MOLGEN 3.5

MOLECULAR STRUCTURE GENERATION

Award winning

scientific software

for chemistry

Reference Guide

Second Edition, March 2009

Ralf Gugisch - Adalbert Kerber - Axel Kohnert - Reinhard Laue
Markus Meringer - Christoph Rücker - Alfred Wassermann

Preface

The program system MOLGEN is devoted to the generation of all structural formulae (= connectivity isomers) that correspond to a given molecular formula, optionally under further constraints, e.g. the necessary presence or absence of substructures. It arose from the idea to provide an efficient and portable tool for molecular structure generation and visualization in chemical industry, research, and education.

MOLGEN serves very well as the mathematical heart of a program system for molecular structure elucidation, since it provides all mathematically possible candidates that correspond to a given set of chemical data. It is the outcome of research projects supported by the Deutsche Forschungsgemeinschaft and the BMBF, to which we like to express our thanks herewith. The system consists of the following components:

1. the generator MOLGEN for generation of connectivity isomers,

2. the structure editor MOLED for drawing input for the generator,

3. the module MOLVIEW for three-dimensional placements calculated by the built-in optimizer using a simplified version of the MM2 energy model,

4. a generator of configurational isomers corresponding to a constitutional isomer and their 3D realizations.

This manual is divided into seven chapters:

- The first chapter describes installation and basics of the program running on a first example.

- The second chapter comprises important notes for the user that should prevent possible errors due to misunderstanding the mathematical concepts underlying the algorithms.

- In the third chapter, the reader finds a description of the details important for an efficient use of MOLGEN .

– The fourth chapter mentions some strategies that are useful for more demanding cases and some peculiarities that arise from the underlying concepts.

– The fifth chapter is an appendix, wherein the historical development of the isomerism problem and the energy model used for three-dimensional placement of atoms in space are outlined.

– The sixth chapter is an appendix that describes error messages and gives some system limits.

– In the final seventh chapter we collect the literature concerning MOLGEN that was published in connection with the research projects that lead to the implementation of MOLGEN and its special versions.

Bayreuth, Freiburg, Munich, March 2009

Ralf Gugisch
Adalbert Kerber
Axel Kohnert
Reinhard Laue
Markus Meringer
Christoph Rücker
Alfred Wassermann

Contents

Chapter 1

First Steps

1.1 An introduction to MOLGEN

The program system MOLGEN enables you to generate the complete set of connectivity isomers (=structural formulae) corresponding to a given molecular formula and (optionally) further constraints. Hence you have to enter the molecular formula; for atoms unknown to MOLGEN , their valences are prompted.

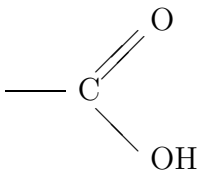
The construction is *free of redundance*, that is, you will not get any structural formula twice. Moreover, the construction is *complete*, which means that you will get the full set of all possible structural formulae that correspond to the given molecular formula and the prescribed valences of the atoms.

A MOLGEN run on the molecular formula C_6H_6 of benzene shows that there are altogether 217 isomers of benzene. And if you enter the molecular formula

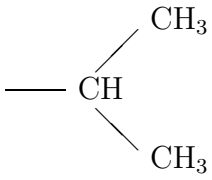
$$C_8H_{16}O_2$$

then MOLGEN will provide within a fraction of a second exactly 13,190 pairwise different connectivity isomers. These examples show that, in general, the total number of structural formulae corresponding to a given molecular formula is very large. Therefore you may want to reduce the output by imposing additional restrictions. For this purpose you can enter, together with the molecular formula, prescribed substructures e.g., a hydroxyl group, that have to be contained in the isomers, as well as forbidden substructures e.g., a ring of size 4, which you do not allow to show up. For example, if you prescribe, together

with the molecular formula $C_8H_{16}O_2$, the existence of a carboxyl group



MOLGEN will generate exactly 39 structures. If, in addition, you forbid that the isomers contain an isopropyl group



then from the former 39 just 27 isomers will remain.

An important point is, of course, how far MOLGEN can reach. The present version 3.5 allows generation of structures of up to 100 atoms. It must, however, be mentioned that due to the sometimes astronomical number of solutions, MOLGEN is not able to generate the complete set of structural isomers for all molecular formulae in a reasonable time, although the 13,190 isomers of $C_8H_{16}O_2$ are evaluated in a fraction of a second on a standard PC.

1.2 Installing MOLGEN

1.2.1 Hardware requirements

MOLGEN runs under all MS Windows 32 or 64 bit operation systems (Windows 95, 98, ME, NT4.0, 2000, XP, Vista). The following hardware is required:

1. An IBM-compatible PC (80486 or higher) running MS Windows 95 (or higher),
2. An interface for accessing the installation file, as for instance a CD-ROM or DVD-ROM drive, or alternatively a network connection.
3. A mouse is required as some parts of the program can be used exclusively by a mouse.
4. About 5.0 MB storage space on the hard disk. The space which is needed for a run depends, of course, on the particular problem and its complexity. For saving 1,000 structures one needs (depending on the number of atoms) between about 80 and 500 kB of space.
5. A printer is desirable.

1.2.2 Installation

If all these requirements are fulfilled you may now start installing MOLGEN . To do this, carry out the following steps:

1. Insert the MOLGEN installation medium in to the appropriate drive. In case you received the installation files as a compressed zip archive, uncompress these files into an arbitrary temporary folder.

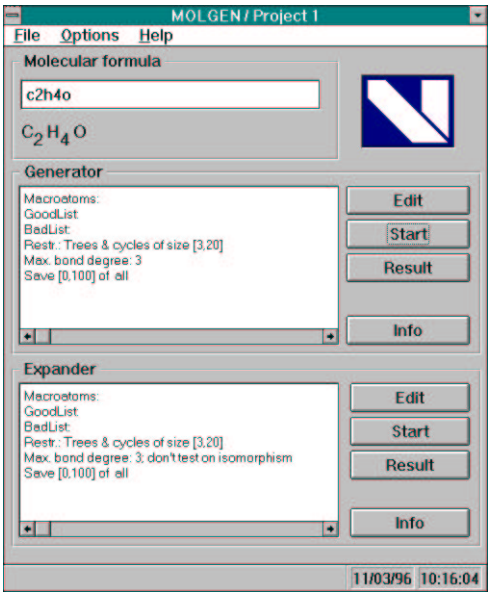
2. Open the EXPLORER and click on the folder of your installation medium (CD-ROM or DVD-ROM, or temporary folder with installation files). Then double-click on the Setup program icon to start the installation. Follow the instructions of the MOLGEN -MS installation program.

