match no. 27 pp. 49-85 1992

Structure Generator GEN

Simona Bohanec, Jure Zupan

Institute of Chemistry, Hajdihova 19, 61115 Ljubljana, Slovenia

(received: June 1992)

Abstract

A structure generator for the generation of all structures from given structural fragments is presented. The applications of two different systems that include the described generator will be described and discussed. These two systems can use up to 30 structural fragments with free bonds defined to a different extent of completeness. The generator can consider different constraints like molecular weight, molecular formula, or structural fragments, etc. The generation is fast, exhaustive, irredundant, and effective. Our generator is compared to some other generators[1-3] with respect to the number of generated output structures and required generation times. The generator is implemented on IBM personal computers and compatibles.

1 Introduction

The structure generation problem is one of the most common problems in chemist's everyday work. A lot of analytical methods in the organic chemistry are based on the decomposition of the sample on smaller structural parts, on the combination of structural fragments, or on the suggestion of some. These structural parts, called also fragments, are usually analysed separately.

The size of fragments depends on the analytical technique which is performed. The analysis of elements gives the smallest fragments. These fragments consist of only one atom and have no additional information about the connections among themselves. Spectroscopic and spectrometric methods like NMR, IR, mass, etc., give the information about larger fragments or functional groups which usually consist of two or more carbon or

hetero atoms with hydrogens attached to them. The modern spectroscopic tehniques give also some suggestions about the connections between these atoms' groups within the structure.

Usually, the chemists use more than one of the mentioned methods and as the result they obtaine the sets of fragments of different sizes which are more or less accuratelly defined. Afterwards, the chemist seeks the structure in which these fragments are connected together to give the solution of the analysed sample.

Many fragments can overlap each other and many different constraints can complicate the conditions under which the fragments link together. The connection of fragments together on all possible ways is a combinatorial problem, which usually can not be efficiently solved by the humans. All possible combinations of fragments should be tried and the generated structures must be checked against all suggestions about the connections (constraints) derived from different analyses. Such complex procedure can be efficiently made only by a computer program called the *structure generator*[4].

A good structure generator must be:

- **exhaustive**: to generate *all* possible chemical structures consistent with a given set of fragments and various constraints,
- **irredundant**: to avoid the generation of identical structures,
- **effective**: to percieve the generation of chemically or topologically impossible, identical, or non-connected structures in advance or as soon as possible, and
- **fast**: to generate as many structures as possible per time unit.

Due to the fact that during the **fast** generation process the generation of identical structures cannot be completely avoided, the final elimination of duplicate structures must be added at the end of the process.

By now, a lot of structure generators and systems in which the generators are incorporated are known. Some of these systems are DENDRAL[5–8], CHEMICS[3,9,10], CASE[11–15], Robien's system[16], MOLGRAPH[2], ISOGEN[17], etc. They have incorporated the generators like CONGEN[6–8], GENOA[8], ASSEMBLE[11,12], COCOA[15], etc. They use different input fragments and constraints, generate different types of structures, and they are made for different computers, etc. Some of them are designed for special purposes (for generation of the macromolecules, large biological systems[18], for generation of only aliphatic[5] or only ring systems[19]) and are mainly implemented on large computer systems or specific mini/microcomputers[3,5–15] what prevents their much wider use. Some of them use only predefined fragments[3,9,10,20], the others use only one-atom fragments[17] or large fragments[6–8]. The fragments can overlap or not, can be with or without connected monovalent atoms, etc. They also differ regarding to the constraints they consider and the generation methods they use.

In order to achieve the wide applicability of our generator, called GEN, we were trying to consider as many 'good' characteristics of the mentioned generators as possible. Therefore:

- it can be used for the generation of arbitrary structures and also substructures,
- it can generate the structures from the input set of arbitrary fragments of any sizes with free bonds defined to a different extent of completeness,
- it can use many different input data which can be regarded as constraints,
- it is designed for personal computers, etc.

2 Generation with Generator GEN

The generation process consists of three parts:

1. the *preprocessing procedure* which prepares the input data for the generation performed by the generator GEN,
2. the *generation procedure* performed by the generator GEN, and
3. the *procedure for a display* of generated structures.

In the continuation the first two parts will be described in detail.

The *preprocessing procedure* is used for the preparation of any input data for the generation in the narrow sence. The input data may consist of any fragments with free bonds defined to a different extent of completeness, a molecular formula, molecular weight, and/or other constraints. On the basis of the available input data the preprocessing procedure determines **all** possible fragments, distributes them into one or more sets, and rewrites each set of fragments and all constraints in the standardized form acceptable for the generator GEN.

Due to different types of input data required by different generation systems, various preprocessing procedures are used. The *generator GEN* generates the structures from the sets that were previously determined by the preprocessing procedure. The generation process, performed by the generator GEN, is further divided into three steps which are implemented in turn:

1. determination of all possible connections between fragments – making the *connection matrix*,
2. generation of all possible structures – the *filling of the matrix* and considering the constraints, and
3. checking of the generated structures against constraints in order to eliminate the duplicates.

2.1 Preparation of Sets of Fragments

How different preprocessing procedures obtain the sets of fragments from different types of input data will be described in detail at the 'Applications'. At this point, the preprocessing of already determined set of fragments into various sets useful for the generator GEN will be explained.

Each set of fragments must consist of at least two or more different or identical fragments. The set, which is going to be used in the generation process, is called the *actual set*. The actual set is a sub-set of the *complete set* of possible fragments obtained by the preprocessing procedure.

The *fragment* is a part of a chemical structure and contains one or more non-hydrogen atoms with attached hydrogen atoms and *free bonds*. Each fragment has one or more free bonds available for forming connections to other fragments.

Free bonds can be of different types: single, double, or triple. They can or cannot be strictly defined. During the generation process the non-strictly defined free bonds can be merged into multiple bonds (for example, two single bonds into one double bond). The free bonds of the **same type**, which are on the **same atom** in the fragment, are called the **equivalent free bonds**. In some cases the free bonds, which are not on the same atom, are also equivalent, since they are in the same environment within the (symmetrical) fragment. In GEN, such free bonds are not treated as equivalent.

Before starting the generation of structures with the generator GEN, all fragments from the actual set must be given in a uniformed representation which we called a *connection table of fragments*[21]. Besides the standardization of the input format, the connection table of fragments has another very important role. It provides the necessary data for checking the final structures and all structural constraints. This is particularly important because some of the constraints may involve more fragments or even overlapping of some.

The connection table consists of as many rows as there are non-hydrogen atoms in **all** fragments from the actual set. Each row has the following information:

- identification number (ID) of the atom i,
- chemical symbol of the atom,
- identification numbers of neighboring atoms and the corresponding bond types (1, 2, or 3),
- free bonds and their bond types.

Free bonds are labeled as b_i. The maximal number of all free bonds is B_{all}. B_{all} is the sum of all single, double, and triple free bonds in all fragments from a given actual set:

$$B_{all} = B^s + B^d + B^t \tag{1}$$

where B^s, B^d, and B^t are total numbers of single, double, and triple free bonds in the actual set, respectively.

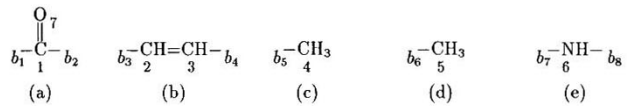
The link between **two** free bonds of the same type is called a *bond*. The result of the connection of two single free bonds is the *single bond*, while the result of the connection of two double or triple free bonds are the *double* or the *triple bonds*, respectively.

B_o is the **exact number** of bonds which are **necessary** to form the structure from a given set of fragments:

$$B_o = \frac{B^s}{2} + \frac{B^d}{2} + \frac{B^t}{2} \tag{2}$$

In the connection table of fragments, all atoms are arranged alphabetically (C, N, O, etc.) and within this sequence the atoms of the same type having larger number of free bonds have the priority. If, after following these rules, the ambiguity exists, the atoms are arranged arbitrarily.

The set of five fragments *a*, *b*, *c*, *d*, and *e* and the corresponding connection table is displayed in Figure 1. Fragments *c* and *d* are equal and have only one free bond. Each of the fragments *a*, *b*, and *e* are different and have two free bonds. Within the fragment *a* and the fragment *e* both free bonds are equivalent, while the two free bonds in the fragment *b* are not, because they are not on the same atom. The *ID* determine also the sequence of non-hydrogen atoms' data in the connection table of fragments.



Fragment	ID	Chemical Symbol	Neighbors Atom	Neighbors Bond	Free bonds Bond	Free bonds Type	Free bonds Bond	Free bonds Type
a	1	C	7	2	b_1	1	b_2	1
b	2	C	3	2	b_3	1		
b	3	C	2	2	b_4	1		
c	4	C			b_5	1		
d	5	C			b_6	1		
e	6	N			b_7	1	b_8	1
a	7	O	1	2				

Figure 1: The connection table of fragments *a*, *b*, *c*, *d*, and *e*. Up to six neighbors and free bonds can be attached to each atom

The set of fragments, rewritten in a form of the connection table of fragments, enters into the generator GEN. The generator should connect free bonds between themselves

to form the structure. After the generation process the labels b_i are exchanged with the identification numbers of atoms to which the particular bond is attached. Therefore, at the end of the process the connection table of fragments represents the topological connectivity of atoms in the generated structure.

From the connection table (Figure 1) some other valuable information about the free bonds can be retrieved (Table 1). These information are:

- the identification number of the atom on which the free bond b_i is attached and
- the sequential number k distinguishing the equivalent free bonds on the atom.

