2</sub>	5	1	1
	C <sub>3</sub> O	2	0	0
	C <sub>4</sub> H <sub>4</sub>	11	7	1
53	C <sub>3</sub> H <sub>3</sub> N	19	6	1
54	C <sub>2</sub> H <sub>2</sub> N <sub>2</sub>	19	4	0
	C <sub>3</sub> H <sub>2</sub> O	9	3	0
	C <sub>4</sub> H <sub>6</sub>	9	9	7
55	CHN <sub>3</sub>	6	1	0
	C <sub>2</sub> HNNO	11	1	0
	C <sub>3</sub> H <sub>5</sub> N	21	13	2
56	CN <sub>2</sub> O	4	1	0
	C <sub>2</sub> O <sub>2</sub>	3	1	0
	C <sub>2</sub> H <sub>4</sub> N <sub>2</sub>	27	9	0
	C <sub>3</sub> H <sub>4</sub> O	13	13	2
	C <sub>4</sub> H <sub>8</sub>	5	5	4
57	CH <sub>3</sub> N <sub>3</sub>	13	0	0

<i>m</i>	$\beta$	<i>MG</i>	<i>BS</i>	<i>MS</i>
	C <sub>2</sub> H <sub>3</sub> NO	26	6	2
	C <sub>3</sub> H <sub>7</sub> N	12	12	6
58	CH <sub>2</sub> N <sub>2</sub> O	18	4	0
	C <sub>2</sub> H <sub>2</sub> O <sub>2</sub>	9	3	1
	C <sub>2</sub> H <sub>6</sub> N <sub>2</sub>	18	10	1
	C <sub>3</sub> H <sub>6</sub> O	9	9	6
	C <sub>4</sub> H <sub>10</sub>	2	2	2
59	CHNO <sub>2</sub>	8	2	0
	CH <sub>5</sub> N <sub>3</sub>	11	4	1
	C <sub>2</sub> H <sub>5</sub> NO	22	10	3
	C <sub>3</sub> H <sub>9</sub> N	4	4	4
60	CO <sub>3</sub>	1	1	0
	CH <sub>4</sub> N <sub>2</sub> O	21	7	2
	C <sub>2</sub> H <sub>4</sub> O <sub>2</sub>	10	9	3
	C <sub>2</sub> H <sub>8</sub> N <sub>2</sub>	6	5	4
	C <sub>3</sub> H <sub>8</sub> O	3	3	3
	C <sub>5</sub>	6	0	0
61	CH <sub>3</sub> NO <sub>2</sub>	15	5	1
	CH <sub>7</sub> N <sub>3</sub>	4	1	0
	C <sub>2</sub> H <sub>7</sub> NO	8	7	3
62	CH <sub>2</sub> O <sub>3</sub>	4	2	0
	CH <sub>6</sub> N <sub>2</sub> O	8	2	0
	C <sub>2</sub> H <sub>6</sub> O <sub>2</sub>	5	5	2
	C <sub>5</sub> H <sub>2</sub>	21	0	0
63	CH <sub>5</sub> NO <sub>2</sub>	8	1	0
	C <sub>4</sub> HN	27	0	0
64	CH <sub>4</sub> O <sub>3</sub>	3	3	0
	C <sub>3</sub> N <sub>2</sub>	14	0	0
	C <sub>4</sub> O	7	0	0
	C <sub>5</sub> H <sub>4</sub>	40	8	0
65	C <sub>4</sub> H <sub>3</sub> N	87	7	0
66	C <sub>3</sub> H <sub>2</sub> N <sub>2</sub>	86	8	1
	C <sub>4</sub> H <sub>2</sub> O	36	2	0
	C <sub>5</sub> H <sub>6</sub>	40	20	4
67	C <sub>2</sub> HN <sub>3</sub>	34	3	0
	C <sub>3</sub> HNNO	46	1	0
	C <sub>4</sub> H <sub>5</sub> N	116	12	5
68	CN <sub>4</sub>	6	0	0
	C <sub>2</sub> N <sub>2</sub> O	20	2	0
	C <sub>3</sub> O <sub>2</sub>	7	1	1
	C <sub>3</sub> H <sub>4</sub> N <sub>2</sub>	155	19	5
	C <sub>4</sub> H <sub>4</sub> O	62	19	2
	C <sub>5</sub> H <sub>8</sub>	26	25	16
69	C <sub>2</sub> H <sub>3</sub> N <sub>3</sub>	99	10	2
	C <sub>3</sub> H <sub>3</sub> NO	136	13	4
	C <sub>4</sub> H <sub>7</sub> N	85	30	6
70	CH <sub>2</sub> N <sub>4</sub>	31	4	1
	C <sub>2</sub> H <sub>2</sub> N <sub>2</sub> O	114	8	2
	C <sub>3</sub> H <sub>2</sub> O <sub>2</sub>	34	5	1
	C <sub>3</sub> H <sub>6</sub> N <sub>2</sub>	136	23	2
	C <sub>4</sub> H <sub>6</sub> O	55	34	15
	C <sub>5</sub> H <sub>10</sub>	10	10	10