When the installation is carried out, MOLGEN is ready to start and you can execute it by double-clicking on the MOLGEN icon in the program manager or in the MOLGEN installation directory. You are now ready to run a first, simple example.

HINT: On some operation systems the default directory proposed by the installation program contains blanks in its file name, e.g. C:\Program Files\Molgen3.5. This can harm the failure-free operation of MOLGEN . It is strongly recommended to use an installation directory without blanks in the path name, e.g. C:\Programs\Molgen3.5.

1.3 A first example

Here we want to generate all isomers corresponding to the formula C_2H_4O. We therefore start the program system by double-clicking on the MOLGEN icon in the program manager. After the opening screen, you will see the following project window of MOLGEN :



Inside this project window you can enter the molecular formula in the Molecular formula field. Below, you can see which restrictions apply to the generator and expander. They will be discussed later.

Activate the Edit field in the molecular formula section by entering

c2h4o or C2H4O

The molecular formula should now be visible immediately below the edit field, formatted as C_2H_4O. This is all the necessary input.

Start the generator by pressing the START button in the generator section. A new window then appears displaying the number of isomers computed, the time needed so far and an estimate of the percentage of isomers already constructed. You might have missed this dialog since only 3 isomers were generated and it disappeared before you got the chance to abort it by pressing the CANCEL button.

After generation, information on the result is displayed on the screen. In the present example, 3 isomers are generated and stored. Further details of this info window will be discussed later. Here we leave this window via OK, returning to the project window of MOLGEN .

In principle you have finished the present example in which you have generated the three isomers that exist. To see them select the RESULT button. The generated isomers are displayed in three numbered boxes.

If you wish to quit, closing the window by a double-click on the upper left corner or by choosing 2D GENERATOR - CLOSE from the menu brings you back to the project window, and from there you may leave MOLGEN .

Perhaps now you think that the effort for obtaining these three isomers is considerable, since in this case you are certainly able to find the desired solutions with pencil and paper. Try a larger example such as $C_8H_{16}O_2$. You will be astonished how many structural isomers exist, and you will notice that you are **not** able to solve this task with pencil and paper!

In this introductory example we described every step in detail; hopefully many options of the program are self-explanatory and you will quickly get used to them in further program sessions. On the other hand you should have gained confidence in the structure generator in a very simple case, because you will have to rely on its correctness in more complex cases.

Chapter 2

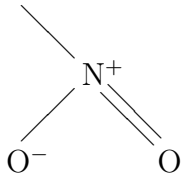
Important notes

Comments from users of MOLGEN showed that in some cases MOLGEN does not produce the result expected. To avoid this, a short list is given which you should consult before contacting the authors. Most of the cases occur due to the mathematical representation of molecules within MOLGEN . It is often very useful to make up your mind about how MOLGEN handles them.

The mathematical model behind the molecules in MOLGEN is that of molecular graphs. This means that for MOLGEN structures consist of nodes (atoms) with certain attributes (valence, atom type) and connections between these nodes (bonds) whose value is the degree of the bond, an integer. If you keep this fact in mind while working with MOLGEN , it should be easier to understand some "strange" behavior.

2.1 Charges

For MOLGEN , every atom type has a prescribed valence. This is why one element cannot occur with different valences in a molecule. MOLGEN cannot deal with charges. A well known example for charged atoms is the nitro group:



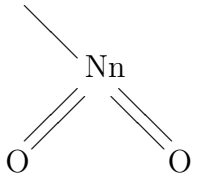
The user should carefully note that

MOLGEN *cannot handle (sub)structures with charges such as this directly!*

Hint: If a nitro group should be a part of the compound in question, it is easiest to reduce the molecular formula by two oxygen atoms and one nitrogen atom and to introduce a

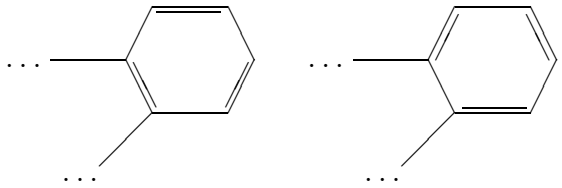
user-defined atom "Ua" of valence 1. Though the NO_2 group will not explicitly occur, you can identify it at once.

Another possibility is to define two new atom types "Np" for N^+ of valence 4 and "Om" for O^- of valence 1. If you prefer this solution, you have to remember that these new atom types can also be connected to other atoms in your molecular formula. Or: Simply define atom type "Nn" of valence 5 and use the following macroatom:



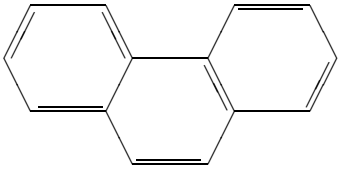
2.2 Aromatic structures

During the construction process, MOLGEN does not have any information about aromatic parts of a structure. Moreover, it does not know a particular bond type "aromatic" since, as mentioned above, bonds have an integer degree. This should be no problem as long as you do not expect MOLGEN to take care of aromaticity information. The following two structures, though completely different from the graph theoretical point of view, represent the same molecule:

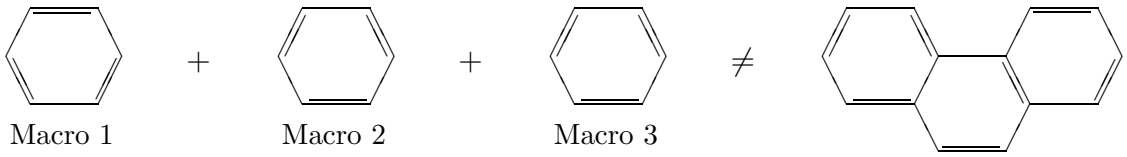


> MOLGEN *therefore offers a reliable* filter *to delete aromatic duplicates from the set of constructed isomers.*

You can run into difficulties, if you want to get fused aromatic systems such as phenanthrene:

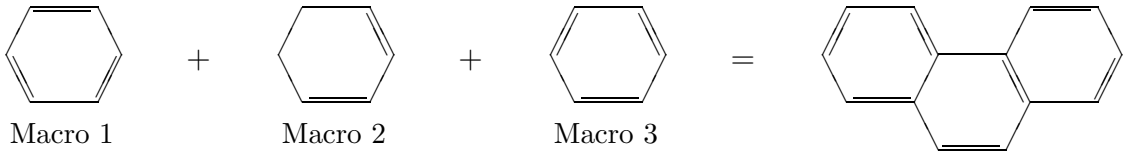


If you know that this system of three aromatic rings is contained in the unknown structure, you **must not** prescribe three benzene rings



as macroatoms, since for MOLGEN one of the rings has two double bonds only and you would not find the above example within the constructed isomers. If you use one benzene in the goodlist, phenanthrene is constructed properly.