Table 1: Data for eight free bonds from the fragments shown in Figure 1. The sequential number k of equivalent free bonds goes from 1 to the maximal number of equivalent free bonds on the same atom

i	Label	ID[i]	k
1	b_1	1	1
2	b_2	1	2
3	b_3	2	1
4	b_4	3	1
5	b_5	4	1
6	b_6	5	1
7	b_7	6	1
8	b_8	6	2

To generate **all** possible structures from the given actual set the generator has to try **all possible connections** of free bonds. In order to achive the fast and efficient generation the *connection matrix of bonds* is used.

2.2 Connection Matrix of Bonds

The connection matrix of bonds E is a matrix of the free bonds. Each row and each column in E corresponds to exactly **one** free bond, so the size of the matrix E is $B_{all} \times B_{all}$.

The values of the elements e_{ij} of the matrices E are either T (true) or F (false). The element e_{ij} is T if the connection between the free bond b_i and the free bond b_j exist. Otherwise, e_{ij} is equal F. E is symmetrical across the main diagonal because the elements e_{ij} and e_{ji} represent the same connection.

Four types of the connection matrices E are distinguished:

- an *empty connection matrix* E_o: a matrix without connections (with all elements equal F),

- a *full connection matrix* E_f: a matrix with all possible connections,
- a *basic connection matrix* E_b: a matrix with all chemical meaningful connections, and
- a *connection matrix with a final structure* E_s: a matrix with exactly one T (true) element in each row and in each column.

The process of obtaining the basic connection matrix E_b starts with the full connection matrix E_f. E_b is obtained after the elimination of some meaningless and useless connections in the E_f. The elimination of any connection e_{ij} is made by turning its T value to F.

The elimination of the connections is related directly to the strategy of the generation of structure. Therefore, the detail description of the eliminations will be given in a special section "Elimination of the Useless Connections – Obtaining a Basic Connection Matrix" after the description of the generation strategy.

The basic connection matrix E_b for the set of fragments (Figure 1) is shown in Figure 2a. The connections e_{ij} in the E_b are marked with the consequtive numbers from 1 to 20 (Figure 2b). The eliminated connections are **not** marked.

The connections between the free bonds on different atoms of the same type are treated as identical. Each group of **identical connections** is labeled with a number called *identity parameter* EQ (Figure 2c). The parameter $EQ[e_{ij}]$ is calculated for all elements e_{ij} using the values $ID[i]$ and $ID[j]$ given in Table 1:

$$EQ(e_{ij}) = (ID[i] - 1)B_{all} + ID[j] - 1 \tag{3}$$

Once the EQ values are substituted into the connection matrix E_b (Figure 2c) it can be seen that some e_{ij} have identical values of EQ. For example, $e_{1,4}$ and $e_{2,4}$, $e_{5,7}$ and $e_{5,8}$.

When the basic connection matrix E_b is defined and all consequtive numbers of connections and parameters EQ are stored in special arrays, the generation can start.

The number of connections in the E_b has the greatest effect on the generation time. If there are many connections e_{ij} with T values in the E_b, a lot of time is needed for the generation. Therefore, the E_b must contain only the connections which are necessary for the generation of **all** structures.

2.3 Generation

The generation of structures is a combinatorial process of linking the fragments from a given set together in such a way that all possible structures are obtained. In order to generate only one structure from the set shown in Figure 1 four ($B_o = 4$) connections from E_b (Figure 2a) should be used. All sets of four connections among 20 possible ones

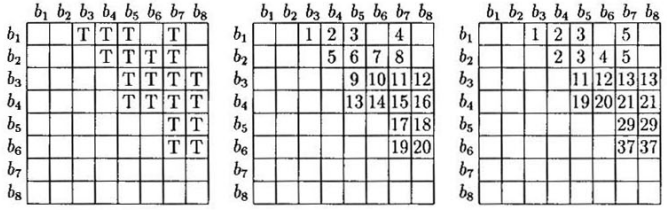


Figure 2: a) the basic connection matrix E_b obtained from the connection table of fragments shown in Figure 1, b) possible connections substituted with consequtive numbers, and c) possible connections grouped with respect to the dentity parameter EQ

are established during the process of generation. Therefore, to find all possible final structures 2295 combinations (combinations of the fourth order of 20 elements) should be tried.

The number of combinations increases very quickly with the number of free bonds, i.e., with the increasing B_o (equation 2) and the number of possible connections.

In order to avoid the combinatorial explosion a *controlled combinatorial process* is used. During the controlled combinatorial process, the empty connection matrix E_o is filled with the connections from the basic connection matrix E_b according to some rules. The rules in question are called the *rules for filling the matrix*. The goal of filling the empty matrix is to obtain the connection matrix with a structure E_s. When B_o connections are added into the matrix on such way that in each row and in each column exactly one T element is placed the connection matrix with a structure E_s is obtained. During the process of filling the connection matrix is neither empty (like E_o) nor is equal to the matrix with a structure E_s. Therefore, we should mark it as E.

Each new generated stucture has its own E_s, e.i., the structure is uniformly defined by the E_s.

The *rules for filling the matrix* E in order to obtain the E_s are:

1. each free bond should be used exactly once for forming the bond, i.e., in each row and column exactly one connection should exist,
2. the filling starts in the first (upper) row (the free bond b_1 should be the first free bond connected to any other free bond) and then continues to the lower rows,
3. the next connection to be added to the matrix must be in the first lower row with no connections,

4. each new added connection should have the **same** or **larger** EQ parameter compared with the previous connection. The same EQ is allowed only when merging of single bonds into multiple ones takes place.

By the consideration of the above rules, all possible rearrangement of connections should be tried to obtain all possible E_s, i.e., all possible structures. It is evident from the rules that each connection added into the matrix, reduces the posibility for placeing another connections. Therefore, the number of connections in the upper rows in E_b, which have more possible positions for placing the connections than the lower ones, have the greatest effect on the number of possible rearrangements of connections, i.e., on the number of possible E_s. This also effects on the generation time. This means, that from the upper rows of E_b as many connections as possible should be eliminated.

By adding a new connection into the matrix E two fragments are linked into a larger one. In order to continue the generation each new fragment **and** the remaining ones must satisfy the following three conditions:

1. each fragment must have at least one free bond, i.e., it should not be a complete structure,
2. the number of fragments with one free bond must be smaller than or equal to the number of still allowed connections, and
3. none of the atoms should have more free bonds than there are still allowed connections to complete the final structure.

If the fragments satisfy all above conditionds, the generation can continue. Otherwise, the last established connection must be erased and in the same row of the matrix the next one should be tried.

An example of filling an empty matrix E_o with the connections from E_b (Figure 2b) is shown in Figure 3. The filling starts in the upper (first) row of the matrix shown in Figure 2b and 3a. The first possible connection is the connection $e_{1,3}$ with the consequtive number 1. This connection is input into the matrix E_o. From two free bonds, b_1 and b_3, a new bond is formed. The 1-st and 3-rd row and column of E cannot accept any further connections, so the elements in these rows and columns are marked with '$\times$' (Figure 3b). The added connection is marked with a circle '$\bigcirc$'.

The next connection must be placed in the second row of E, i.e., the free bond b_2 should used in the next bond. The first useful connection in the second row is the connection 5 (Figure 3b). The generated fragment does not satisfy the condition 1 which prohibits the generation of fragments with no free bonds. The connections $e_{1,3}$ and $e_{2,4}$ with consequtive numbers 1 and 5, respectively, would make from fragments a and b (Figure 1) a small structure.

The next possible connection in the second row is the connection 6. It passes all tests and it is added into the matrix E. The elements in the 2-nd and 5-th rows and columns in E are marked as useless (Figure 3c).

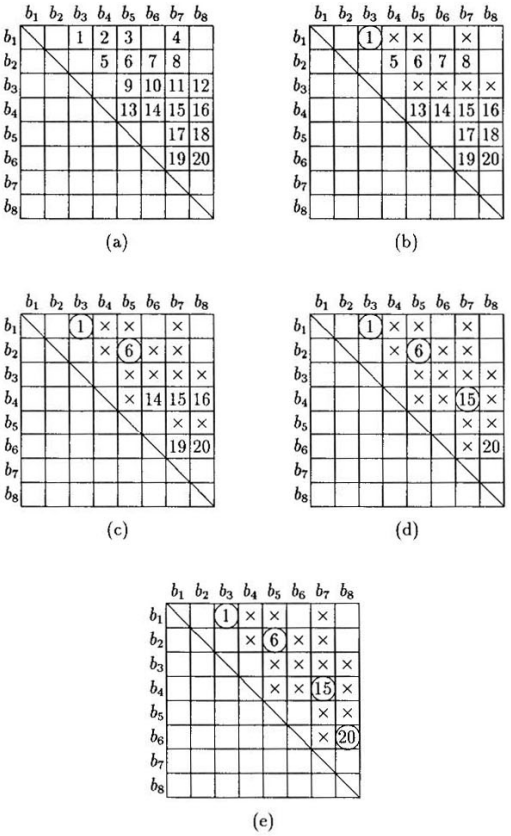


Figure 3: The example of filling the empty connection matrix E_o with the connections from E_b in order to obtain the matrix with a final structure E_s. The connections which should or should not be used for the generation of a final structure are marked with the consequtive numbers from the E_b or with '×', respectively. The connections with the consequtive numbers in the circles '◯' are added into the matrix E. Empty places (elements) represent useless connections eliminated during the process of obtaining the E_b

The procedure is repeated in the 4-th row, where the connection 14 would cause the generation of a small structure. It is eliminated and the next connection 15 is tried successfully (Figure 3d).

