MOLGEN-CID - A Canonizer for Molecules and Graphs Accessible through the Internet

Joachim Braun, Ralf Gugisch, Adalbert Kerber, Reinhard Laue, Markus Meringer, and Christoph Rücker*

Department of Mathematics,
University of Bayreuth,
D-95440 Bayreuth, Germany

The MOLGEN Chemical Identifier MOLGEN-CID is a software module freely accessible via the Internet. For a molecule or graph entered in molfile format it produces, by a canonical renumbering procedure, a canonical molfile and a unique character string that is easily compared by computer to a similar string. The mode of operation of MOLGEN-CID is detailed and visualized with examples.

INTRODUCTION

A chemical compound should be unambiguously identifiable by a unique label. For decades structure-describing traditional chemical nomenclature served this purpose more or less well. However, with compounds under study becoming more and more complex chemical names also became ever more complex, as a result most chemical names now are lengthy, difficult to pronounce and unwieldy. In chemists' everyday-life chemical names were therefore superseded by structure drawings, considered by many the natural language of molecular science. On the other hand, a structure can be drawn in various ways, such that there is no 1:1 correspondence between a compound and a particular drawing. Further, the atoms in a structure drawing may be numbered in many ways (n! numberings for a compound containing n atoms), so that derived computer representations (connection tables, adjacency matrices), though unambiguous, are not unique.

For some time registry numbers seemed to be a solution to the problem, at least for the bench chemist and the layman, in that a new registry number (CAS-RN or BRN) is attributed to a compound when it is first registered by Chemical Abstracts Service or Beilstein. This number then serves as the compound's unique ID. This procedure, of course, leaves to the agency the problem to compare a seemingly new compound to all those already present in the database.[1] As a further principal limitation, for an unpublished compound a RN is not available.

Nowadays, in the computer age and the time of combinatorial chemistry, when chemical companies and even individuals establish their own databases of real or virtual compounds and reactions, the problem of identifying compounds has become more urgent than ever. The problem can be described as the problem of canonization, that is to attribute to a compound, by a set of rules, a standard representation, a unique character string easily comparable by computer or manually to the corresponding strings of other compounds. This is equivalent to producing a unique numbering of the atoms in a molecule, a canonical numbering. Any molecule generator software such as MOLGEN[2] or SMOG[3] necessarily contains a canonizer to avoid redundant generation.

Many canonization methods have been proposed. For procedures described early in the chemical literature see the paper by Jochum and Gasteiger and references cited therein.[4] Randic considered that adjacency matrix canonical that results in the minimum binary number

* Corresponding author phone: +49 921 553386; fax: +49 921 553385; e-mail: christoph.ruecker@uni-bayreuth.de

when the rows of its upper half are concatenated.[5] Hendrickson instead used the maximum number obtained from the upper half matrix.[6] Kvasnicka and Pospichal prefer the maximum number obtained from the lower half matrix.[7] Though such an extremality requirement obviously leads to a unique numbering, extremality is not necessary. Rather, the goal may be achieved by one out of many procedures, provided it is well-defined, i.e. in application to each particular graph (molecule) it does not leave room for arbitrariness. New canonization procedures are still being developed.[8,9]

An often used but inferior method to discriminate molecules is by means of graph invariants, numbers obtained from a structure in some well-defined way. Similarly, the atoms in a molecule may often be distinguished using vertex-in-graph invariants. The most important procedure of this kind probably is the Morgan algorithm, in which the atoms in a molecule are distinguished by their extended connectivities, numbers obtained by repeated summation of the connectivity values over all neighbors of a particular atom. This method still seems to be the basis of the Chemical Abstracts registry system.[10] An improved version was proposed by Balaban, Mekenyan and Bonchev.[11] The Weiningers published a method largely based on graph invariants to obtain a unique form of a SMILES notation.[12] Though graph invariant-based methods sometimes work surprisingly well,[13] all graph invariants are degenerate, i.e. there are nonisomorphic graphs (nonidentical molecules) having the same numerical value of a particular graph invariant or even identical values for a combination of several graph invariants. This problem even today is occasionally ignored.[14]

The real merit of graph invariants in the present context is that they often allow the nonidentity of two compounds to be easily perceived without the need for a rigorous isomorphism test. Similarly, vertex-in-graph invariants, though sometimes identical for nonequivalent vertices, often allow easy perception of the nonidentity of graph vertices, whereby an ensuing rigorous canonization is rendered far less difficult.

The extraordinary value of a canonizer became apparent to us again when we recently found that even among simple graphs of no more than 8 vertices, there are some that cannot be differentiated by the highly discriminant combination of Balaban's index J and distance matrix eigenvalues. For the MOLGEN canonizer it was no problem at all to resolve these degeneracies.[15]

Recently, the International Union of Pure and Applied Chemistry (IUPAC) has recognized the need for a canonization procedure available to every chemist, and is undergoing a major effort to develop a corresponding software tool, the IUPAC Chemical Identifier, IChI.[16] At present the project is in the β test phase, software may be obtained from Stephen Stein or Stephen Heller, the developers, for local installation and use.[17] Details of their procedure are not yet published.