Hint: As macroatoms in general provide a much more efficient way to reduce the number of constructed isomers, every information available should be used there. So if you know about the aromaticity of a structure, for up to two benzene rings, enter the above macroatom benzene, but for any additional one, use the following fragment. This will, of course, result in a larger set of isomers than intended, but since the resulting number is quite small, it can easily be filtered afterwards by adding the original number of benzene structures to the goodlist, as outlined above.

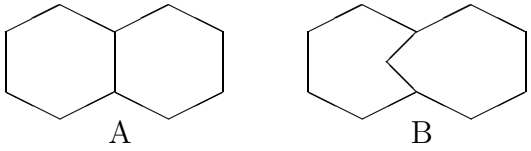


2.3 Ring sizes

The size of rings in isomers constructed by MOLGEN is measured as follows:

> *A ring is a closed path in the molecular graph such that no atom is immediately connected to more than two others in the path.*

The following picture should clarify this definition:



Structure A consists of two 6-membered rings, whereas in example B, in addition to the two 7-membered rings, the enveloping 10-membered ring is also perceived. So any time you want to add an upper ring size limit, keep this example in mind.

2.4 H distribution and macroatoms

Note that a H distribution is used only for atoms not being part of a macroatom. For the generation, it is recommended to use exclusively either the H distribution feature or the specification of macroatoms. If you have to use both, decrease the number of prescribed atoms with a certain H distribution by the number of those occurring within macroatoms.

Chapter 3

User reference

3.1 The structure generator

3.1.1 The molecular formula

After the brief introductory example in chapter 1, we are now going to explain the main procedures of MOLGEN with the aid of a large family of isomers. For this purpose, we consider the molecular formula $C_8H_{16}O_2$. The molecular formula is the first and main piece of input to MOLGEN . It comprises the following notions:

- The **atom type**, by which we mean the chemical element the atom belongs to. It is entered by the symbol used in the periodic table of elements.

- The **frequency** of an atom type, i.e. the number of atoms of this element contained in the molecular formula.

- The **valence** of an atom type, the total number of covalent bonds that connect an atom of that type to other atoms in the molecule (a double bond requires at least valence 2). The valence is the number of valence electrons or the number of missing electrons compared with the next noble gas configuration. Thus, the valence of H is 1, of O is 2, of C is 4, etc.

3.1.2 Entering a molecular formula

The molecular formula is entered in the project window of MOLGEN

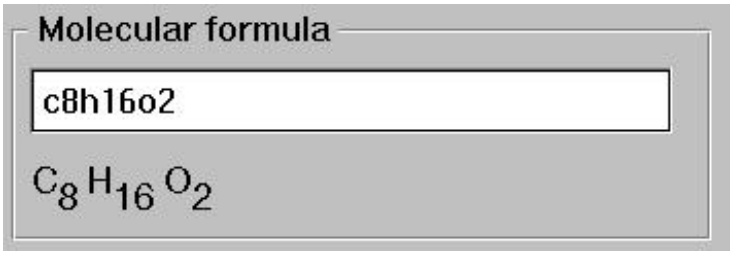


Figure 3.1 Molecular formula field

Enter the molecular formula $C_8H_{16}O_2$ as

c8h16o2 or C8H16O2

Optionally, a blank or a " - " character can be entered between the atom type and its frequency, but such blanks are no longer required. As usual, the omission of a frequency defaults to 1. For instance, the input of C2H6O is interpreted as $C_2H_6O_1$. Atom symbols must fulfill the following convention:

- An atom symbol consists of one or two letters. While in the first case capitalization is not critical, in the second case **the first character must be a capital letter**.

Valences are known for C, H, O, N, F, Cl, Br and I. If you want to use other atom types, you have to specify the valence explicitly:

There is also the possibility to define atom types not known to the system!

In this case the frequency (even if it is 1) and the valence of an atom must be added. Valence and frequency have to be separated by either a blank " " or a " - " character. As an example insert

Qs3 2c4o

In the area just below the edit field, the correct representation of this molecular formula Qs_3C_4O will show up. Your input is partially checked automatically, and an error message will be shown if you insert anything wrong. The user-defined atom Qs of valence 2 will now occur three times in each resulting structure.

3.2 Running the generator

3.2.1 Starting the generator

Start generation of structures by pressing the START button in the MOLGEN project window generator section.

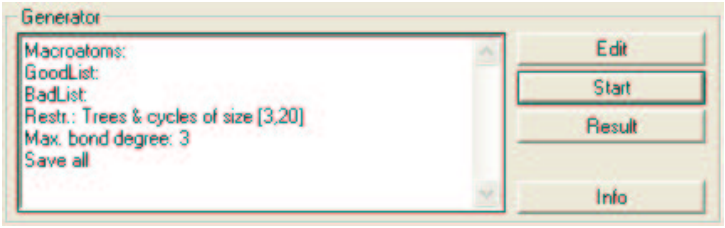


Figure 3.2 Generator section

During the construction, an estimation for the number of solutions is computed. This value is used for displaying the *approximate percentage* of molecules already constructed, indicating the progress of the generation.

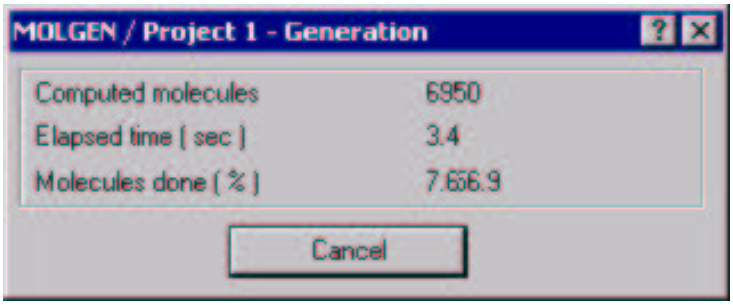


Figure 3.3 Generation status window

You may stop the generator at any time by pressing the CANCEL button. All the resulting structures (up to the limit set in the generator section) computed so far will be saved.