The last possible connection is 20 (Figure 3e). After adding this connection into the matrix the final structure, consisting of all fragments, is generated.

To examine **all possible combinations** of connections the procedure of finding the next structure starts with the already established E_s in the following actions:

1. all B_o connections from E_s are ordered as a list,
2. the last two connections from the list are erased,
3. the next connection in the row with the last but one erased connection is tried,
4. if at least one connection in a row was already examined, the next connection in this row must have larger EQ than the examined one,
5. after all connections in one row were tried, the procedure is finished if this row is the first one. Otherwise, the search for a new connection continues with action 2.

For example, the connections from Figure 3e with consequtive numbers 1, 6, 15, and 20 form the list of all B_o connections (action 1). The last two connections 15 and 20 from the list are erased (action 2) and the next connection following the connection 15 in the same row (4-th row) is tried (action 3). This is the connection 16, but it has the same parameter EQ as the connection 15 (action 4). Because all connections in the 4-th row are examined, the procedure is repeated with action 2: the connections 16 and 6 are erased and the next connection (connection 7) in the 2-nd row is tried, etc.

The final result of this generation are three different E_s which represent three different structures. All E_s, matrices with structures are shown in Figure 4.

The number of the output structures and the speed of the generation process depend on the number, type, and symmetry of the input fragments. In the case of completely different and non-symmetrical fragments, each new combination of connections (bonds) leads to a new structure. On the other hand, if the input set consists of many identical and/or highly symmetrical fragments, many identical structures should be generated. The generation of all duplicates should be avoided as early as possible.

In GEN, two actions are applied to prevent the generation of identical structures. First, some connections are erased during the determination of the basic connection matrix E_b. And second, during the generation process the parameter EQ is examined before the addition of each new connection into the matrix.

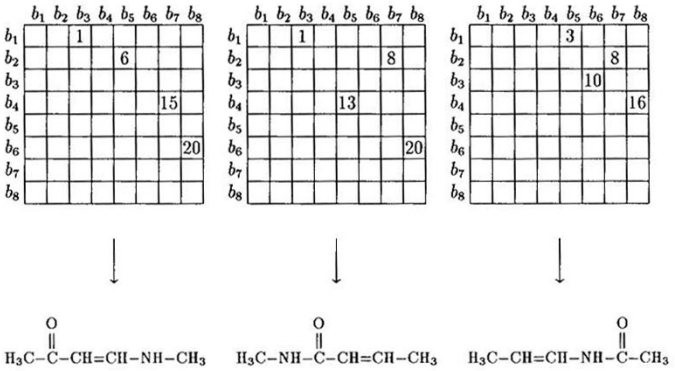


Figure 4: All possible structures generated from fragments shown in Figure 1

2.4 Checking of Output Structures

Due to the fact, that during the fast generation process the generation of identical structures cannot be **completely** avoided, the final elimination of duplicate structures is included at the end of the process. The elimination of duplicates is based on the substructure search[22]. The same substructure search routine is used also for checking of the output structures if they comply with the structural constraints (if there are any).

3 Elimination of the Useless Connections – Obtaining a Basic Connection Matrix

As mentioned before, the basic connection matrix E_b is obtained after the elimination of meaningless and useless connections from the full connection matrix E_f. The number of connections that remain in the E_b, specially those on upper rows, influences the generation time. Therefore as many connections as possible should be eliminated to establish the exhaustive and irredundant, but also a fast generation.

The elimination process is divided into two parts: the *general* eliminations performed on all fragments and the *specific* ones which are performed on the identical fragments only. The more equivalent bonds the fragments have and the more identical fragments

are in the set, the better are the effects of both types of eliminations (more connections are eliminated). The eliminations are tightly related to the generation process, i.e., to the rules for filling the matrix, and to the sequence order of atoms with free bonds in the connection matrices.

All types of eliminations are explained and illustrated in Figures 5, 6, 7, and 8. The empty places mark the useless and already eliminated connections (the elements e_{ij} with F values). The places marked with points '•' are elements e_{ij} which represent the possible and meaninful connections between the free bonds b_i and b_j. The places marked with the encircled points '⊙' or the circles '○' represent the connections which are going to be or have already been eliminated by some of the previously described eliminations, respectively. Parenthesis at the top of matrices mark the equivalent free bonds.

The general eliminations are explained on the actual set of fragments shown at the top of Figure 5.

The *general eliminations* involve seven cases:

1. All e_{ij}, where $i > j$. The connections e_{ij} and e_{ji} form the same bond (Figure 5a).
2. All e_{ij}, where $i = j$. Free bonds cannot be connected to themselves (Figure 5b).
3. The free bonds on the same initial fragment. For example, the connections $e_{1,2}$, $e_{1,3}$, $e_{2,3}$, and $e_{4,5}$ represent the connections between the free bonds on the same fragment (Figure 5c).
4. All e_{ij} connecting two fragments having only **one** free bond (Figure 5d). Such connections generate small structures.
5. Some identical connections e_{ij} leading to the generation of the identical intermediate fragments. Basically, these are the connections between the first and any other atom from the set providing the connections have the identical parameters EQ (Figure 5e).
6. All e_{ij} connecting the first equivalent free bond on the last atom having such bonds and all fragments having only one free bond. If not eliminated, such connections would lead to the generation of small structures (Figure 5f).
7. All connections e_{ij} between the last free bond in the set and the all but the last equivalent free bonds on any atom. The same is true also for the connections between the last but one free bond in the set and to the last or to the last but one free bond on any other atom from the set, etc (Figure 5g).

The first four eliminations are obvious, while the 4-th, 5-th, 6-th, and 7-th will be briefly discussed. The actual execution is shown in Figures 5d-5g.

b_1–CH< b_2 b_3 b_4–CH_2–b_5 b_6–CH_3 b_7–NH_2 b_8–OH

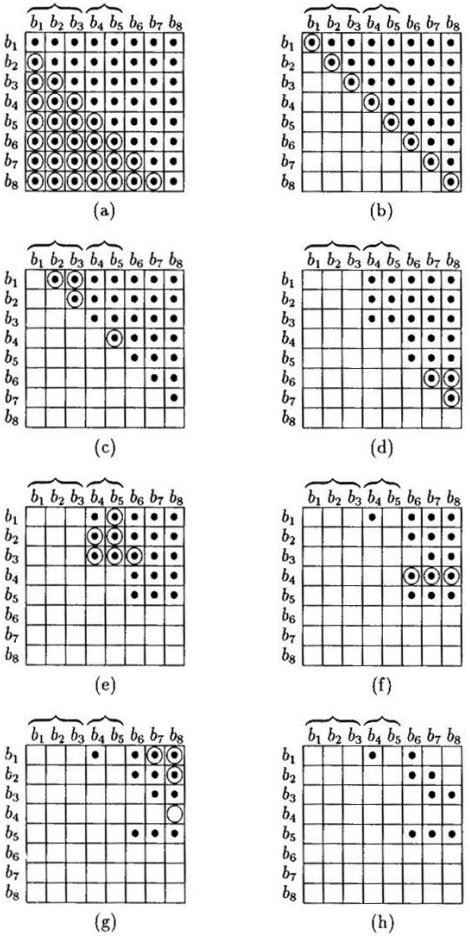


Figure 5: The general eliminations of connections

ad 4) The connections $e_{6,7}$, $e_{6,8}$, and $e_{7,8}$ are eliminated because they cause the generation of structures CH_3-NH_2, CH_3-OH, and NH_2-OH which do not contain all required fragments (Figure 5d).

ad 5) Very often some of the identical intermediate fragments lead to the generation of identical final structures. To prevent the generation of such duplicates these identical connections should be found and eliminated. According to the rules for filling the matrix, the generation path leading from the connections between the equivalent free bonds on the first atom and the equivalent free bonds on any other atom can be predicted: the **first** free bond b_1 must be **always** connected to the **first** free bond on **any** other atom. If exists, the **second** free bond b_2 on the **first** atom must be always connected to the **first** free bond on any other atom from the set. For example, in Figure 5e b_1, b_2, and b_3 are the free bonds on the first atom and b_4 and b_5 are the free bonds on the second atom. All connections $e_{1,4}$, $e_{2,4}$, $e_{3,4}$, $e_{1,5}$, $e_{2,5}$, and $e_{3,5}$ have the same EQ, therefore, all connections axcept the connection $e_{1,4}$ should be eliminated. For the same reason, the connection $e_{3,6}$ that links the free bond b_3 and the free bond b_6 on the third atom, should be eliminated too.

The number of identical connections that should be eliminated due to the above reason depends on the type of the output structures and on the size of the initial fragments. If the output structures can have multiple bonds or/and if the initial fragments consist of more than one non-hydrogen atom having free bonds lower number of identical connections should be eliminated. For example, if the structures with double bonds are allowed to be generated from the fragments shown in Figure 5, the connection $e_{2,5}$ must **not be** eliminated.

ad 6) In Figure 5f the connections $e_{4,6}$, $e_{4,7}$ and $e_{4,8}$ are eliminated, because, together with allowed connections $e_{5,6}$, $e_{5,7}$, and $e_{5,8}$, they lead to the generation of small structures like $HO-CH_2-CH_3$, $HO-CH_2-NH_2$, or $NH_2-CH_2-CH_3$.

ad 7) In Figure 5g the connections $e_{1,8}$ and $e_{2,8}$ are eliminated because b_8 is the last free bond in the set but b_1 and b_2 are not the last free bonds on the first atom. The connection $e_{1,7}$ is eliminated, because the free bond b_7 is the last but one free bond in the set and b_1 is the last but two equivalent free bond on the first atom, etc.