RESULTS AND DISCUSSION

In the present article we report on the MOLGEN Chemical Identifier (MOLGEN-CID), a software installed at the University of Bayreuth and freely accessible to everyone for use via the Internet.[18,19] In short, a (molecular or non-molecular) graph in molfile format (arbitrary initial numbering) is uploaded to MOLGEN-CID, a canonical numbering is performed, and a unique and unambiguous character string as well as a molfile are returned that describe the canonized structure. Since canonization, as a rule, will be sought for testing the identity of two compound representations, web pages taylored for this purpose are provided: Two molfiles uploaded separately are both canonized, and the resulting character strings are automatically compared, resulting in the answer "identical" or "nonidentical". Since many chemists will not be able to provide molfiles, a molecule can alternatively be drawn in the

freely available ACD molecule editor[20] which then transforms the drawing into a molfile to be processed as described.
For those not wishing to transfer structures via the Internet and for those wishing to canonize a whole database of compounds, an inhouse version of MOLGEN-CID is available.
By default, MOLGEN-CID works on hydrogen-suppressed graphs, at least if the molecule is entered as hydrogen-suppressed molfile or drawn without hydrogens in the ACD editor. Information on bond multiplicity is used from the beginning by MOLGEN-CID, while in IChI multiple bonds are removed before the canonization process is started.
The output character string from MOLGEN-CID by default does not contain hydrogens. The heavy atoms are given in the order of their canonical numbering, each atom is followed by a list of bonds of indicated kind (s = single, d = double, t = triple, a = aromatic) to its neighbors identified by their canonical numbers. The string is easily reconverted to the structure even manually. For example, the canonical strings for benzyl alcohol and anisole are
Os8Cs8a3a4Ca5Ca6Ca7Ca7CC and
Os2s8Ca3a4Ca5Ca6Ca7Ca7CC, respectively.
The benzyl alcohol string translates: There is an oxygen atom (number 1) that is singly bonded to atom number 8. Atom number 2 is carbon and has a single bond to atom 8 and aromatic bonds to atoms 3 and 4. Atom number 3 is carbon and has an aromatic bond to atom 5, and so on.

Canonization Procedure.

Step 1, initial classification. As in many other canonization procedures, our method starts with partitioning the graph vertices into classes according to some vertex-in-graph invariants. The purpose of this step is to restrict the number of numberings to be considered from n! to $n_1!\cdot n_2!\cdot\ldots\cdot n_k!$, where n_1, n_2, …, n_k are the cardinalities of the first, second, …, kth vertex class, so that $n_1 + n_2 + \ldots + n_k = n$.
The criteria used for initial classification are easily obtained non-numerical and numerical vertex properties. They are hierarchically ordered as follows:

1. Nature of an atom (C, N, O, …). All atoms of a higher atom number in the periodic system have priority over (will get lower canonical numbers than) all atoms of a lower atom number.
2. Atom attributes such as an atomic mass other than default (isotope), a charge other than zero, an unpaired electron (free radical), or a valency other than default (e.g. the default valency for carbon is four, including bonds to hydrogen atoms).
3. Ring or chain nature. Ring atoms have priority over chain atoms.
4. For chain atoms their skeleton/nonskeleton property. A chain connecting two rings is considered part of the molecular skeleton, in contrast to a side chain which is not. An atom in a skeleton chain has priority over an atom in a side chain.
5. The number of aromatic, triple, double and single bonds (not counting those to hydrogen) in which an atom is engaged, in this order. E.g., a carbon atom engaged in three aromatic bonds has priority over one having two aromatic bonds, a carbon atom in a triple bond has priority over a central allenic C atom which has priority over a carbon engaged in one double bond, and a C atom with four single bonds to non-hydrogen atoms has priority over those with three, two, or one such bonds.

Step 2, iterative refinement. The initial classification is iteratively refined according to each atom's immediate neighbors, as far as a neighbor is already "unique" (forms a class for itself).[21,22] Each unique atom in turn is used to split nonunique classes, and each atom becoming unique thereby joins the queue to be used itself.
By steps 1 and 2 a discrete partition is often obtained, in particular for molecular graphs.

Example. Consider the structure of the Pymetrozine analogue **1** shown in Figure 1 as hydrogen-suppressed graph with an arbitrary initial vertex numbering. Its treatment is indicated in Scheme 1.