Carrying out the above example, after a short time which depends on the performance of your computer and the operating system, you get another window, displaying some information on the generation process. It tells you that there exist exactly 13,190 distinct structures with that molecular formula, that 100 of these were saved and that 0.21 seconds of cpu time were needed to do these calculations.

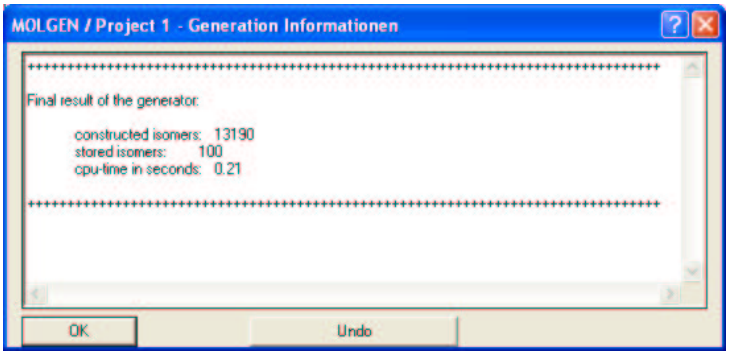


Figure 3.4 Info Window

You can either accept this message with OK or get back to the status of the previous generation by clicking the UNDO button. Note that in this case all changes made in MOLGEN after the previous generator run are definitely lost.

This window can also occur when the generation cannot be started due to some wrong input. Normally, these errors are described inside the window and at some point there may also be a CONTINUE button that overrides the error messages and starts the generation despite of these warnings.

At any time after the generation, you can display this window showing the result of the previous generator run by simply pressing the INFO button (see Figure 3.2).

Of course, you can enter a lot of constraints on the structures to be generated. All these options are described in the next section.

3.3 Restricting the number of isomers

Apart from the molecular formula, MOLGEN provides many possibilities to restrict the number of structures that are generated. Since even a rather small molecular formula such as $C_{10}H_{10}$ leads to the surprisingly large number of 369,067 isomers, you can see that restricting the output is an essential part of MOLGEN .

To allow input of generator restrictions, press the EDIT button in the generator section of the MOLGEN project window (see Figure 3.2).

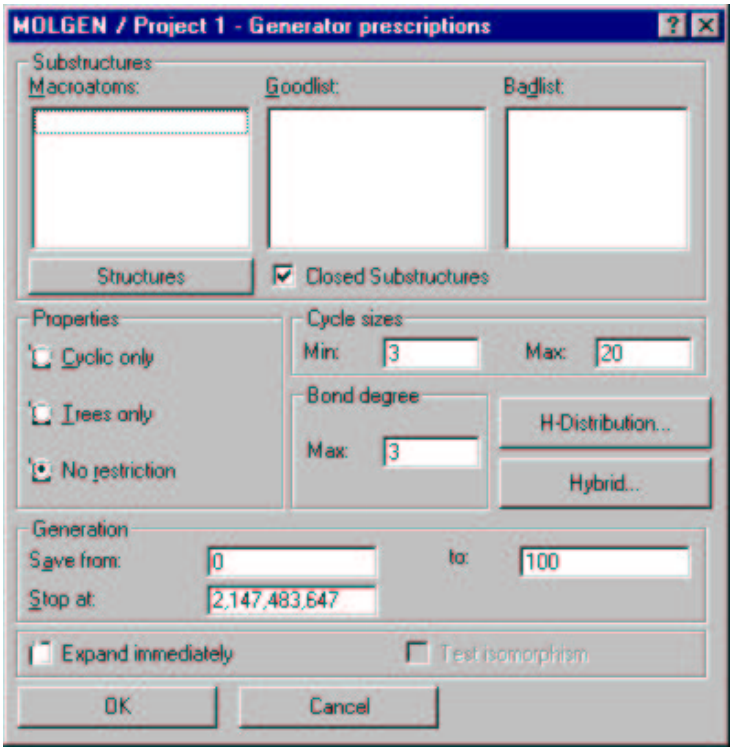


Figure 3.5 Generator prescriptions

3.3.1 Prescribing structural properties

First, we take a look at some of the easier restrictions. In the PROPERTIES section, you can enter some additional information about the structure of the resulting isomers.

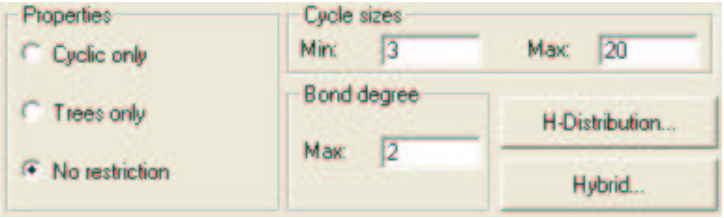


Figure 3.6 Structural restrictions

- TREES ONLY means that rings must not occur; with CYCLIC ONLY selected, no isomers without rings are generated. NO RESTRICTION says that both possibilities are allowed. In our example a restriction to trees reduces the result from 13,190 to 5,899 isomers.

- Moreover, you can limit the size of rings. If you leave MIN and MAX to the default values 3 and 20, respectively, rings of any size between these bounds are possible. Otherwise, minimum and maximum size of rings are determined by the values entered.

- The item BOND DEGREE controls the maximum allowed bond degree. The value 2, for instance, admits single and double bonds only.

3.3.2 Number of structures computed and saved

Since the whole set of isomers is generated by MOLGEN and since this set may be extremely numerous, it is also possible to abort the generation after a certain number of structures.

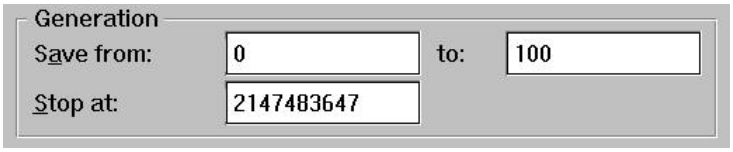


Figure 3.7 Generation number restrictions

- By default this number is set to the maximum integer representable by your computer. A different number can be entered in the STOP AT field.

- The number of structures saved on your disk can also be edited here. The amount of disk space used for a single structure depends mostly on the molecular formula.

There are some general rules on using numerical input fields found in MOLGEN . You can insert the maximum value possible in any numerical field simply by pressing the PAGE UP button on your keyboard. The minimum value available is reached by pressing PAGE DOWN. With the CURSOR UP and CURSOR DOWN keys, a certain amount is added to the number displayed in the field. Almost every numerical entry field in MOLGEN can be manipulated this way, so all values are easily changed in order to fulfil your requirements. Of course, you can also insert the correct number directly.