TABLE 1. Numbers of generated and of retrieved molecular graphs of specified molecular formula. For the meaning of columns *MG*, *BS*, *MS* see text

<i>m</i>	<i>MF</i>	<i>MG</i>	<i>BS</i>	<i>MS</i>
51	1	7	1	1
52	3	18	8	2
53	1	19	6	1
54	3	37	16	7
55	3	38	15	2
56	5	52	29	6
57	3	51	18	8
58	5	56	28	10
59	4	45	20	8
60	6	47	25	12
61	3	27	13	4
62	4	38	9	2
63	2	35	1	0
64	4	64	11	0
65	1	87	7	0
66	3	162	30	5
67	3	196	16	5
68	6	276	66	24
69	3	320	53	12
70	6	380	84	31
71	5	373	66	16
72	9	403	75	27
73	5	335	61	16
74	8	369	65	27
75	6	320	44	12
76	9	419	42	9
77	4	520	11	1
78	6	877	42	7
79	4	1112	21	2
80	7	1645	100	23
81	3	2074	51	11
82	6	2610	206	52
83	6	2851	147	19
84	10	3221	276	78
85	6	3143	189	32
86	10	3380	256	77
87	9	3174	166	33
88	13	3525	184	65
89	8	4150	114	20
90	11	5849	107	28
91	9	7521	55	4
92	12	10894	137	19
93	6	14591	93	12
94	9	18860	314	45
95	7	22393	153	22
96	12	26445	550	112
97	6	28352	389	35
98	11	31099	787	177
99	10	31233	502	61
100	16	33627	710	154

<i>m</i>	<i>MF</i>	<i>MG</i>	<i>BS</i>	<i>MS</i>
101	10	38086	412	62
102	15	48014	527	113
103	13	59788	280	41
104	18	83934	373	56
105	11	113690	124	13
106	15	149575	413	43
107	12	186287	210	24
108	17	228976	854	126
109	9	260828	450	41
110	14	295401	1420	192
111	12	313498	927	63
112	18	340265	1963	278
113	11	381354	1052	75
114	17	450823	1724	250
115	16	545029	908	81
116	22	732127	1315	175
117	15	984109	618	52
118	20	1297523	829	107
119	18	1666914	394	51
120	24	2105911	948	88
121	15	2506173	465	64
122	20	2927092	1659	137
123	17	3249694	1060	66
124	24	3607136	2810	230
125	14	4062183	1733	73
126	21	4676436	3627	303
127	18	5533238	2021	116
128	26	7151422	3149	344
129	17	9413835	1718	123
130	24	12342645	2423	241
131	22	16092977	1206	94
132	30	20713339	1883	166
133	20	25402700	950	77
134	27	30490819	2039	201
135	24	35085124	1199	107
136	32	39929496	3375	351
137	20	45530360	1888	96
138	27	52127166	4938	360
139	24	61065969	2752	96
140	32	76676102	6028	389
141	21	98693127	3152	117
142	29	128022986	5393	350
143	26	167704025	2682	122
144	35	218055323	4151	337
145	24	273040710	2127	131
146	32	335050591	3518	245
147	30	396768206	1887	134
148	39	462214697	3553	271
149	27	535197826	2178	122
150	35	615977591	5300	376

TABLE 2. Numbers of generated and of retrieved molecular graphs of specified molecular weight. For the meaning of columns *MG*, *BS*, *MS* see text

described above. Finally, column *MS* shows the number of corresponding molecular graphs existing as real compounds with a mass spectrum available in one of the largest collections of mass spectra, the NIST MS library<sup>3</sup>. Comparison shows that up to  $m \approx 50$  columns *MG* and *BS* more or less agree, while at higher  $m$  the number of molecular graphs derived from existing compounds falls short of the number of potential molecular graphs. The gap between *MG* and *BS* seems to widen for increasing  $m$ . A mass spectrum is available for a small fraction of existing compounds only.