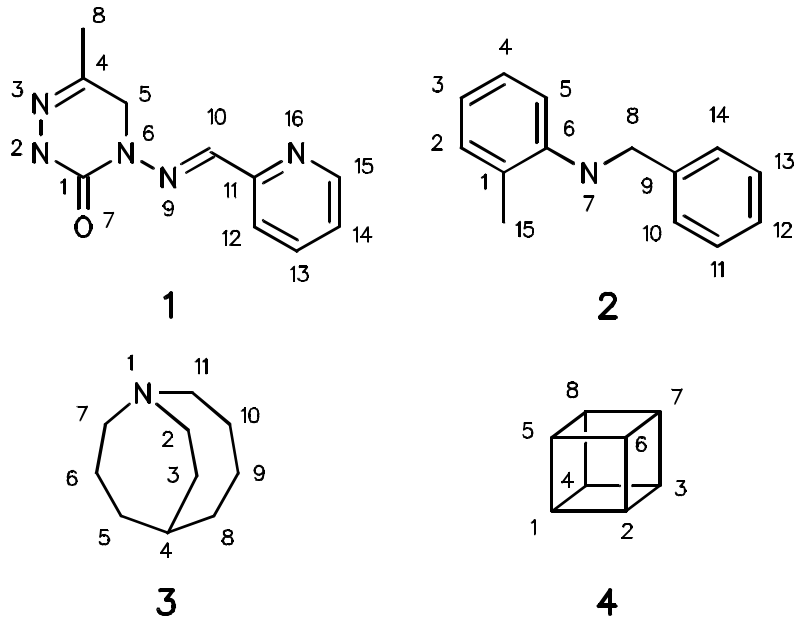

Figure 1: Compounds 1 – 4 used as examples, with arbitrary initial vertex numbering.

```
Pymetrozine analogue 1
Initial numbers 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

partition by
criterion 1    |7|2 3 6 9 16|1 4 5 8 10 11 12 13 14 15|
criterion 3    |7|2 3 6 16|9|1 4 5 11 12 13 14 15|8 10|
criterion 4    |7|2 3 6 16|9|1 4 5 11 12 13 14 15|10|8|
criterion 5    |7|16|3|6|2|9|11|12 13 14 15|1 4|5|10|8|  initial
classification

refined by 7   |7|16|3|6|2|9|11|12 13 14 15| 1| 4| 5|10|8| newly
unique: |1|4|
refined by 16  |7|16|3|6|2|9|11|15|12 13 14| 1| 4| 5|10|8| newly
unique: |15|
refined by 11  |7|16|3|6|2|9|11|15|12|13 14| 1| 4| 5|10|8| newly
unique: |12|
refined by 15  |7|16|3|6|2|9|11|15|12|14|13| 1| 4| 5|10|8| newly
unique:|14|13|
               ↓  ↓ ↓ ↓ ↓ ↓  ↓  ↓  ↓  ↓  ↓  ↓  ↓  ↓  ↓  ↓ ↓
Canon. numbers  1  2 3 4 5 6  7  8  9 10 11 12 13 14 15 16
```

Scheme 1

Step 1: The vertex classification obtained using criteria 1 and 3-5 is given in the first lines in Scheme 1. Criterion 2 is of no use in this example. In the partition so obtained ("initial classification") there are only two classes containing more than one atom, one comprised of atoms 12-15, the other of atoms 1 and 4.

Step 2: Unique atom 7 allows splitting of atom 1, a neighbor, from 4, not a neighbor. Unique atom 16 allows splitting of 15, its neighbor, from 12-14. Unique atoms 3,6,2,9 do not lead to any further splitting. Unique atom 11 allows to split 12 from 13,14. Unique atoms 5,10,8,1,4 do not split the remaining pair. Finally, unique atom 15 allows to split 14, its neighbor, from 13. Now the partition is discrete, and canonical numbers are assigned as shown in the last two lines of Scheme 1.

Step 3, backtracking. If a discrete partition is not yet achieved, either for insufficient resolving power of steps 1-2,[23] or for symmetry equivalence of certain vertices, discrete partitions (numberings) not contradicting the initial classification are generated by a backtracking procedure. The first class of lowest cardinality >1 is chosen,[24] and an arbitrarily selected vertex in it is artificially marked to be preferred and is made the root of a branch. By this distinction of a particular vertex other vertices may become distinguishable, so that again by iterative classification a finer partition is obtained. Step 3 is recursively repeated until a discrete partition is achieved (a depth-first search) by backtracking, marking an atom, and iterative refinement applied in turn. Backtracking ensures that at each branching point (in principle) each eligible atom is marked and treated at some time in the process, so that, in fact, there is no arbitrariness.[25]

N-benzyl-*o*-toluidine **2**

Initial numbers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
step 1	|7|	1	6	9	|	2	3	4	5	10	11	12	13	14	| 8 |15|
step 2	|7|	6	|9|	1	|5|	10	14	|2|	4	|3|	11	12	13	| 8 |15|	
btl1, 10 marked	|7|	6	|9|	1	|5|	10	|14|	2	4	|3|	11	12	13	| 8 |15|	
refined by 10	|7|	6	|9|	1	|5|	10	|14|	2	4	|3|	11	|12|	13	| 8 |15|	
refined by 14	|7|	6	|9|	1	|5|	10	|14|	2	4	|3|	11	|13|	12	| 8 |15|	*1
backtrack																
btl1, 14 marked	|7|	6	|9|	1	|5|	14	|10|	2	4	|3|	11	12	13	| 8 |15|	
refined by 14	|7|	6	|9|	1	|5|	14	|10|	2	4	|3|	13	|11|	12	| 8 |15|	
refined by 10	|7|	6	|9|	1	|5|	14	|10|	2	4	|3|	13	|11|	12	| 8 |15|	*2 a
candidate 1 kept	|7|	6	|9|	1	|5|	10	|14|	2	4	| 3	|11|	|13|	12	| 8 |15|	
	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	
Canon. numbers	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	

Scheme 2

Example. An unsubstituted phenyl residue is a typical case of both symmetry (two *ortho* and two *meta* atoms) and insufficient resolution of steps 1-2 (*meta* vs *para* position). In *N*-benzyl-*o*-toluidine **2** (Figure 1, Scheme 2) two unresolved classes remain after steps 1 and 2, one containing atoms 10 and 14, the other atoms 11-13 (arbitrary numbering given in Figure 1). The two-member class is chosen, and atom 10 is preliminarily marked on backtrack level 1 (btl1). Thereby atom 14 also becomes unique, and refinement by 10 and then 14 leads to a discrete partition, candidate 1 (*1) for canonical numbering. Backtracking and alternative marking of 14 followed by refinement results in another discrete partition which however leads to the same adjacency matrix as the first (an automorphism, the symmetry of the phenyl residue). Therefore the first candidate is kept and used for assigning canonical numbers, as shown.