Besides the general eliminations the specific eliminations should be considered even more thoroughly. They are particularly successful on the sets having a considerable number of small and identical fragments, i.e., fragments consisting of only one carbon atom with attached hydrogens and equivalent free bonds. In the continuation, this type of eliminations will be exaplained on three different sets of such fragments:

1. set: >C<, −CH<, and five $-CH_3$ (Figure 6),
2. set: −CH<, three $-CH_2-$, and three $-CH_3$ (Figure 7),
3. set: five $-CH_2-$ and two $-CH_3$ (Figure 8).

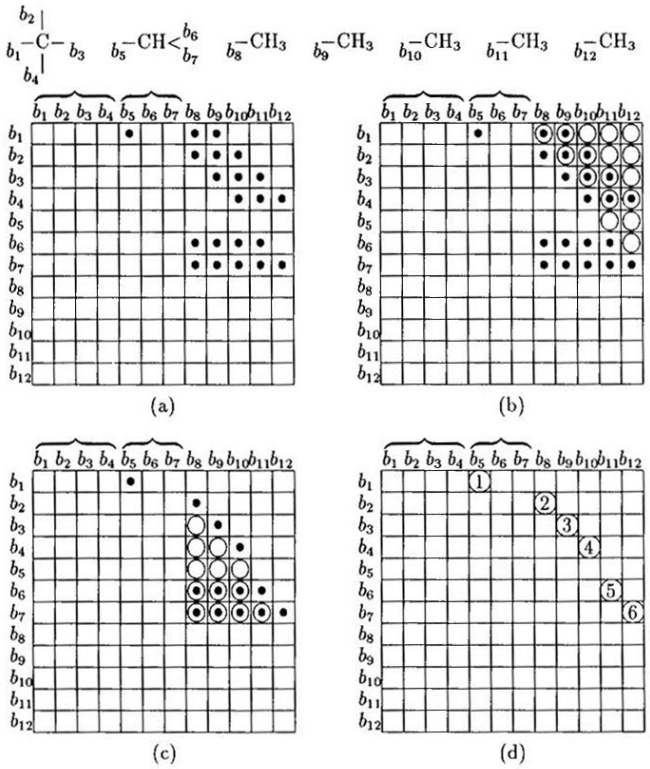


Figure 6: a) the general eliminations, b) and c) the specific eliminations of connections of $-CH_3$ with other fragments, and d) the basic connection matrix E_b

In Figures 6d, 7d, and 8d the basic connection matrices of each of these sets are shown. The possible connections are marked with their consequtive numbers. The numbers of connections which are performed in the final structures are in the circles '○'.

The *specific eliminations* are performed in order to prevent the generation of small and/or identical structures. These eliminations involve the following connections:

1. The connections e_{ij} linking **$-CH_3$ fragment** with any other fragment **too early** in the generation process. The use of such connections at the beginning of the generation leads to the formation of small structures. For example, if the connection $e_{1,8}$ (Figure 6) is used as the first conection in the generation, the only possible structure that should be generated is the small structure $C(CH_3)_4$.

   In order to prevent such situations, the connections e_{ij}, where b_j is the free bond on any $-CH_3$ fragment and b_i is the first equivalent free bond on the first fragment in the set, are eliminated. To the second equivalent free bond on the first fragment only the first $-CH_3$ fragment can be attached, to the third bond the first or the second $-CH_3$ fragment can be attached, etc. (Figure 6b).

2. The connections e_{ij} linking **$-CH_3$ fragment** with any other fragment **too late** in the generation process. The use of $-CH_3$ fragments too late in the generation can lead to small structures due to the excessive crosslinking of their mutual bonds, while leaving the $-CH_3$ fragments out of the generation process.

   The connections that must be eliminated to prevent such situations are connections e_{ij}, where b_j is the free bond on any except on the last $-CH_3$ fragment and b_i is the last equivalent free bond on the last fragment having such bonds (Figure 6c).

   In the same sense as described, the specific eliminations 1 and 2 are extended from the first towards the last, and from the last towards to the first fragment with equivalent bonds, respectively.

3. All e_{ij} which connect the **first** equivalent free bond b_i on **all $-CH_2-$ fragments** (except on the first $-CH_2-$ fragment in the set at all) with any free bond b_j on either $-CH_2-$ or $-CH_3$ fragments. Such connections would lead to the generation of small structures like $CH_3-(CH_2)_n-CH_3$ ($n = 1, 2, 3, \ldots$) and therefore, they are eliminated (Figure 7b).

4. Some connections e_{ij} which connect the bond b_i on the **first** fragment with any other bond b_j on the fragments which are **identical** among themselves (i.e. not necessariely identical to the first one). The use of too many such connections leads to the generation of the identical intermediate fragments and, therefore, also to the generation of the identical final structures. For example, the result of making any of the following three connections: $e_{1,4}$, $e_{1,6}$, or $e_{1,8}$ is the same fragment $>CH-CH_2-$. Therefore, the last two connections $e_{1,6}$ and $e_{1,8}$ are eliminated (Figure 7c).

5. Some connections e_{ij} between the free bonds on the **identical** fragments forming the same intermediate results. For example, the result of any two connections between the first three fragments in Figure 8 ($-CH_2-$) is the fragment $-CH_2-CH_2-CH_2-$. The same fragment is obtained by using any other possible connections among the actual five identical fragments. Therefore, the connections between the first and the fourth, the first and the fifth, the second and the fifth fragment are eliminated (Figure 8b).

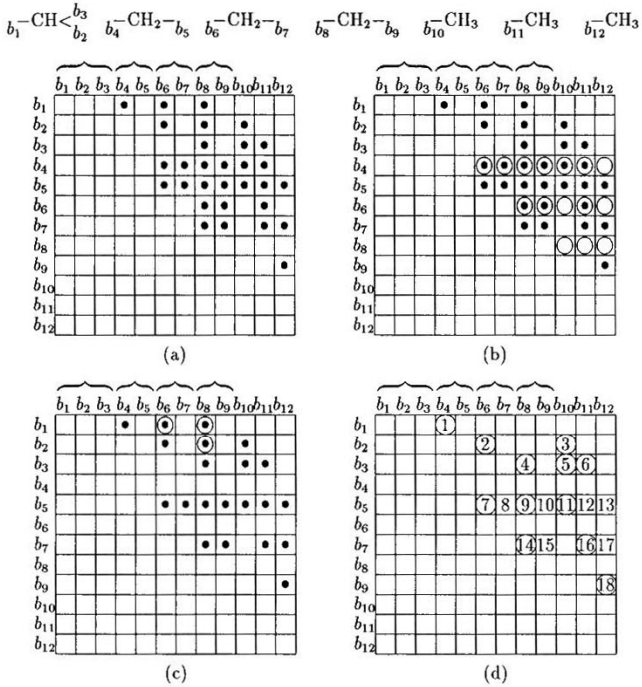


Figure 7: a) the general and the specific eliminations of connections of $-CH_3$ fragment, b) the specific eliminations of connections of $-CH_2-$ fragments, c) of the first and other carbon atoms identical to themselves, and d) the basic connection matrix E_b

6. Some connections e_{ij} between the free bonds on the **identical** fragments forming 3, 4, or more – member rings. For example, Figure 8c shows, that after linking the first with the second and the first with the third fragment (by making the connections $e_{1,3}$ and $e_{2,5}$, respectively), the use of any additional connection between the second and the third fragment leads to the generation of the 3-atom ring. The same is true also for the connections between the third and the fourth, or between the fourth and the fifth fragment. These later connections between the fragments should be eliminated to prevent the generation of such small rings.

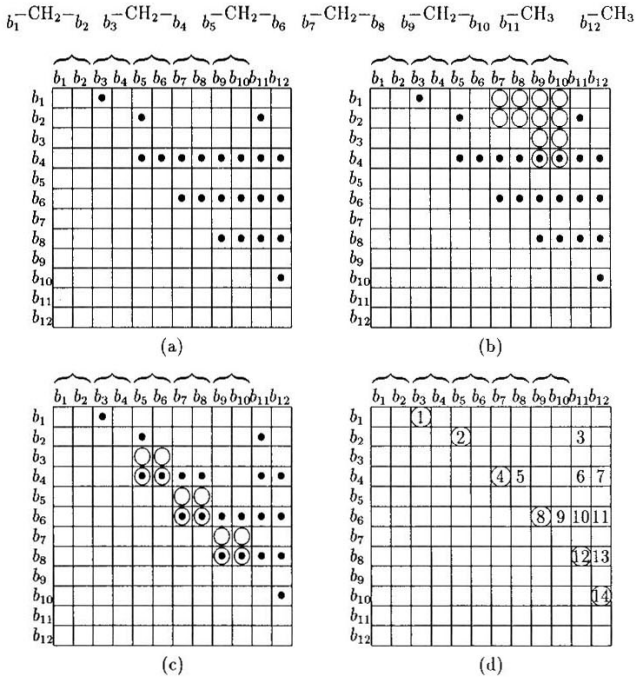


Figure 8: a) the general and the specific eliminations of connections of $-CH_3$, $-CH_2-$ fragment, and connections between the first atom and other identical atoms, b) and c) the specific eliminations of connections between the identical fragments, and d) the basic connection matrix E_b

After the examination of all connections and after the elimination of the useless ones the basic connection matrix E_b is obtained.