A more condensed view of the data for  $m$  between 51 and 150 is presented in Table 2. Here column *MF* gives the number of molecular formulae corresponding to a particular  $m$ , the meaning of the other columns is as before. The observations made in Table 1 are fully confirmed by the data in Table 2. Over the whole  $m$  range considered,  $m = 12, \dots, 150$ , 3699858517 molecular graphs were generated (time for generation: 2 h 11 min on a 2.6 GHz Pentium 4 PC), 103036 realized molecular graphs were retrieved from Beilstein, and 9136 molecular graphs correspond to a compound with a mass spectrum available in the NIST MS library. This is an approximate ratio of 404976 : 11 : 1.

A graphical representation of the data from Table 2 is shown in Figure 1. Two unexpected features are easily observed:

- (i) While *MG* increases for increasing  $m$  (with a few exceptions in the low  $m$  range), *BS* shows a remarkable even/odd  $m$  zig-zag behavior reminding to that noticed above for *MF*. Thus, a particular molecular graph of molecular weight  $m$  and another one of molecular weight  $m + 1$  seem to have roughly the same chance to exist as a known compound independently of  $m$  being even or odd.
- (ii) There is a wave-like pattern both in *MG* and in *BS* with a period of 14 mass units, the mass increment of a  $\text{CH}_2$  group. We may therefore speculate that all influences governing the number of molecular graphs in a family of a given number of C atoms will faithfully be reproduced in the next-higher homologous family.

A glance at Figure 1 reveals the most important fact: Without any doubt  $MG(m)$  will continue to exponentially increase for higher  $m$ , whereas  $BS(m)$  will eventually decrease for increasing  $m$ . The latter follows trivially from the fact that the number of known existing organic compounds, though increasing over time, is finite at any fixed time. Thus for increasing  $m$  the gap between *MG* and *BS*, being a factor of more than  $10^5$  for  $m = 150$  already, will further increase. We may

---

<sup>3</sup>NIST/EPA/NIH Mass Spectral Library, NIST '98 version, U.S. Department of Commerce, National Institute of Standards and Technology

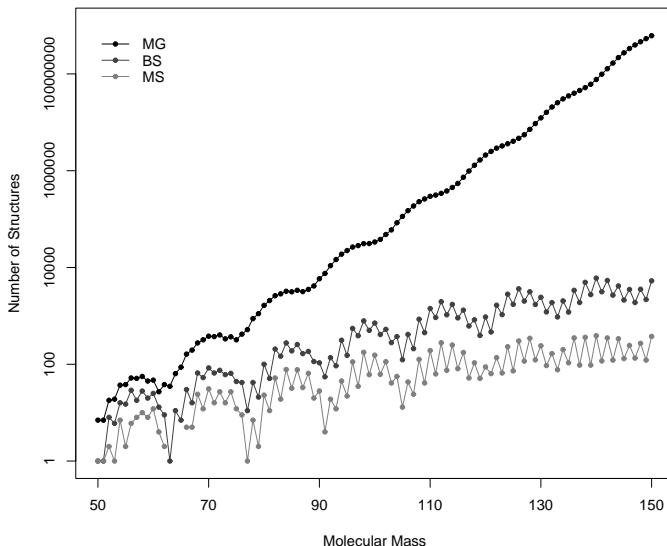


FIGURE 1. Graphical representation of the data from Table 2

conclude stating that the overwhelming majority of organic compounds is unknown.

How will the picture change if instead of C, H, N, O we consider the more important heteroelements such as Si, P, S, Halogens, as well, or even all elements? Rather than to repeat the whole analysis for a large set of elements, we performed a rough estimation for molecular graphs containing at least one C atom and no elements other than {C, H, N, O, <sup>28</sup>Si, P, <sup>32</sup>S, F, <sup>35</sup>Cl, <sup>79</sup>Br, I} in their standard valences (3 for P, 2 for S, 1 for Halogens), applying the same restrictions as above, and for molecular weights 100 and 150 only. The numbers of valid molecular formulae corresponding to  $m = 100$  and 150 are 83 and 628, respectively, the numbers of generated molecular graphs are 62105 and 1052647246. These numbers are lower bounds, due mostly to the unrealistic valence restrictions for P and S (Table 3).