Pruning the backtrack tree. It is of decisive importance to devise the procedure so that not all possible numberings have to be constructed, that on the contrary as many branches of the backtrack tree as possible are pruned. In our procedure, this goal is achieved by a combination of two features. First, for the comparison of candidate adjacency matrices an extremality criterion is used, maximization of the number obtained from concatenation of lines in the *lower* half of the matrix. This choice has the advantage that when entries in a certain line of the matrix are changed, the lines further up are not affected, i.e. the first digits of the number to be maximized are not changed thereby. Second, *for the purpose of comparing adjacency matrices* the atoms are re-numbered in the order of when an atom becomes unique in the process. Therefore, if a partial numbering results in a concatenated number smaller than the current favorite with respect to its first i digits, then any permutation in the remaining labels is unnecessary since it cannot change the first i digits, i.e. the backtracking tree is pruned at once.

```
1-azabicyclo[4.3.2]undecane 3
Initial numbers  1  2  3 4  5  6 7 8 9 10 11

                 |1|4| 2 3  5  6 7 8 9 10 11|
refined by 1     |1|4| 2 7 11| 3 5 6 8  9 10|
refined by 4     |1|4| 2 7 11| 3 5 8 |6 9 10|
btl1, 2 marked   |1|4|2| 7 11| 3 5 8 |6 9 10|
refined by 2     |1|4|2| 7 11|3|5 8 |6 9 10|
btl2, 7 marked   |1|4|2| 7|11|3|5 8 |6 9 10|
refined by 7     |1|4|2| 7|11| 3|5 8 |6|9 10|
refined by 11    |1|4|2| 7|11| 3|5 8 |6|10|9|
refined by 6     |1|4|2| 7|11| 3|5|8 |6|10|9|  *1
backtrack
btl2, 11 marked  |1|4|2|11| 7| 3|5 8 |6 9 10|
refined by 11    |1|4|2|11| 7| 3|5 8 |10|6 9|
refined by 7     |1|4|2|11| 7| 3|5 8 |10|6|9|
refined by 6     |1|4|2|11| 7| 3|5|8 |10|6|9|  *2
backtrack
btl1, 7 marked   |1|4|7|2 11|  3 5 8 |6 9 10|
refined by 7     |1|4|7|2 11|  3 5 8 |6|?  ?| pruned
backtrack
btl1, 11 marked  |1|4|11|2 7|  3 5 8 |6 9 10|
refined by 11    |1|4|11|2 7|  3 5 8 |10|? ?| pruned

candidate 2 kept |1|4|2|11|7| 3|5|8|10| 6| 9|
                  ↓ ↓ ↓  ↓ ↓  ↓ ↓ ↓  ↓  ↓  ↓
Canon. numbers    1 2 3  4 5  6 7 8  9 10 11
```

Scheme 3

Example. The hypothetical 1-azabicyclo[4.3.2]undecane **3** (Figure 1) has no symmetry. In Scheme 3 its treatment is given step by step. By criteria 1 and 5 atoms 1 and 4 become unique, respectively, all other atoms are in one class. Refinement by 1 and then by 4 allows some splitting but does not result in another unique atom. Therefore in the first class of lowest cardinality (2,7,11) atom 2 is artificially marked to become unique (backtrack level 1, btl1). Refinement by 2 results in atom 3 becoming unique. Refinement by 3 has no effect. Therefore now (backtrack level 2, btl2) in the first class of lowest cardinality (7,11) atom 7 is marked

unique, whereby atom 11 also becomes unique, and by refinement by 7 and then by 11 atoms 6, 10, and 9 also become unique. Refinement by 6 leads to the first discrete partition (candidate 1). In Scheme 3, in each line the atom(s) becoming unique is (are) *italicized*. Renumbering in the order of becoming unique gives the following mapping

```
initial numbering 1 4 2 3 7 11 6 10 9  5  8
renumbered        1 2 3 4 5  6 7  8 9 10 11,
```

corresponding to the following adjacency matrix:

```
   | 1  2  3  4  5  6  7  8  9 10 11
 ------------------------------------
 1 |
 2 | 0
 3 | 1  0
 4 | 0  1  1
 5 | 1  0  0  0
 6 | 1  0  0  0  0
 7 | 0  0  0  0  1  0
 8 | 0  0  0  0  0  1  0
 9 | 0  0  0  0  0  0  0  1
10 | 0  1  0  0  0  0  1  0  0
11 | 0  1  0  0  0  0  0  0  1  0
```

Now after backtracking to btl2 atom 11 is marked, whereby atom 7 also becomes unique. Refinement by 11 and then by 7 results in atoms 10, 6, and 9 becoming unique in this order. Thus now the partial renumbering scheme is

```
initial numbering 1 4 2 3 11 7 10 6 9
renumbered        1 2 3 4  5 6  7 8 9,
```

corresponding to the following partial adjacency matrix:

```
   | 1  2  3  4  5  6  7  8  9
 ---------------------------
 1 |
 2 | 0
 3 | 1  0
 4 | 0  1  1
 5 | 1  0  0  0
 6 | 1  0  0  0  0
 7 | 0  0  0  0  1  0
 8 | 0  0  0  0  0  1  0
 9 | 0  0  0  0  0  0  1  0
```