The measure of the performance of the eliminations can be given by the ratio between the number of connections in E_b that must be used in the final structures and the number of all connections in E_b. This ratio, given in percentage, varies with respect to the number of identical fragments, the number of equvalent free bonds, the presence of hetero atoms, the sizes of the fragments in the set, etc.

For three sets given in Figures 6, 7, and 8, respectively, the performance of the elimination procedure is given in Table 2. From the first, the second, and the third set one, three, and one structure can be generated, respectively. For all three sets the consequtive numbers of connections that link fragments into the final structures are in the circles '○' in Figures 6d, 7d, and 8d which represent the E_b.

Table 2: The performance of the elimination procedure

No. of the set	No. of all connec. in E_b	No. of used connections	performance [%]	No. of gener. structures	consequtive numbers of used connections
1	6	6	100	1	1 2 3 4 5 6
2	18	12	67	3	1 2 4 11 16 18 1 2 5 9 16 18 1 3 6 7 14 18
3	14	6	43	1	1 2 4 8 12 14

4 Constraints

The generator GEN can consider the following types of constraints:

- the molecular formula,
- the molecular weight,
- the number of fragments in the actual set,
- the number and type of free bonds, if the output structure is a substructure, and
- the structural constraints, i.e., the fragments which must be present or the fragments which are not allowed to apear in the output structures.

Before these constraints are input into the generator GEN, they must be rewritten into the standardized form of the connection table of fragments (see the section 'Preparation of Sets of Frgaments') acceptable for the generator.

All fragments, determined from different sources, which will be used for generation, constitute the complete set. From this set one or more actual sets are made due to the constraints. More constraints are given, less actual sets are made and also less structures are generated. On the contrary, if no constraints are given, a lot of actual sets are made and used for the generation of many final structures.

Molecular formula

The molecular formula of the final structure consists of the numbers of different atoms that must be in the final structure. For each atom the exact or the approximate number of atoms can be given. The approximate number is given by the interval. The exact molecular formula or the exact number of at least some atoms reduces the number of actual sets more than wide intervals of atoms.

Molecular weight

If the molecular weight is given as the constraint, the molecular weights of all fragments of each set are calculated and sumed up. Only the sets consisting of fragments, which sum of the molecular weights is identical to the defined molecular weight, will be used for the generation. The molecular weight is one of the most efficient constraint, i.e., it reduces the number of actual sets very much.

Number of fragments in the actual set

To reduce the number of possible actual sets the user should input the number of fragments that must be in one actual set. Otherwise, the actual sets can have from two up to all fragments from the complete set.

Number and types of free bonds of the final substructure

If the final structure should be really a substructure, the user must give the numbers and types of free bonds of the final substructure.

Structural constraints

The generator GEN distinguishes between structural fragments that **must be present** and those that **must be absent** in the resulting structure. Different *Mandatory Present Structural Fragments* (MPSF) and different *Mandatory Absent Structural Fragments* (MASF) are allowed to overlap, while the overlapping of identical MPSFs or MASFs is not allowed. Both types of structural fragments are used as constraints before and during the generation process.

The generator can consider up to **eight** different MPSFs and up to **eight** different MASFs. The number of identical structural constraints is not limited.

The structural constraints can contain one, two, or more non-hydrogen atoms with bonded hydrogens and free bonds.

For enhancing the description of structural constraints, two different labels are available for free bonds. Free bonds labeled 'A' can be connected to any atom **including** hydrogen, while free bonds labeled 'X' can be connected to any atom **except** hydrogen. During the generation the free bonds 'A' can be merged to form multiple bonds, while the free bonds 'X' cannot be. For example, all possible forms of the structural constraint $A-CH_2-A$ are four: $=CH_2$, $-CH_2-$, $-CH_3$, and CH_4, while the only possible form of the fragment $X-CH_2-X$ is $-CH_2-$.

Two free bonds 'X' on **one** fragment are not allowed to be linked together or to be linked to the **same** atom on another fragment. The free bonds can be only single.

The number of non-hydrogen atoms together with the types and definitions of the free bonds of the structural constraints have largest effect on the use of them in the generation process.

The MPSF or MASF with only **one** non-hydrogen atom can **always** be identified in the preprocessing step, i.e., in the initial fragments that form the actual set. If MPSF or MASF is identified in the actual set, this set is either accepted or rejected for the generation, respectively.

The MPSF with **two** or **more** non-hydrogen atoms and well defined free bonds causes the deletion of some free bonds in the connection table before the generation process, i.e., the reduction of the connection matrix. The MASF with **two** non-hydrogen atoms and well defined free bonds can (besides being in the actual set) reduce the connection matrix by prohibiting the connections that could produce the structure leading to MASF.

During the last step of the generation process in the final structure, the presence or the absence of MASFs with more than one non-hydrogen atom and constraints having free bonds of type 'A' (ring for example) is examined.

All constraints and the connections, that must be eliminated because of the MASFs, are rewritten into the connection table of fragments by the preprocessing procedure.

The MPSFs with well defined free bonds reduce the number of free bonds in the actual set. Therefore, the size of the connection matrix is reduced too. The MASFs cause the elimination of some connections in E_b. As the consequence of this reductions and eliminations, the generation time is also reduced.

5 Applications

As it was said before, the generator GEN is not a stand alone routine. For a full operation, it needs the preprocessing procedure, which prepares the input data for the generator, and the procedures for the display and for the manipulation with the generated structures. The input data can be fragments of any sizes with more or less defined free bond and/or any constraints described in the previous section. Due to the various input data, different preprocessing procedures are used.

At the moment, the generator GEN is included into two systems. In both of them the generator GEN is the central part, but the input fragments for the generation are determined and/or selected from different input data and the sets of fragments are chosen by different types of constraints. The first system GENSTR is implemented for the use of fragments, which have fixed and exactly defined free bonds (free bonds labeled with 'X'), and fragments having all hydrogen atoms bonded on carbon and/or hetero atoms. GENSTR system can employ the exact molecular formula or appropriate molecular formula with given interval of some atoms, the molecular weight, the number of fragments in the set, and/or the number and type of the free bonds of the output structures as constraints. The other system GENMAS system uses the fragments from the molecular formula as the main input data. The fragments retrieved from the molecular formula do not have strictly defined free bonds (they are labeled with 'A') and do not have hydrogen atoms located on carbon and hetero atoms. GENMAS system can consider the structural fragments, MPSF and MASF, as constraints.

5.1 GENSTR System

The main input data for the GENSTR system are fragments obtained from various spectroscopic and analytical experiments. These fragments consist of at least one non-hydrogen atom with defined free bonds and bond types and with hydrogens bonded to them.

In order to collect a set of fragments for the generation of all possible structures the system GENSTR offers various possibilities:

- selection the fragments from the table of standard fragments built into the system (Table 3),
- building arbitraray fragments with a built-in structure editor[21,22],
- selecting fragments which were obtained by the CARBON[23] interpretational system for ^{13}C NMR spectral data, and/or
- combination of any of the above three ways.

All fragments, selected by the user and input into the system, constitute the complete set. All fragments must have fixed and well-defined free bonds. This means, that although the free bonds can be single, double, or triple, only the bonds of the same type can be connected together during the generation process. No merging of single into multiple bonds is allowed.

The *actual set* is a set which is used for the generation. Each output structure must consists of **all** fragments from the actual set. In the resulting structures the fragments must not overlap. Using the constraints from the complete set one or more actual sets

Table 3: Standard fragments built in the GENSTR system

$-CH_3$	$=C<$	$-CHO$	$-N=$
$-CH_2-$	$=C=$	$-COO-$	$\equiv N$
$=CH_2$	$\equiv C-$	$-COOH$	$-Ph$
$=CH-$	$-OH$	$-CN$	$-F$
$>CH-$	$-O-$	$-NH_2$	$-Cl$
$\equiv CH$	$=O$	$-NH-$	$-Br$
$>C<$	$-CO-$	$-N<$	$-J$

are selected. The more fragments are in the complete set and the less constraints are determined, the more actual sets can be selected.

As the constraints, which can reduce the number of possible actual sets and the number of resulting structures, the following data can be used:

- molecular weight of the final structures,
- exact or approximate molecular formula with a given interval of atoms of the final structures,
- the number of fragments in the actual set, and/or
- the number and the type of free bonds in the final structures if they are supposed to be radicals.

The actual sets are selected automatically from the complete set on the basis of the constraints. If there are many constraints, only a few actual sets can be produced.

In Figure 9 the automatic selection of the actual set from the complete set of 13 fragments on the basis of molecular weight and the interval of numbers of C, N, and O atoms is shown.

Before the generation starts, the selected actual sets are displayed to the user for confirmation or rejection. The manual confirmation or rejection is enabled because some structural constraints can not be automatically detected and considered. Therefore, the user is always advised to check the selected set very carefully.

The fragments from each actual set that passes the user's examination are written in the form of a connection table. The connection table of each actual set is entered into the generator GEN separately and the generation process can start.