On the other hand, from Beilstein (version BS0404PR, 12% increase in compounds compared to BS0302PR) we retrieved the uncharged one-component compounds corresponding to the valid molecular formulae for  $m = 100$  and 150, altogether 1620 and 10874 compounds, respectively (Table 3, column  $BS^*$ ). These numbers are upper bounds,

$m$	$MF$	$MG$	$BS^*$
100	83	62105	1620
150	628	1052647246	10874

TABLE 3. Numbers of valid molecular formulae, generated molecular graphs, and retrieved compounds for molecular weights 100 and 150, for the set of 11 elements

in that they are numbers of compounds (including radicals, stereoisomers, isotopically labeled compounds, compounds containing atoms in nonstandard valences, and duplicates) rather than of molecular graphs. Thus in going from the 4-element set to the 11-element set the numbers of possible molecular graphs and the numbers of molecular graphs representing known compounds increase by a remarkably constant factor of about 2 for both  $m = 100$  and 150. Thus for the 11-element set considered our conclusion remains valid, the overwhelming majority of organic compounds is unknown.

#### 4. EXAMPLE: $C_6H_6$ ISOMERS

The mathematically possible molecular graphs include those often considered chemically impossible due to all kinds of molecular pathologies, such as highly condensed small rings, double or triple bonds in small rings, chains threaded through small rings, and so on. Such structural elements will result in severe distortions of the usual atom and bond geometries and thus are associated with a high energy content. Such thermodynamic instability may cause an attempted synthesis to fail, in that the reactants are likely to find alternative less demanding ways of stabilization.

However, there are other possible reasons for an organic compound not to be known:

- (i) A compound may be kinetically unstable (not surrounded by high potential walls). Though it may form in an attempted synthesis, it tends to immediately decay and thus to remain elusive.
- (ii) Trivially, nobody may have attempted to synthesize it, though a synthesis would be perfectly feasible.
- (iii) Trivially, nobody may have looked for it, or not in the correct organism, though it exists as a natural compound in some organism.

Therefore one cannot expect a strict association between a compound's high/low energy content and its (non)existence as a known compound. Nevertheless, we were curious whether the energy content

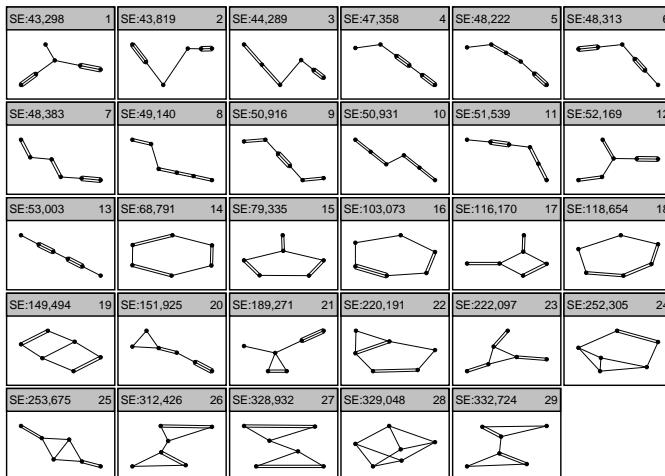


FIGURE 2.  $C_6H_6$  Isomers in the Beilstein file and their calculated steric energies

can serve at least within a family of isomers as a soft predictor of a structure's recognition as a known compound.

Of the 217 mathematically possible  $C_6H_6$  structures no more than 29 are registered in the Beilstein file. Neglect by researchers in this case cannot be the primary reason for the high ratio of unknowns, since the benzene isomers have elicited many theoretical and synthetic efforts [9, 10, 11].

A compound's position on an energy scale can be calculated by a number of methods. The most simple type of calculation is force field calculation (molecular mechanics)[12], and a rather primitive purely steric force field similar to, but even simpler than Allinger's MM2 [13], is incorporated in MOLGEN. The following results were obtained using this force field, therefore they are very crude approximations at best.