Here by entry “1” as matrix element (9,7) (*italic*) it becomes evident that the next discrete partion to be found, candidate 2, will be better (in the sense of our extremality criterion) than candidate 1. In fact, while refinement by 10 has no effect, refinement by 6 results in candidate 2, whose renumbering scheme and adjacency matrix are

```
initial numbering 1 4 2 3 11 7 10 6 9  5  8
renumbered        1 2 3 4  5 6  7 8 9 10 11
```
and

	1	2	3	4	5	6	7	8	9	10	11
1											
2	0										
3	1	0									
4	0	1	1								
5	1	0	0	0							
6	1	0	0	0	0						
7	0	0	0	0	1	0					
8	0	0	0	0	0	1	0				
9	0	0	0	0	0	0	1	0			
10	0	1	0	0	0	0	0	1	0		
11	0	1	0	0	0	0	0	0	1	0	

This renumbering scheme is kept as the currently best one.
Thereby btl2 is exhausted (Scheme 3), and after backtracking to btl1 atom 7 is marked, refinement by 7 results in atom 6 becoming unique, so that now the current partial renumbering scheme and partial adjacency matrix are as follows

```
initial numbering 1 4 7 6
renumbered        1 2 3 4
```

and

	1	2	3	4
1				
2	0			
3	1	0		
4	0	***0***	1	

Here entry “0” as matrix element (4,2) (*italic*) determines that all discrete partitions to be derived from this partial numbering will be worse than candidate 2. Therefore this part of the backtrack tree can immediately be pruned. In exactly the same manner the last alternative at btl1, marking atom 11 with atom 10 also becoming unique after refinement by 11, is found to be worse than candidate 2.
Figure 2 shows the backtrack tree corresponding to this example, pruned parts of the tree are drawn as broken lines.

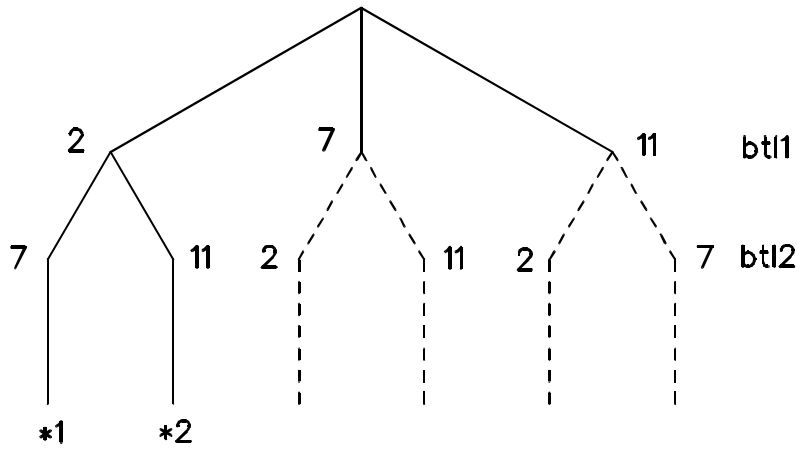

Figure 2: The backtrack tree for compound **3**. Parts of the tree that are not visited are drawn in broken lines.

Candidate 2 thus is the basis of the canonical numbering finally obtained as in the previous examples and shown at the bottom of Scheme 3.

Profitting from symmetry. Large parts of the search tree can be pruned in cases of higher symmetry. If two labelings obtained at different positions in the tree turn out to result in one and the same adjacency matrix, then a symmetry (automorphism) has been found. The information on automorphisms accumulating in the process finally defines the graph's or molecule's complete automorphism group. It is stored in the form of a set of generators (a Sims chain[26,27]). This information is used to cut parts of the backtrack tree found to be equivalent to others already visited.

cubane **4**									
Initial numbers	1	2	3	4	5	6	7	8	
	1	2	3	4	5	6	7	8	
btl1, 1 marked	***1***	2	3	4	5	6	7	8	
refined by 1	1	2	4	5	3	6	7	8	
btl2, 2 marked	1	***2***	4	5	3	6	7	8	
refined by 2	1	2	4	5	3	6	7	8	
btl3, 4 marked	1	2	***4***	***5***	3	6	7	8	
refined by 4	1	2	4	5	***3***	***6***	***8***	***7***	*1
backtrack									
btl3, 5 marked	1	2	***5***	***4***	3	6	7	8	
refined by 5	1	2	5	4	***6***	***3***	***8***	***7***	*2 a
backtrack									
btl2, 4 marked	1	***4***	2	5	3	6	7	8	
refined by 4	1	4	2	5	3	8	6	7	
btl3, 2 marked	1	4	***2***	***5***	3	8	6	7	
refined by 2	1	4	2	5	***3***	***8***	***6***	***7***	*3 a
backtrack									
btl2, 5 marked	1	***5***	2	4	3	6	7	8	
refined by 5	1	5	2	4	6	8	3	7	
btl3, 2 marked	1	5	***2***	***4***	6	8	3	7	
refined by 2	1	5	2	4	***6***	***8***	***3***	***7***	*4 a
backtrack									
btl1, 2 marked	***2***	1	3	4	5	6	7	8	
refined by 2	2	1	3	6	4	5	7	8	
btl2, 1 marked	2	***1***	3	6	4	5	7	8	
refined by 1	2	1	3	6	4	5	7	8	
btl3, 3 marked	2	1	***3***	***6***	4	5	7	8	
refined by 3	2	1	3	6	***4***	***5***	***7***	***8***	*5 a
backtrack									
btl1, 3 marked	***3***	1	2	4	5	6	7	8	
etc.									
candidate 1	1	2	4	5	3	6	8	7	
	↓	↓	↓	↓	↓	↓	↓	↓	
Canon. numbers	1	2	3	4	5	6	7	8	