The first step of the generation process is the elimination of idle connections to obtain the E_b. The first elimination, that should be performed on the set with exactly defined free bonds, is the elimination of connections which connect the free bonds of different

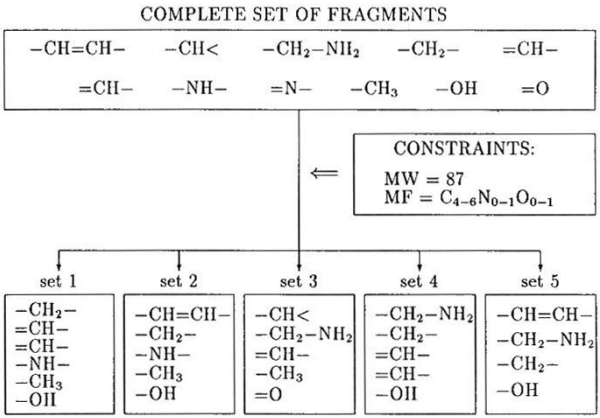


Figure 9: The selection of actual sets from the complete set according to the given molecular weight and the interval of the possible numbers of carbon, nitrogen, and oxygen atoms of the final structures

types. In Figure 10 the eliminations of connections between single and double free bonds of the first actual set are shown. The similar eliminations can be made also with e_{ij}-s describing the connections between single and triple or double and triple free bonds. After these usually very efficient eliminations the other general and specific can follow to obtain the basic connection matrix E_b.

The next two steps are the filling of the matrix and the examination of the output structures. The result is a set of different structures that are generated from **all** fragments of **one** actual set.

The same generation procedure is then performed sequentially on all actual sets.

In the GENSTR system it is possible that two or more **identical** structures emerge from different actual sets. To remove duplicates, the eliminating routine is used once again after all structures generated from all actual sets have been obtained.

The result of the generation from the actual sets given in Figure 10 are structures shown in Figure 11a. From the first and the second as well as from the fourth and the fifth sets identical structures are generated (Figure 11b). The generation of identical structures was caused because the fragment −CH=CH− in the second and in the fifth actual set consists of two identical fragments, −CH=, from the first and from the fourth set. The eliminating routine detects the duplicates and drops one of them from the final list.

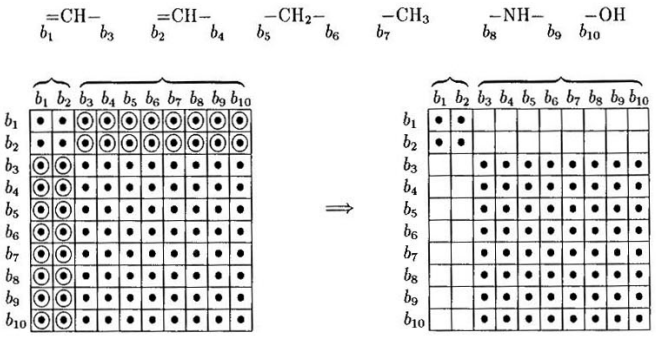
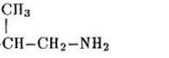


Figure 10: The elimination of connections which connect the single and double free bonds. The single and double free bonds are grouped with parenthesis

1) HO−NH−CH_2−CH=CH−CH_3
2) HO−CH=CH−CH_2−NH−CH_3
3) HO−NH−CH=CH−CH_2−CH_3
4) HO−CH_2−CH=CH−NH−CH_3
5) HO−CH=CH−NH−CH_2−CH_3
6) HO−CH_2−NH−CH=CH−CH_3



8) HO−CH=CH−CH_2−CH_2−NH_2
9) HO−CH_2−CH=CH−CH_2−NH_2

(a)

SET	STRUCTURES
1	1 2 3 4 5 6
2	1 2 3 4 5 6
3	7
4	8 9
5	8 9

(b)

Figure 11: (a) The generated structures and (b) a table of sets and belonging structures

5.2 GENMAS System

The GENMAS system is designed to use only the molecular formula as the input. To prevent the combinatorial explosion the structural constraints should be given as well, specially, for molecular formulas with more than 10 non-hydrogen atoms when more structures will be produced than the user can handle efficiently.

The molecular formula can consist of up to 30 non-hydrogen atoms like C, N, O, S, halogenes, and the generic fragment 'other'. The term 'other' refers to any atom or atom's fragment that can be determined and entered by the user. The user determines the generic fragment 'other' with up to two characters long symbol and a number of free bonds. If there are many free bonds, they must be equivalent. For examples, some fragments that can be treated as 'other'-s are: $-CH_2-CH_2-O-CH_3$ (one free bond), $>CH-NH-CH_3$, $>N-CH_2-Ph$, 3,4-disubstituted pyrrole ring (two equivalent free bonds), or 1,3,5-substituted benzene (three equivalent free bonds).

The generic fragment 'other' is treated like any other hetero atom.

The GENMAS system uses the structural fragments (see section 'Constraints'), MPSF and MASF, as constraints. These fragments should have the free bonds defined to a different extent of completeness using signs 'A' or 'X'.

Table 4: Table of standard constraints that are built-in the GENMAS system

$-CH_3$	$-CH=CH_2$	$-O-CH_2-$	$-NH-$	C7-ring	6-ring
$-CH_2-$	$-C\equiv C-$	$-CO-$	$-CH_2-NH-$	C6-ring	5-ring
$>CH-$	$-C\equiv CH$	$-CHO$	$-N<$	C5-ring	4-ring
$>C<$	$-OH$	$-COO-$	$-CN$	C4-ring	3-ring
$-CH_2-CH_2-$	$-CH_2-OH$	$-COOH$	Ph-ring	C3-ring	=
$-CH_2-CH_3$	$-O-$	$-NH_2$	benzene	8-ring	$\equiv$
$-CH=CH-$	$-O-CH_3$	$-CH_2-NH_2$	C8-ring	7-ring	

MPSF and MASF structural constraints can be input by the structure editor[21,22]. Besides the standard structure editing commands the structure editor in GENMAS has an additional command, STANDC, for selecting the built-in fragments (Table 4). The selected built-in fragments can be further on changed by the structure editor commands.

As said in the section 'Constraints', free bonds 'X' must be the single bonds only. The exception are last eight standard constraints (Table 4), where label 'X' represents arbitrary non-hydrogen atoms in rings or double or triple free bond.

After the molecular formula and the structural constraints are entered, the complete set of fragment is formed. The complete set consists of all atoms (including hydrogens) from molecular formula. The free bonds on atoms are labeled with 'A'. The number of free

bonds on the atom corresponds to the valence number of the atom. For example, the complete set for the molecular formula $C_4H_5NO_2$ consists of four >C<, five −H, one −N<, and two −O− fragments with free bonds labeled 'A'.

The next action of the preprocessing procedure is the substitution of hydrogens on the free bonds of carbon and hetero atoms. The hydrogens are distributed in **all** possible ways hence producing **all** possible sets. Using MPSFs and MASFs the number of all possible sets is reduced. Without any MPSFs or MASFs the number of possible sets (which is usually very large) is equal to the number of actual sets. Table 5 shows how five hydrogens can be placed on four carbons, one nitrogen, and two oxygens in order to obtain 37 possible sets. Among the possible sets, the selection of actual sets for the generation is made by the preprocessing procedure.

For example, for the pruning of 37 possible sets, two MPSFs (−CO− and $-CH_2CH_2-$) and one MASP ($-CH_3$) were used (Table 5). The sets number 6, 19, 20, 24, and 25 include all atoms for forming both MPSFs and exclude the MASF. Therefore, the fragments from each set are written in a form of connection table with some connected free bonds in order to form the MPSFs with well defined free bonds. The presence of 5-ring MPSF will be considered during the generation process by the eliminating routine when the structures will be generated.

The example of the connection table of fragments from set 6 before and after the connections of some free bonds are given in Table 6a and b, respectively. The −CO− fragment is formed by connecting free bonds b_5, b_{14}, and b_6, b_{15} as shown in Table 6b. The labels of connected free bonds (b_5, b_{14} and b_6, b_{15}) are substitute by the ID numbers of atoms which are attached to the bonds (6, 2 and 6, 2). The double bond between 2C and 6O is generated. If there is more than one equivalent free bond on the atom, always the free bond with the smallest sequential number k is used to form the MPSF. If there are more identical fragments in the set (like 1C and 2C), the last is used. At the present example, 2C is used and its free bonds are b_5 and b_6. On the same way, the second MPSF ($-CH_2-CH_2-$) is formed. After the consideration of all MPSFs, the remaining free bonds are renumbered. As a consequence of the formation of MPSFs, the number of free bonds is reduced from 16 to 10, i.e., the size of E_b is reduced from 16×16 to 10×10.

The connection tables of all selected actual sets (6, 19, 20, 24, 25) are input into the generation process sequentially. If the generation of cyclic structures is allowed, the user is asked about the number of rings that can constitute the output structures. The number of rings is defined by the number of all non-hydrogen atoms (ID^{all}) in the set and the number of all bonds B, that should be formed to obtain the output structure. B is given by the equation 4:

$$B = B_{all} + B_c \tag{4}$$

where B_c is the number of bonds that were made in the connection table to form all MPSFs.