For the buttons H DISTRIBUTION... and HYBRID... see Sections 3.7 and 3.8.

3.4 Using macroatoms

Now the most powerful feature of MOLGEN will be discussed: MOLGEN knows *three different types of substructures* - macroatoms, goodlist and badlist structures.

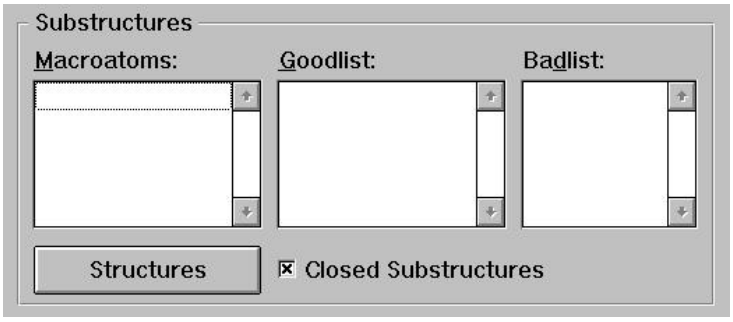
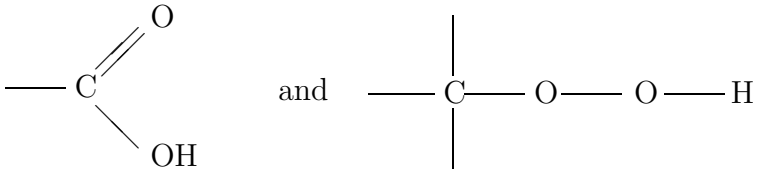


Figure 3.8 Substructure section

Generating thousands or even millions of structures by only using the restrictions described in the previous chapters is of course still inappropriate for further considerations. Usually, chemical analysis can detect certain substructures occurring in the molecule in question. The structure generator offers the option to take such a substructure into consideration, e.g. in the form of a so-called *macroatom*. A macroatom symbolizes a group of several connected atoms which are treated as a single one by the generator. This technique, which might be hard to understand at first glance, will be illustrated immediately by an example.

Assume that the desired molecule of our family of isomers should contain a carboxyl group COOH. The generator, however, which works far more generally and which uses only little chemical information, does not know by itself which one of the following two structures is meant,



For a substructure, the generator only uses the name under which the prescribed molecule part is stored together with the information about the atoms contained therein and their connections.

Several substructures are saved in the subdirectory BIB. Besides, as we shall see later, there is the possibility to input structural groups of your own. The existing file named COOH consists of the carboxyl group. So let us include it as described in the next section.

3.4.1 Including a stored substructure

You have several possibilities to insert into one of the three listboxes a substructure (normally stored in the BIB subdirectory).

- If the name of the substructure is known to the user, it can be inserted manually. Just click on the first free entry or the very top of one of the three listboxes. A popup-menu will occur looking like

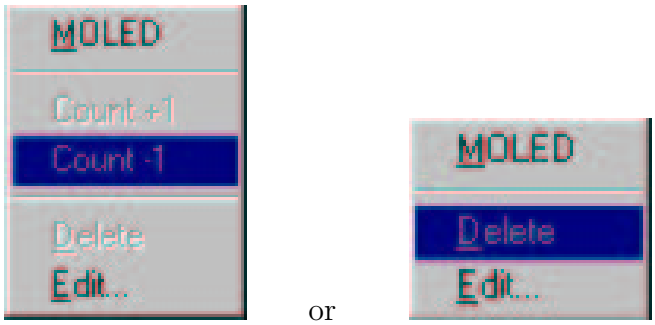


Figure 3.9 Substructure popups

In the badlist, where substructures are inserted that must not occur in the resulting structure, the options COUNT+1 and COUNT-1 are senseless and therefore missing.

In order to insert a structure now, click on the EDIT... option of the menu and the following dialog will appear:

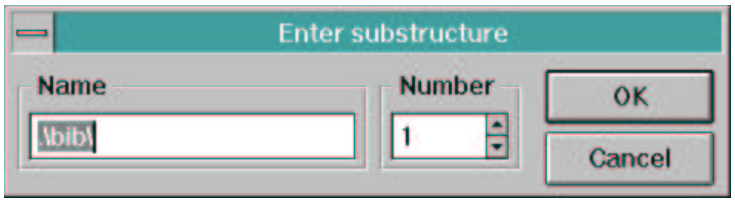


Figure 3.10 Substructure insert by name

Now you can insert the path and filename of the desired substructure in the NAME field. In the NUMBER numeric entry field, enter the substructure's occurrence number. If you enter this dialog from the badlist, the NUMBER option is unavailable.

To remove a substructure from this list, click on the DELETE option in the popup-menu above.

You can also start the molecule editor MOLED (see Section 3.12) to draw a new structure. Just click on the MOLED option to start it.

Usually, the substructures have a MOLGEN – internal format, but it is also possible to insert structures in the well known MDL MOLFILE format. They are automatically converted to the MOLGEN format. How this is done will be described later in this chapter.

- All listboxes provide full DRAG & DROP functionality with respect to files. So you can drop any substructure file for example from the file-manager into one of these boxes. The occurrence number will be set to one except for a badlist entry. Furthermore, entries can easily be moved from one list into another, using the DRAG & DROP mechanism.

- Click on the STRUCTURES button in Figure 3.4 and the following dialog will enable you to browse through the filesystem and to select the substructure you want.

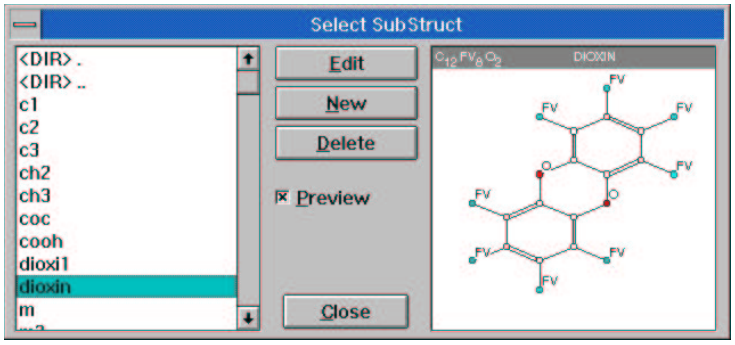


Figure 3.11 Substructure select dialog

In addition to that, a small preview window is provided where you can take a look at a substructure before dragging it out of the file listbox into one of the three substructure boxes.