Scheme 4

Example. The cubane molecule **4** (Figure 1) is highly symmetric. In Scheme 4 its treatment in our procedure is shown. Steps 1 and 2 do not achieve any splitting. Step 3 by marking atoms

1, 2, and 4 in btl1, btl2, and btl3, respectively, soon finds candidate 1 (*1 in Figure 3). At this stage the renumbering scheme (renumbering atoms in the order of their becoming unique) and adjacency matrix are

```
initial numbering 1 2 4 5 3 6 8 7
renumbered        1 2 3 4 5 6 7 8    and
```

	1	2	3	4	5	6	7	8
1								
2	1							
3	1	0						
4	1	0	0					
5	0	1	1	0				
6	0	1	0	1	0			
7	0	0	1	1	0	0		
8	0	0	0	0	1	1	1	

By backtracking, marking atom 5 on btl3, and refinement by 5 candidate 2 is found, the renumbering scheme now is

```
initial numbering 1 2 5 4 6 3 8 7
renumbered        1 2 3 4 5 6 7 8,
```

wherefrom the same adjacency matrix as before originates, an automorphism is found, the leftmost “a” in Figure 3.

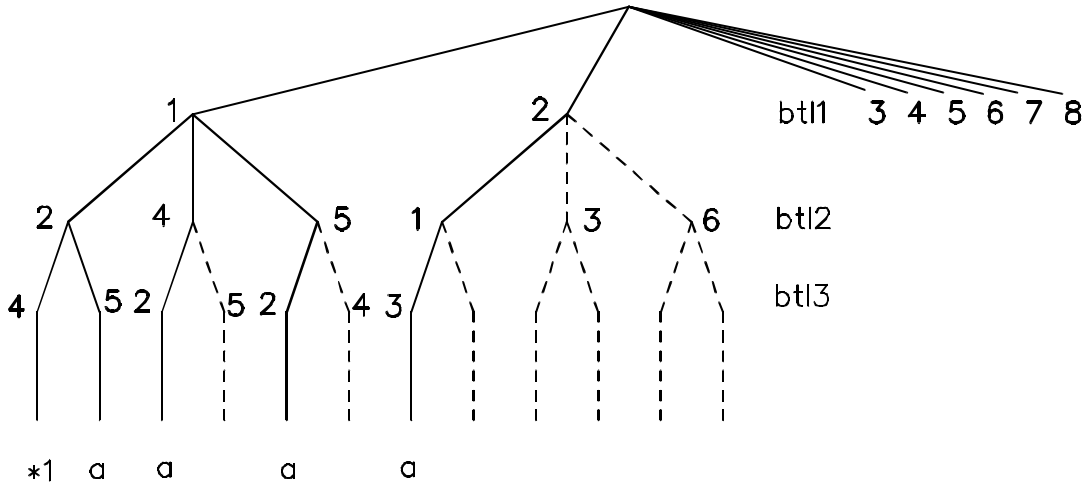

Figure 3: The backtrack tree for compound **4**, cubane (schematic). From each vertex 3-8 on btl1 branches are pending as from vertex 2.

Backtracking, marking atoms 4 and 2 on btl2 and btl3, respectively, leads to candidate 3, which again produces the same adjacency matrix as candidate 1 (second “a” in Figure 3). This automorphism derived from atom 4 marked on btl2 means that there must be another automorphism to be found as a branch originating in that node of the backtrack tree, just as there are two automorphic leaves below atom 2 on btl2. Therefore that whole branch of the tree can be pruned.
Backtracking and marking atom 5 on btl2 results in candidate 4, again automorphic to candidate 1, and as before a branch is now pruned.

Backtracking to btl1, marking atom 2, etc. finds candidate 5, again automorphic to candidate 1 (fourth "a" in Figure 3). It follows that all branches originating in atom 2 on btl1 must be equivalent to all branches originating from atom 1 on btl1, they are therefore pruned. For atoms 3,4,5,6,7,8 on btl1 things are exactly as for atom 2 on btl1. Thus candidate 1 is kept as best till the end of the procedure.
Note that, as evident from the above examples, in our method renumbering schemes only are stored, not matrices, resulting in a rather low memory requirement.
A detailed description of the canonization procedure was given earlier in German.[22]

Scope and Limitations. At present MOLGEN-CID treats covalently bonded compounds only, made either of one or of several components (connected or disconnected undirected graphs). Stereoisomerism is not yet treated. MOLGEN-CID is not restricted to molecular graphs, in particular, vertex degrees are not restricted to ≤ 4.[28]