For each of the 217 isomers a 3D atom arrangement was constructed and energy-minimized starting from random initial atom coordinates. To avoid trapping in a local minimum, the procedure was repeated ten times for each isomer, starting from different random coordinates each time, and the overall steric energy minimum obtained was kept. In Figure 2 the 29 Beilstein  $C_6H_6$  isomers are displayed in the order of their steric energy values. Note that the 13 lowest-energy isomers according to that force field are the open-chain hexadiynes, hexadienynes, and hexatetraenes. The next higher energy isomer among those known is

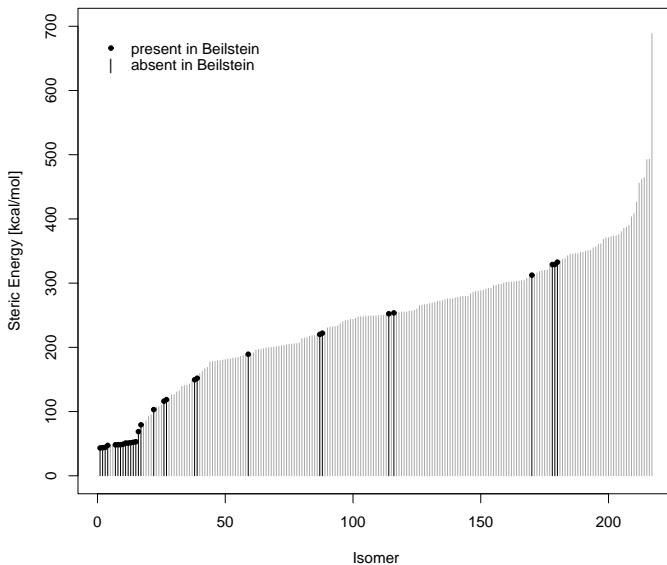
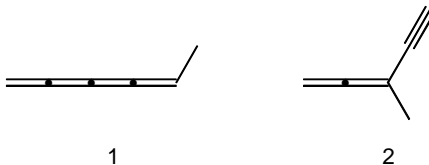


FIGURE 3. Steric energy (force field) of the  $C_6H_6$  isomers, in the order of increasing value

cyclohexa-1,3,5-triene or benzene (the force field does not know the concept of aromatic stabilization) followed by fulvene. The isomers highest in steric energy among those known are the three bi(cyclopropenyl) isomers and prismane.

In Figure 3 the steric energy values of all 217 isomers are shown, the 29 Beilstein compounds are indicated in black. Almost all of the low energy, open-chain isomers are known (13 out of 15), the exceptions being hexa-1,2,3,4-tetraene (**1**) and 3-methylpenta-1,2-dien-4-yne (**2**).



The remaining 16 known  $C_6H_6$  isomers are cyclic or polycyclic and are scattered over the intermediate energy range. The highest energy range is not populated by any known isomer. Not unexpectedly,

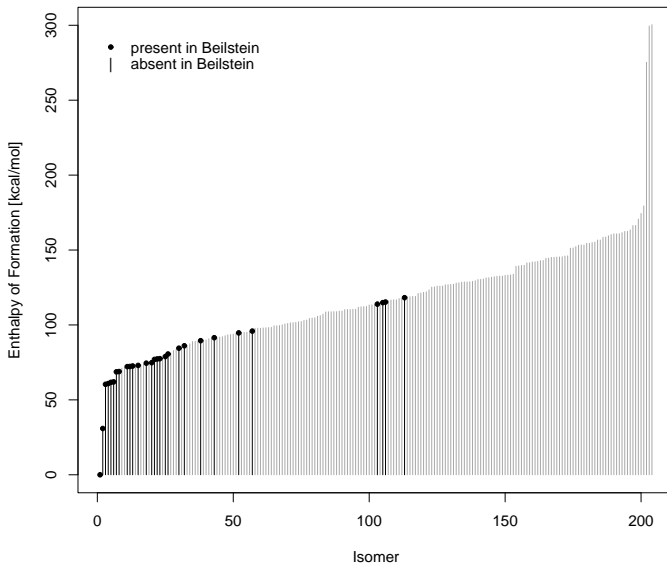
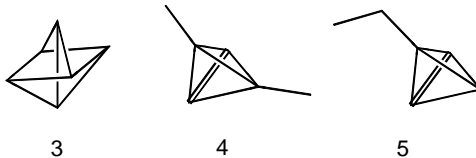


FIGURE 4. Enthalpy of formation of the  $C_6H_6$  isomers according to reference [11], in the order of increasing value. Enthalpy values are given relative to benzene.

the highest-energy isomers (by this force field) are the only graph-theoretically nonplanar [14] isomer tetracyclo[2.2.0.0<sup>2,5</sup>.0<sup>3,6</sup>]hexane (**3**) and the tetrahedrenes **4** and **5**. Lowest in steric energy among those still unknown are cyclohexa-1,2,4-triene, cyclohex-4-en-1-yne, and 1- and 3-ethynylcyclobutene.