- Clicking on the EDIT button will open the MOLED application (see Section 3.12) where a substructure can be edited if it is not of the kind you want. Similar to that is the NEW button that brings up MOLED without loading a structure. The DELETE button will erase the structure selected in the listbox on the left-hand side, that is it will physically be removed from the file system. With the PREVIEW checkbox, the preview window can be displayed or removed.

MDL MOLFILE structures are not displayed in the preview window since they may contain more than one structure. Nevertheless, they can be dragged into the substructure section and are identified and split automatically.

- Clicking on a molecule picture in the preview mode will bring up the following popup menu:

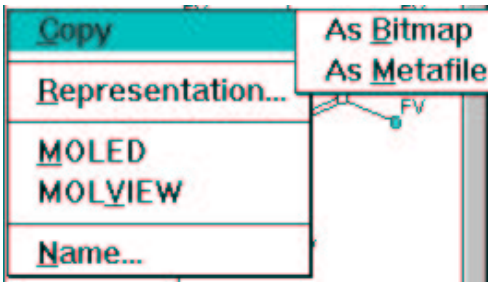


Figure 3.12 Molecule preview popup

The COPY option provides the possibility to copy the structure displayed in the window to the Windows clipboard. You can choose between a BITMAP and a METAFILE representation. Most Windows applications will accept the bitmap format, but some of them refuse to work with a metafile. Nevertheless, the metafile format gives you a wider variety of manipulating the structure in a program using vector graphics, such as Corel Draw.

The REPRESENTATION... allows you to change the layout of the structure inside the preview window. That dialog will be discussed later (see Section 3.12.9, Figure 3.42).

MOLED will start the MOLED application as mentioned before. MOLVIEW calls the MOLVIEW application which enables you to compute a 3D representation of the current structure. MOLVIEW is described later (see Section 3.13).

- The NAME... option gives you the possibility to name the current structure. The name is stored internally and may differ from the filename where the structure is stored. The dialog for entering a name is:

Figure 3.13 Internal molecule name

If you insert a file containing MDL MOLFILE substructures into one of the three boxes, the following dialog will be displayed:

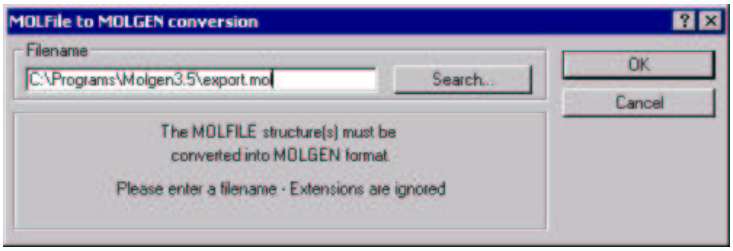


Figure 3.14 Molfile conversion dialog

The extension of the filename you enter will not be used. The structures described inside the MDL MOLFILE file are automatically saved as e.g. export.1, export.2, ... You can choose an appropriate file using the default system dialog by pressing the SEARCH... button. The files created by the program are automatically inserted into the box instead of the original MDL MOLFILE file.

For our example of computing all isomers with molecular formula $C_8H_{16}O_2$, please include the macroatom COOH. After a new run of the generator you will get the message: *Isomers constructed*: 39. This clarifies the effect of adding a macroatom. The number of possible isomers is usually reduced drastically. The generator constructs the remainder of the molecule only, where of course fewer possibilities exist than in the complete family. Let us have a look at the structures generated with the COOH macroatom prescribed:

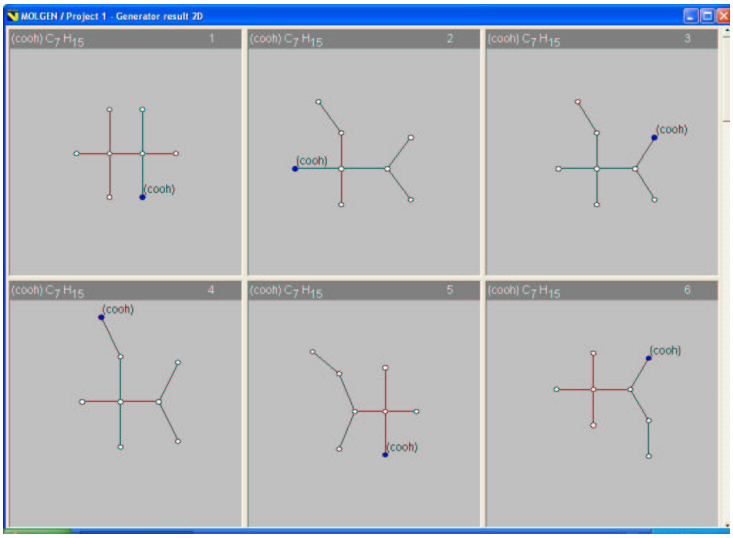


Figure 3.15 A generator result

As you see, the COOH group is represented by a single atom and the molecular formula is reduced by the atoms used in COOH. To show this molecule without the (COOH) atom but with the whole substructure you have to expand it. The next paragraph will show you how this is done.

3.5 Expanding macroatoms

If atoms were combined to macroatoms, this process has to be reversed after generation. This is the task of the *expander* module. One "atom" COOH, for example, is blown up to a substructure consisting of one C, one H and two O atoms connected by the correct bonds. At the same time, the generated structures are tested for identity, because after use of certain macroatoms expansion may produce the same isomer more than once. To use the expander, press the EDIT button in the Expander Section of the MOLGEN project window.

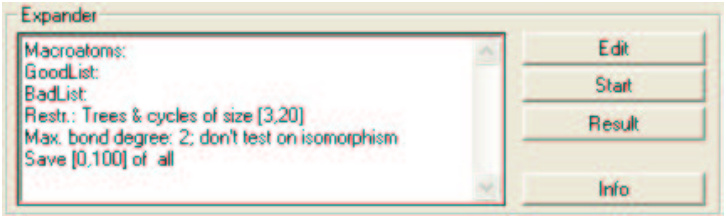


Figure 3.16 Expander section

Make sure that you have used the generator with macroatoms, since otherwise nothing can be expanded. The window appearing now is very similar to the generator prescriptions window (see Figure 3.5).