Tests. For test purposes, the vertices of many graphs were routinely renumbered randomly five times, and in all cases all five renumbered graphs resulted in the same canonical numbering.
Databases such as the NIST Mass Spectral Library (107216 organic compounds, 5943 duplicates or stereoisomeric pairs detected) or the Maybridge Combinatorial Chemistry Database (MayDec02CCeus, 13410 compounds, 19 such cases found) were processed by MOLGEN-CID.
All the pairs of hard-to-distinguish molecules or graphs appearing in references 13b and 15 were correctly found nonidentical by MOLGEN-CID. Conversely, different drawings of the same graph (2 nontrivial cases in reference 15) were correctly identified.
Molecular graphs, as a rule, contain vertices easily differentiated (heteroatoms) and often edges easily differentiated (multiple bonds). Most molecular graphs contain rather few bonds or cycles compared to the number of atoms, and thus most molecular graphs are planar graphs.[29] All this adds to molecular graphs being rather easy to handle for canonization, symmetry perception and isomorphism test algorithms. In the words of a classic: "… most graphs present no great problem even to badly designed algorithms. The test of a graph isomorphism algorithm is how it behaves under 'worst case' conditions, i.e. how it handles the really recalcitrant graphs …"[30] Samples of such really recalcitrant mathematical graphs were compiled by Weisfeiler[31] and Mathon[32] to challenge such algorithms. These are graphs without multiple edges or special vertices, of high vertex degrees, many of them regular (all vertices of the same degree) and of high or seemingly high symmetry. These graphs were used for a further test of MOLGEN-CID.

The 20 Mathon graphs contain between 25 and 50 vertices of degrees up to 16, among them 14 regular graphs, e.g. there are 8 regular graphs of 29 vertices of degree 14. The 39 Weisfeiler graphs are all regular: They are made of 10–28 vertices of degree 3–12, e.g. there are 15 regular graphs of 25 vertices of degree 12. These 59 graphs were canonized by MOLGEN-CID within 14.0 sec on an Athlon XP1600 PC, 1.4 GHz, and no doublettes were shown by MOLGEN-CID to exist within both the Mathon and the Weisfeiler sample. However, across the samples two doublettes were correctly found:[33] Mathon's graph A^1_{25} is identical to Weisfeiler's graph 251210, and Mathon's B^1_{25} is identical to Weisfeiler's 25123.[13a,34] Further, Weisfeiler's graphs 1662 and 1661 were found isomorphic to Shrikhande's graph[35] and its twin, respectively. These latter two graphs are depicted in reference 15.

REFERENCES AND NOTES

(1) If this process, due to a lack of proper canonization, goes astray, then a compound will be registered under more than one registry numbers. Such happens both in the CAS Registry ("alternate registry number", "deleted registry number") and notoriously in the Beilstein system. The reverse may also happen: Two similar compounds may erroneously be identified in a database. Such cases are more difficult to detect.
(2) (a) Wieland, T.; Kerber, A.; Laue, R. Principles of the Generation of Constitutional and Configurational Isomers. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 413-419. (b) Benecke, C.; Grüner, T.; Kerber, A.; Laue, R.; Wieland, T. MOLecular Structure GENeration with MOLGEN, new Features and Future Developments. *Fresenius' J. Analyt. Chem.* **1997**, *359*, 23-32.
(3) Molchanova, M.S.; Zefirov, N.S. Irredundant Generation of Isomeric Molecular Structures with Some Known Fragments. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 8-22.
(4) Jochum, C.; Gasteiger, J. Canonical numbering and Constitutional Symmetry. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 113-117.
(5) Randić, M.; Brissey, G.M.; Wilkins, C.L. Computer Perception of Topological Symmetry via Canonical Numbering of Atoms. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 52-59.
(6) Hendrickson, J.B.; Toczko, A.G. Unique Numbering and Cataloguing of Molecular Structures. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 171-177.
(7) (a) Kvasnicka, V.; Pospichal, J. Canonical Indexing and Constructive Enumeration of Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 99-105. (b) Kvasnicka, V.; Pospichal, J. An Improved Method of Constructive Enumeration of Graphs. *J. Math. Chem.* **1992**, *9*, 181-196.
(8) Broadbelt, L.J.; Stark, S.M.; Klein, M.T. Computer Generated Reaction Modelling: Decomposition and Encoding Algorithms for Determining Species Uniqueness. *Comput. Chem. Engng.* **1996**, *20*, 113-129. (9) Warth, V.; Battin-Leclerc, F.; Fournet, R.; Glaude, P.A.; Côme, G.M.; Scacchi, G. Computer Based Generation of Reaction Mechanisms for Gas-Phase Oxidation. *Comput. & Chem.* **2000**, *24*, 541-560.
(10) Morgan, H.L. The Generation of a Unique Machine Description for Chemical Structures – A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107-113.
(11) Balaban, A.T.; Mekenyan, O.; Bonchev, D. Unique Description of Chemical Structures based on Hierarchically Ordered Extended Connectivities (HOC Procedures). *J. Comput. Chem.* **1985**, *6*, 538-551.
(12) Weininger, D.; Weininger, A.; Weininger, J.L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97-101.
(13) (a) Rücker, G.; Rücker C. On Using the Adjacency Matrix Power Method for Perception of Symmetry and for Isomorphism Testing of Highly Intricate Graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 123-126. (b) Rücker, G.; Rücker, C. On Finding Nonisomorphic Connected Subgraphs and Distinct Molecular Substructures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 314-320.
(14) Ratkiewicz, A.; Truong, T.N. Application of Chemical Graph Theory for Automated Mechanism Generation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 36-44.
(15) Rücker, C.; Rücker, G.; Meringer, M. Exploring the Limits of Graph Invariant- and Spectrum-Based Discrimination of (Sub)structures. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 640-650.
(16) Freemantle, M. Unique Labels for Compounds. *Chem. Engineer. News* December 2, **2002**, 33-35.
(17) stephen.stein@nist.gov , stephen.heller@nist.gov
(18) www.mathe2.uni-bayreuth.de/molgen4
(19) Various preliminary versions of the canonizer have been available at our website for