Table 5: All actual sets from the molecular formula $C_4H_5NO_2$. The selection is made by four structural constraints: $-CO-$, $-CH_2CH_2-$, and 5-ring as MPSF and $-CH_3$ as MASF

SET	C C C C N O O	$-CO-^P$	$-CH_2CH_2-^P$	$-CH_3{}^A$	5-ring^P	No.Str.
1	0 0 0 3 2 1 1					
2	0 0 1 2 2 1 1		√			
3	0 0 1 3 1 1 1					
4	0 0 1 3 2 0 1	√				
5	0 0 2 2 1 1 1		√	√		
6	0 0 2 2 2 0 1	√	√	√		
7	0 0 2 3 0 1 1					
8	0 0 2 3 1 0 1	√				
9	0 0 2 3 2 0 0	√				
10	0 0 3 3 0 0 1	√				
11	0 0 3 3 1 0 0	√				
12	0 1 1 1 2 1 1			√		
13	0 1 1 2 1 1 1			√		
14	0 1 1 2 2 0 1	√	√			
15	0 1 1 3 0 1 1					
16	0 1 1 3 1 0 1	√				
17	0 1 1 3 2 0 0	√				
18	0 1 2 2 0 1 1		√	√		
19	0 1 2 2 1 0 1	√	√	√	√	3
20	0 1 2 2 2 0 0	√	√	√	√	3
21	0 1 2 3 0 0 1	√				
22	0 1 2 3 1 0 0	√				
23	0 1 3 3 0 0 0	√				
24	0 2 2 2 0 0 1	√	√	√	√	2
25	0 2 2 2 1 0 0	√	√	√		
26	0 2 2 3 0 0 0	√	√			
27	1 1 1 1 1 1 1			√		
28	1 1 1 1 2 0 1			√		
29	1 1 1 2 0 1 1			√		
30	1 1 1 2 1 0 1			√		
31	1 1 1 2 2 0 0			√		
32	1 1 1 3 0 0 1					
33	1 1 1 3 1 0 0					
34	1 1 2 2 0 0 1		√	√		
35	1 1 2 2 1 0 0		√	√		
36	1 1 2 3 0 0 0					
37	1 2 2 2 0 0 0		√	√		

The number of rings or double bonds R, that can constitute the output structures, are calculated from the equation 5:

$$R = B - ID^{all} + 1 \tag{5}$$

Table 6: The connection table of the fragments from the sixth set (a) before the consideration of the MPSFs and (b) after the consideration of $-CO-$ and $-CH_2-CH_2-$ and the renumbering of remaining free bonds

1C	b_1	1	b_2	1	b_3	1	b_4	1	
2C	b_5	1	b_6	1	b_7	1	b_8	1	
3C	b_9	1	b_{10}	1					
4C	b_{11}	1	b_{12}	1					
5N	b_{13}	1							
6O	b_{14}	1	b_{15}	1					
7O	b_{16}	1							

(a)

1C	b_1	1	b_2	1	b_3	1	b_4	1	
2C	6	2	b_5	1	b_6	1			
3C	4	1	b_7	1					
4C	3	1	b_8	1					
5N	b_9	1							
6O	2	2							
7O	b_{10}	1							

(b)

For example, if R is less then zero, more than one structure should be generated from the present actual set. If R is zero, one structure with no rings or double bonds should be generated. If R is one, one structure with one ring or one double bond should be generated. If R is two, one structure with one triple bond or two double bonds or two rings or one ring and one double bond should be generated.

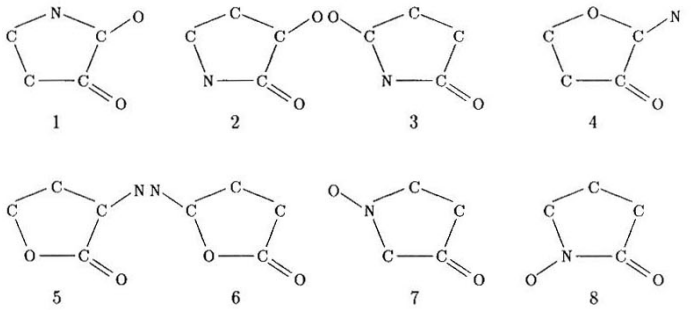


Figure 12: All structures generated on the basis of the molecular formula $C_4H_5NO_2$ and four structural constraints X–CO–X, X–CH_2CH_2–X, and 5-ring as MPSF and X–CH_3 as MASF. The hydrogen atoms are omitted from the display

From sets 19, 20, and 24 three (1, 2, 3), three (4, 5, 6), and two (7, 8) structures, which contain also the 5-ring fragment, are generated, respectively (Figure 12). The structures

generated form sets 6 and 25 include the 4-ring and 6-ring, respectively. They do not satisfy to the 5-ring constraint.

The structures generated from different sets are always different because they consist of different initial fragments (central atom with linked hydrogens and labeled free bonds) that constitute each set. Hence, no additional elimination of duplicates among structures generated from different sets are necessary. The elimination of duplicates should be performed only on structures generated from the same set.

As can be seen from Figure 12, the merging of single bonds to double or triple bonds is allowed, because all fragments that form the actual sets have free bonds of type 'A'. The only exception are free bonds of formed MPSFs that are labeled 'X'. For example, free bonds b_5 and b_6 on Table 6b, that belong to $-CO-$ defined as MPSF, are of type 'X' and, therefore, they cannot be merged.

6 Discussion

To find out the performance and the computational efficiency of the generator GEN the comparisons with other generators have been made. The comparison of the number of isomer structures generated by the systems DENDRAL[1,2], MOLGRAPH[2], CHEMICS[2,3], and GENMAS, which includes the generator GEN, has been made for the same molecular formula (Table 7). All systems except CHEMICS generate an equal number of structures from given molecular formula.

Table 7: Number of isomer structures obtained with different systems

Molecular Formula	DENDRAL	MOLGRAPH	CHEMICS	GENMAS
$C_2H_5NO_2$	84	84	87	84
$C_4H_7NO_2$	764	764	802	764
C_3H_4BrCl	—	10	8	10
C_5H_8BrCl	—	140	108	140
$C_6H_{10}O$	747	747	745	747

The comparison of the generation times between the CHEMICS and the GENMAS system, using identical molecular formula and structural constraints, shows the efficiency of the GENMAS system (Table 8). All structural constraints have free bonds labeled 'X'. The computational were made on MV/2000DC and on various personal computers for CHEMICS and GENMAS, respectively.

The comparison of the efficiency between the later version of generator GEN[24] and the new one with many additional eliminations of useless connections (see Chapter 'Elimination of the Useless Connections'). The new version of generator GEN needs less time for generation structures without given constraints compared with the later version. The

Table 8: The comparison of the generation times between CHEMICS and GENMAS for the generation of various structures. All free bonds of structural constraints are treated as 'X'. '△' is a sign for 1,2-disubstituted 3C-ring

Molecular Formula	MPSF$^{(P)}$ or MASF$^{(A)}$	No. of Structures		CPU time (s)			
				CHEMICS	GENMAS		
		CHEMICS	GENMAS	MV/2000DC	AT[a]	286[b]	386[c]
$C_4H_5O_2Cl$	—	907	907	840	2436	2065	281
C_4H_7NO	—	802	764	1436	402	297	56
$C_4H_5O_2Cl$	$-CO-CH_3{}^{(P)}$ $-CO-Cl^{(P)}$	1	1	6	4	3	2
C_4H_7NO	$-NH-CO-{}^{(P)}$ $-CO-NH_2{}^{(A)}$	5	5	34	15	14	5
C_3H_5NO	—	87	84	52	13	8	2
C_3H_5NO	$-NH_2{}^{(P)}$ △ $^{(A)}$	25	25	44	12	9	4

[a]CI = 6.9; [b]CI = 13.7; [c]CI = 48.5

reason for the decrease of the generation time is in a more efficient procedure for the elimination of connections, the general and the specific ones. Therefore, less duplicates are generated. Unfortunatelly, the new GEN needs more time for generation of structures limited by the structural constraints. In the new version, the additional time (from 2 to 5 second) is needed for making more comparisons in order to eliminate as many connections as possible. If only a few structures are generated, more time is needed for the eliminations than for the generation and the decrease in time is evident. Otherwise, if many structures are generated, the time increase used for the eliminations can be neglected compared to the time gained in a more efficient generation.

The generator GEN does not distinguish resonant or conjugated bonds from the localized single and double bonds that alternate. For example, 1,2-dichlorobenzene (Figure 13a), which have the resonant bonds, can be written in two canonic forms (Figure 13b and c). The generator considers them as different. As the consequence, it generates more structures because all canonic forms are regarded to be different.

In this case, the problem with resonant bonds can easiely be avoided by introducing the concept of the generic fragment 'other' instead of the entire benzene ring with two free bonds. For example, the benzene ring can be labeled as 'R6' (ring 6) with valence number 2. The molecular formula becomes Cl_2R6. Due to the fact, that generic fragments 'other'-s are regarded as any other predefined atom, the MPSF $-R6-$ should be given as well. The resulting structure is only one structure written as $Cl-R6-Cl$. By replacing 'R6' with benzene ring the structure *a* from Figure 13 is obtained.

The introduction of the generic fragment 'other' causes also the decrease of the number of atoms in the molecular formula and this effects directly on the generation time and on the number of output structures.

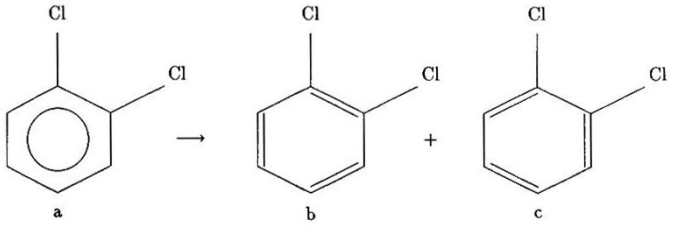


Figure 13: a) 1,2–dichlorobenzene with resonant bonds, b,c) two canonic forms of 1,2–dichlorobenzene with localized single and double bonds

Because the fragment 'other' can be introduced only for atoms' fragment with identical free bonds, very often the problem of resonant bonds cannot be as efficiently and easiely avoided as shown above.