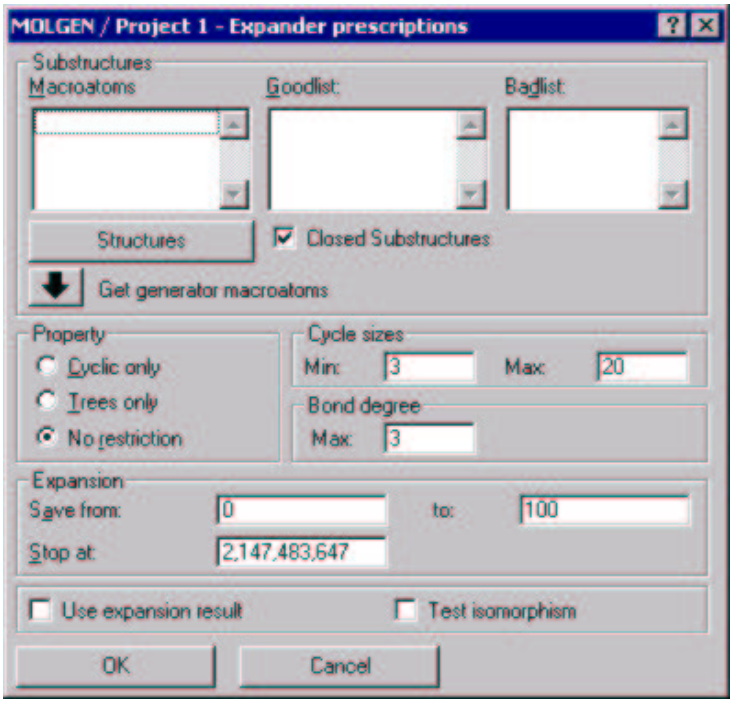


Figure 3.17 Expander prescriptions

Now you can add restrictions or substructures as described in the generator input window. If you want to use the same macroatoms as with the generator, click the button

GET GENERATOR MACROATOMS. Among the restrictions you also find the checkbox TEST ISOMORPHISM that activates and deactivates, respectively, the test on identity of two expanded structures. If this option is activated, each structure generated by the expander will be compared with all previous ones, since it might already exist. The time required for expansion will increase when using this test, so you should apply it only to a relatively small number of isomers. Note that the number of expanded structures may be larger than the number of those generated before. It might, however, decrease by the isomorphism test.

After clicking the START button in the expander section (see Figure 3.16), expansion of the macroatoms will start and a window similar to Figure 3.3 will show the progress of expansion. If you want to run the expander a second time, e.g. for adding further restrictions, you can save time by enabling Use expansion result. In this case the expander processes the structures resulting from the first expander run only, rather than all structures resulting from the generator run.

If you are not yet satisfied with the result (since there are e.g. no isomers left due to restrictions being too strong), you may return to the state before the run just finished by selecting UNDO in the information window arising after the expansion process. Then you may repeat the expansion with a modified input. You can abort the expander at any time by pressing the CANCEL button. All resulting structures saved so far will be kept.

Having prescribed a COOH for our example family, try to expand the constructed isomers now. To do this, use the GET GENERATOR MACROATOMS option in order to avoid entering the macroatom once more, then start the expander.

If you take a look at the 2D drawings now, the COOH group consists of four atoms, as expected.

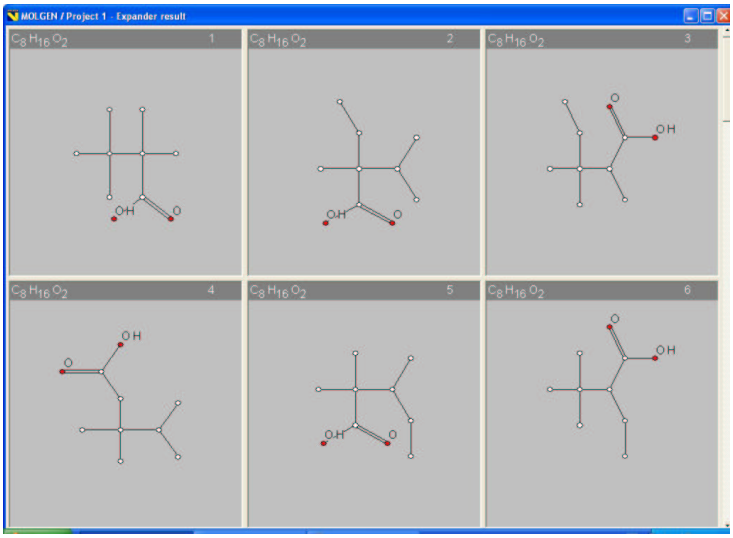


Figure 3.18 An expander result

The expansion may still be a dubious procedure to you. But take into account the enormous saving by using macroatoms, especially with large molecules, since the generator always runs through all possibilities. Their number grows exponentially with the number of atoms and may reach astronomic orders of magnitude. The contrast between 13,190 and 39 in our example and the corresponding time saving make this fact obvious.

One question, however, remains unanswered: Why can the size of the molecule family increase by expansion? Let us take a macroatom with several free valences, e.g. the one called dioxin, $C_{12}H_4O_2$. Its 4 free valences are indistinguishable to the generator, a bond to one of these is equivalent to a bond to any other. Therefore for the dioxins $C_{12}H_4O_2Cl_4$ the generator constructs exactly one structure. After expanding the macroatom the free valences are no longer equivalent, and the number of structures therefore increases from 1 to 22. This example will be treated in more detail later. Besides explicitly using the expander, you can also expand the macroatoms during the generation process. For the beginner, acting this way should be easier to understand. Just use the EXPAND IMMEDIATELY option in the generator input window (see Figure 3.5)

Figure 3.19 Expand immediately

In this context you can, of course, enable the *isomorphism test* as described before. Starting the generator now would immediately lead to the resulting structure in Figure 3.18.

3.6 Using goodlist and badlist

Besides macroatoms there is another possibility to include substructures in the generating process.

3.6.1 The goodlist

By the notion *goodlist* we mean one or several groups of atoms that have to occur in the generated molecules. The difference to macroatoms is that two substructures in the goodlist may overlap. Suppose it is known from spectroscopic analysis that the molecule comprises an alcohol group C-OH and a carbonyl group C=O. If demanded as macroatoms, both have to occur separately. If the same groups are goodlist items, the occurrence of a single carboxyl group COOH fulfills both requests. The goodlist works as a filter,

which *after* construction sorts out all molecules that do not contain the desired groups. A *multiple* occurrence of a goodlist structure can be requested. If CO is required twice, all structures that comprise CO only once or not at all will be rejected. A goodlist entry is also regarded as occurring multiply if two copies overlap in *parts*. Thus, a result containing -C-O-C- fulfills the goodlist demand of *two* C-O groups.