Very recently, high-level computations (ab initio and density functional theory) were performed for the 217  $C_6H_6$  isomers [11]. The authors provided numerical values of the enthalpy of formation for all 204 isomers found to be minima. For comparison, their results are displayed in Figure 4 in the same manner as ours in Figure 3. While

the known compounds are slightly better concentrated in the left-hand part of Figure 4 than of Figure 3, both types of calculations agree that tetrahedrenes **4** and **5** are extremely high in energy (**3** turned out not to be an energy minimum in [11] and therefore is absent from Figure 4) and that the three bi(cyclopropenyl) isomers together with prismane form a group of high-energy isomers quite apart from the other known isomers. There is also agreement that along with **1** and **2** cyclohexa-1,2,4-triene, cyclohex-4-en-1-yne, and the ethynylcyclobutenes should be accessible by present-day synthetic methods.

## REFERENCES

- [1] A. Kerber, R. Laue, M. Meringer, and C. Rücker. *Molecules in Silico: The Generation of Structural Formulae and its Applications*. J. Comput. Chem. Jpn., 3:85–96, 2004.
- [2] R. Grund. *Konstruktion molekularer Graphen mit gegebenen Hybridisierungen und überlappungsfreien Fragmenten*. Bayreuther Mathematische Schriften, 49:1–113, 1995.
- [3] C. Benecke, R. Grund, R. Hohberger, R. Laue, A. Kerber, and T. Wieland. *MOLGEN+, a Generator of Connectivity Isomers and Stereoisomers for Molecular Structure Elucidation*. Anal. Chim. Acta, 314:141–147, 1995.
- [4] C. Benecke, T. Grüner, A. Kerber, R. Laue, and T. Wieland. *Molecular Structure Generation with MOLGEN, new Features and Future Developments*. Fresenius J. Anal. Chem., 358:23–32, 1997.
- [5] R. C. Read. *Everyone a Winner*. Annals of Discrete Mathematics, 2:107–120, 1978.
- [6] C. J. Colborn and R. C. Read. *Orderly Algorithms for Generating Restricted Classes of Graphs*. J. Graph Theory, 3:187–195, 1979.
- [7] J. Braun, R. Gugisch, A. Kerber, R. Laue, M. Meringer, and C. Rücker. *MOLGEN-CID, A Canonizer for Molecules and Graphs Accessible through the Internet*. J. Chem. Inf. Comput. Sci., 44:642–548, 2004.
- [8] M. Meringer. *Mathematische Modelle für die kombinatorische Chemie und die molekulare Strukturauflösung*. Logos-Verlag Berlin, 2004.
- [9] I. Gutman and J. H. Potgieter. *Isomers of Benzene*. J. Chem. Educ., 71:222–224, 1994.
- [10] J. Jeevanandam and R. Gopalan. *MNDO Method of Studies of Isomers of C<sub>6</sub>H<sub>6</sub>*. J. Indian Chem. Soc., 73:109–112, 1996.
- [11] T. C. Dinadayalane, U. D. Priyakumar, and G. N. Sastry. *Exploration of C<sub>6</sub>H<sub>6</sub> Potential Energy Surface: A Computational Effort to Unravel the Relative Stabilities and Synthetic Feasibility of New Benzene Isomers*. J. Phys. Chem. A, 108:11433–11448, 2004.
- [12] N. L. Allinger, J. R. Maple, and T. A. Halgren. In P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. A. Kollman, and H. F. Schaefer, editors, *Encyclopedia of Computational Chemistry*, volume 2, pages 1013–1035. Wiley, Chichester, 1998.
- [13] N. L. Allinger. *MM2. A Hydrocarbon Force Field Utilizing V<sub>1</sub> and V<sub>2</sub> Torsional Terms*. J. Am. Chem. Soc., 99:8127–8134, 1977.
- [14] C. Rücker and M. Meringer. *How Many Organic Compounds are Graph-Theoretically Nonplanar?* MATCH Commun. Math. Comput. Chem., 45:153–172, 2002.