This problem will be solved in one of the next version of the generator by the introduction of a new label for the resonant bonds in the connection table of fragments. This can be made easily always when the fragments have well defined free bonds 'X' (as in GENSTR, for example) and by changing the eliminating routine so that it can find places where the delocalizations can arise. This improved routine should be used for the elimination of the duplicate structures generated from the molecular formula where the free bonds labeled 'A' are not exactly defined.

7 Conclusion

The generator GEN was developed for generating structures from various fragments obtained from different types of spectroscopies. Up to 30 fragments of any sizes can be used. Only the atoms of fragments having free bonds participate in the generation process and effect on the number of generated structures and on the generation time.

The sizes of the input fragments should be reduced by using the generic fragment 'other'. The introduction of the generic fragments 'other'-s also causes a decrease of the time needed for the generation.

The generator GEN enables the generation of structures and substructures (radicals) which can be later used as the input fragments for the next generation. Therefore, the analitical problems involving large structures can be divided into more smaller problems (finding parts of the large structures) which could be solved (identified and/or determined) separately. The union of parts should be done at the end of the analitical process by the generator GEN or even manualy if there are only few of them.

The connection matrices enables the consideration of the majority of structural constraints, even before the actual generation starts. It enables also the elimination of many connections that lead to the generation of duplicates or small structures which consist of only some fragments from the set. As the consequence, the generation time is reduced.

The structural constraints can overlap or not. The free bonds of the structural constraints can be defined to a different extent of completeness ('A' or 'X'). All these possibilities simplify the determination of the output structures. They should be determined and described as precise as possible. In this way only a few structures should be generated. Therefore, the user's decision, which structure is the looked-after one, is much easier.

The consideration of various constraints like exact or approximate molecular formula, molecular weight, number of fragments in the set, number and type of free bonds, if the substructures should be generated, is enabled by different preliminary procedures that prepare the input data for unique generator GEN.

The generator GEN cannot be used only for generation of organic chemical structures, but also for generation or anorganic structures, or general graphs where nodes and branches are atoms and bonds, respectively, ...

The generator GEN can play very important rule as a part of a large structure elucidation system[25] based on various spectroscopic data (like KIBK(R) CARBON SOFTWARE[23] with the GENSTR system based on ^{13}C NMR spectroscopic data). The generator links fragments predicted from different parts of such system into the final structures. Among them, the most probable structures can be selected according to the spectral data obtained by experiments or using different simulators like SIMULA[26], VODIK, etc.

The generator GEN is useful also in the mass spectrometry. It can predict the structures from peaks in mass spectrum where each peak determines a set of fragments with defined mass. Usually more molecular formulas can be predicted from certain mass. The generator (in the combination with preliminary procedure similar to this in the GENMAS system) generates fragments from these molecular formulas and links them together obtaining the possible structures belonging to or producing the query spectra.

8 Technical Characteristics

The generator GEN is designed for personal computers. It is easily transportable. The use of it is simple and very easy to learn. It is designed for everyone not very familiar with the use of computers. The generator has a content–dependent HELP option. All operations are available over menus. The fragments, structural constraints, and output structures are displayed as two-dimensional graphs on the screen.

Each of described two systems, GENSTR and GENMAS, needs about 0.5 MByte of disk space and 640 KBytes of memory. All generated structures are stored on disk and for

about 1000 structures the generator needs additional 1 MByte of space. For saving more structures, correspondingly more space should be available.

The generator (and also the GENMAS system and all systems from KIBK(R) CARBON SOFTWARE) is written in Turbo Pascal 5.5 programming language. It works on IBM PC XT/AT/286/386 personal computers and compatibles with EGA, VGA, or Hercules graphic card. For drawing chemical structures and fragments the TURBO Pascal standard graphic routines are used.

Acknowledgement

The authors acknowledge the financial support from the former Research Community of Slovenia and from the Ministry of Science and Technology of Slovenia.

References

1. Lederberg, J.; Sutherland, G. L.; Buchanan, B. G.; Feigenbaum, E. A.; Robertson, A. V.; Duffield, A. M.; Djerassi, C. Application of Artificial Intelligence for Chemical Inference. I. The Number of Possible Organic Compounds. Acyclic Structures Containing C, H, O, and N. *J. Am. Chem. Soc.* **1969**, *91*, 2973-2976.

2. Kerber, A.; Lane, R.; Moser, D. Ein Strukturgenerator für molekulare Graphen. *Anal. Chim. Acta* **1990**, *235*, 2973-2976.

3. Funatsu, K.; Miyabayaski, N.; Sasaki, S. Further Development of Structure Generation in the Automated Structure Elucidation System CHEMICS. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 18-28.

4. Gray, N.A.B. *Computer-Assisted Structure Elucidation*, John Wiley & Sons, New York, 1986.

5. Gray, N.A.B. Dendral and Meta-Dendral - The Myth and the Reality. *Chem. Intell. Lab. Sys.* **1988**, *5*, 11-32.

6. Masinter, L. M.; Sridharan, N. S.; Lederberg, J.; Smith, D. H. Applications of Artificial Intelligence for Chemical Inference. XII. Exhaustive Generation of Cyclic and Acyclic Isomers. *J. Am. Chem. Soc.* **1974**, *96*, 7702-7714.

7. Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. Application of Artificial Intelligence for Chemical Inference. XVII. An Aproach to Computer-Assisted Elucidation of Molecular Structure. *J. Am. Chem. Soc.* **1975**, *97*, 5755-5762.

8. Carhart, R. E.; Smith, D. H.; Gray, N. A.; Nourse, J. G.; Djerassi,C. GENOA: A Computer Program for Structure Elucidation Utilizing Overlapping and Alternative Substructures. *J. Org. Chem.* **1981**, *46*, 1708-1718.

9. Oshima, T.; Ishida, Y.; Saito, K.; Sasaki, S. Chemics-UBE, A Modified System of Chemics. *Anal. Chim. Acta* **1980**, *122*, 95-102.

10. Abe, H.; Okuyama, T.; Fujiwara, I.; Sasaki, S. A Computer Program for Generation of Constitutionally Isomeric Structural Formulas. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 220-229.

11. Shelley, C. A.; Hays, T. R.; Munk, M. E.; Roman, R. V. An Approach to Automated Partial Structure Expansion. *Anal. Chim. Acta* **1978**, *103*, 121-132.

12. Shelley, C. A.; Munk, M. E. CASE, a Computer Model of the Structure Elucidation Process. *Anal. Chim. Acta* **1981**, *133*, 507-516.

13. Munk, M. E.; Farkas, M.; Lipkus, A. H.; Christie, B. D. Computer-Assisted Chemical Structure Analysis. *Mikrochim. Acta* **1986**, *II*, 199-215.

14. Christie, B. D.; Munk, M. E. The Application of Two-Dimensional Nuclear Magnetic Resonance Spectroscopy in Computer-Assisted Structure Elucidation. *Anal. Chim. Acta* **1987**, *200*, 347-361.

15. Christie, B. D.; Munk, M. E. Structure Generation by Reduction: A New Strategy for Computer-Assisted Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 87-93.

16. Robien, W. Computer-Assisted Structure Elucidation of Organic Compounds III: Automatic Fragment Generation from ^{13}C NMR Spectra. *Mikrochim. Acta* **1986**, *II*, 271-279.

17. Zhu, S.; Zhang, J. Exhaustive Generation of Structural Isomers for a Given Empirical Formula - A New Algorithm. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 34-38.

18. Nilakantan, R.; Bauman, N.; Venkataraghavan, R. A Method for Automatic Generation of Novel Chemical Structures and Its Potential Applications to Drug Discovery. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 527-530.

19. Carhart, R. E.; Smith, D. H.; Brown, H.; Sridharan, N. S. Applications of Artificial Intelligence for Chemical Inference. XVI. Computer Generation of Vertex-Graphs and Ring Systems. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 124-131.

20. Razinger, M.; Zupan, J.; Novič, M. Computer generation of chemical structures from known fragments. *Mikrochim. Acta* **1986**, *II*, 411-421.

21. Zupan, J. *Algorithms for Chemists*, John Wiley & Sons, Chichester, 1989.

22. Zupan, J.; Bohanec, S. Creation and Use of Chemical data Bases with Substructure Search Capability. *Vestn. Slov. Kem. Drus.* **1987**, *34(1)*, 71-81.

23. Zupan, J.; Novič, M.; Bohanec, S.; Razinger, M.; Lah, L.; Tušar, M.; Košir, I. Expert System for Solving Problems in Carbon-13 Nuclear Magnetic Resonance Spectroscopy. *Anal. Chim. Acta* **1987**, *200*, 333-345.

24. Bohanec, S.; Zupan, J. Structure Generation of Constitutional Isomers from Structural Fragments. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 531-540.

25. Bohanec, S.; Tušar, M.; Tušar, L.; Ljubič, T.; Zupan, J. A System for Creating Collections of Chemical Compounds Based on Structures, in 'Data Handling in Science and Technology', *Scientific Computing and Automation (Europe)* **1990**, Elsevier, Amsterdam, *volume 6*, 393-405.

26. Lah, L.; Tušar, M.; Zupan, J. Simulation of ^{13}C NMR spectra. *Tetr. Comput. Method.* **1989**, *2(2)*, 5-15.