more than 2 years.
(20) www.acdlabs.com/download/chemsk.html
(21) McKay, B.D. Computing Automorphisms and Canonical Labelling of Graphs. Pages 223-232 in *Proc. Intern. Conf. on Combinatorial Theory*, Lecture Notes in Mathematics No. 686; Springer: Berlin 1977.
(22) Benecke, C. Objektorientierte Darstellung und Algorithmen zur Klassifizierung endlicher bewerteter Strukturen, *MATCH – Commun. Math. Comput. Chem.* **1998**, *37*, 7-156.
(23) Classification by iterative refinement (steps 1,2) can be made even more powerful by slight variations in the procedure. Thus use of further vertex-in-graph invariants is an obvious option. In step 2 instead of the immediate neighborship, relations of longer distance or even of all distances could be used for better discrimination. This would require to initially construct and then evaluate the graph's distance matrix. Further, the restriction of unique atoms only to be used for refinement could be alleviated. This alternative, however, was shown often not to be of advantage (reference 22). We decided not to exploit iterative classification to its limits since backtracking is needed anyway for cases of symmetry.
(24) Knuth, D.E. Estimating the Efficiency of Backtrack Programs. *Mathematics of Computation* **1975**, *29*, 121-136.
(25) The refinement resulting from artificially marking a vertex usually reduces the number of alternatives on the next backtrack level, thus preventing exponential growth of the backtrack tree.
(26) (a) Sims, C. Computation with Permutation Groups. *Proc. Second Symp. on Symbolic and Algebraic Manipulation* (Petrick, S.R., Ed.), Assoc. Comput. Mach., Los Angeles 1971, pages 23-28. (b) Jerrum, M. A Compact Representation for Permutation Groups. *J. Algorithms* **1986**, *7*, 60-78.
(27) Grund, R. Symmetrieklassen von Abbildungen und die Konstruktion von diskreten Strukturen, *Bayreuther Mathematische Schriften* **1990**, *31*, 19-54.
(28) There are applications even in organic chemistry where undirected graphs containing vertices of degree >4 have to be canonized, see e.g.
so-called macroatoms of high degree in the former DENDRAL project and in MOLGEN 3.5.
(29) Rücker, C.; Meringer, M. How Many Organic Compounds are Graph-Theoretically Nonplanar? *MATCH – Commun. Math. Comput. Chem.* **2002**, *45*, 153-172.
(30) Read, R.C.; Corneil, D.G. The Graph Isomorphism Disease. *J. Graph Theory* **1977**, *1*, 339-363.
(31) Weisfeiler, B. *On Construction and Identification of Graphs*; Lecture Notes in Mathematics No. 558; Springer: Berlin 1976.
(32) Mathon, R. Sample Graphs for Isomorphism Testing. *Proc. 9th S.-E. Conf. Combinatorics, Graph Theory and Computing*, **1978**, 499-517 *(Congressus Numerantium 21)*.
(33) Walter, C.D. Adjacency Matrices. *SIAM J. Alg. Disc. Methods* **1986**, *7*, 18-29.
(34) Surprisingly, the identity of such graphs can successfully be tested (though much more slowly) by application of traditional chemical nomenclature. Thus Weisfeiler's graph 1561 (15 vertices of degree 6) is hentriacontacyclo[7.6.0.0^{1,3}.0^{1,5}.0^{1,6}.0^{2,7}.0^{2,8}.0^{2,11}.0^{2,12}.0^{3,10}.0^{3,13}.0^{3,14}.0^{4,7}.0^{4,10}.0^{4,12}.0^{4,15}.0^{5,8}.0^{5,12}.0^{5,14}.0^{6,10}.0^{6,11}.0^{6,13}.0^{7,13}.0^{7,15}.0^{8,10}.0^{8,14}.0^{9,12}.0^{9,13}.0^{9,15}.0^{11,14}.0^{11,15}]pentadecane. The uniqueness of these so-called von-Baeyer names is ensured by requirements such as "The main ring shall contain as many carbon atoms as possible", "The main bridge shall be as large as possible", "The main ring shall be divided as symmetrically as possible by the main bridge", "The superscripts locating the other bridges shall be as small as possible …" (rule A-32, International Union of Pure and Applied Chemistry, *Nomenclature of Organic Chemistry, Sections A,B,C,D,E,F,H*, Pergamon Press, Oxford, 1979). The above name was composed by the program POLCYC modified for vertex degrees higher than four, see Rücker, G.; Rücker, C. Nomenclature of Organic Polycycles out of the Computer – How to Escape the Jungle of the Secondary Bridges. *Chimia* **1990**, *44*, 116-120.

(35) Shrikhande, S.S. On a Characterization of the Triangular Association Scheme. *Ann. Math. Statist*. **1959**, *30*, 39-47.