> *When generating with* ***macroatoms*** *only the remainder of the molecule is constructed. Using a* ***goodlist, all*** *isomers corresponding to the molecular formula are constructed, but only those are kept that contain the desired groups.*

If two or more substructures are entered in the goodlist, MOLGEN will understand them to be linked by **and**, i.e. only those structures will be kept that contain each of these substructures. However, by clicking on logical or (in the goodlist window) it is possible to compose a list of substructures that are to be understood as linked by **or**. With such a list in the goodlist window, all those structures will be kept that contain at least one of the substructures.

3.6.2 The badlist

The *badlist* has exactly the opposite effect. Only such isomers will be kept in which the listed substructures do *not* occur. When filling the goodlist you add the multiplicity of its occurrence to each item, whereas in the badlist you simply ban the group itself. You can use good- and badlist with the generator as well as with the expander. The function is similar, but not exactly the same. Peculiarities about the effect may be looked up in Chapter 4.

Goodlist or badlist items are entered analogously to macroatoms, as described in Subsection 3.4.1. New groups can be created using the editor MOLED . For our example $C_8H_{16}O_2$, include the isopropyl from MOLED in the badlist. After finishing the construction the generator notifies that now there are only 11,231 isomers, 11,204 without COOH and 27 isomers with COOH. Finally create *n*-hexane with two free valences at its ends and insert it in the goodlist with multiplicity 1.

After running generator and expander again we have reached our goal: By using more and more information we have reduced the family of isomers so that in the end there is a single structure left. Here it is:

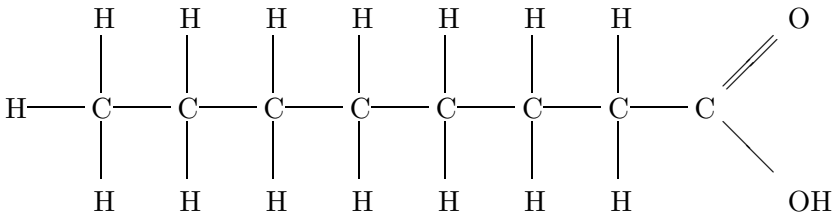
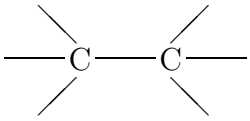


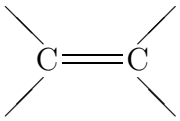
Figure 3.20 Final structure

The CLOSED SUBSTRUCTURES checkbox only affects the use of good- and badlist. If it is enabled, free valences that are bonded to different atoms in a substructure are not allowed to become connected to each other during the generation. So *rings or multiple bonds within a substructure will not be created*. If it is disabled, these restrictions do not apply, so that the number of generated structures increases. Note that substructures used as macroatoms are always closed according to this definition.

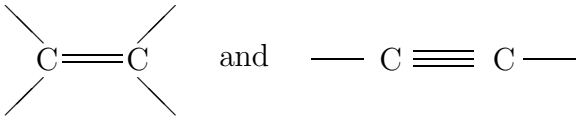
As an example, have a look at the following substructure:



If you use it as a macroatom, free valences cannot be connected to each other inside this structure. So the macroatom is not expanded into a



group. If you use it as a goodlist item, structures containing



groups will be created unless you switch on Closed Substructures.

3.7 Using hydrogen distributions

By pressing the H DISTRIBUTION... button in the window in Figure 3.5, it is possible to enter known H distributions of C, N and O atoms.

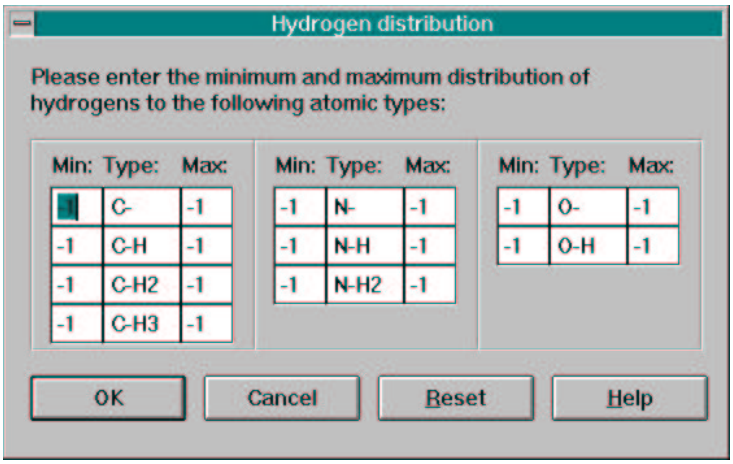


Figure 3.21 Hydrogen distribution dialog

Hydrogen distributions are a common result from spectroscopic analysis. MOLGEN allows to take full advantage of this additional information. Entering the hydrogen distribution in this dialog will strongly decrease the amount of structures generated.

The default value -1 means that any distribution is allowed. If you enter a MINIMUM, the MAXIMUM will be updated accordingly and vice versa. This simplifies the input of exact values. Intervals are also allowed. The RESET button will reset the entries in all fields to the default value -1.

If you know, for example, that the isomers you are looking for have the formula $C_8H_{16}O_2$ and have four primary (4 CH_3), two tertiary (2 CH) and two quarternary carbons (2 C), the total number of solutions corresponding to this data is just 96 — instead of 13,192 from the molecular formula alone.

Please note:

> *When using one or several macroatoms and simultaneously hydrogen distribution restrictions, be aware that the prescribed H distributions affect the atoms outside the macroatoms only.*

For atoms within a macroatom, MOLGEN in its present version is unable to take H distributions into account. In this case it is recommended to proceed as follows: Since a given H distribution decreases the number of possible solutions drastically, you can put your substructure into the goodlist. This way you take advantage of both the information about hydrogens and about substructures.

3.8 Using hybridizations

Pressing the HYBRID... button in the window in Figure 3.5, it is possible to enter known hybridizations of C, N and O atoms:

Hybridizations can often be found by examining spectroscopic data. Depending on the molecular formula, hybridizations can be entered for any C,O or N–atom. The small picture on the left side of the dialog shows a possible hybridization state for an atom. Using the scrollbar beside it, you browse through all hybridization states available with respect to the given molecular formula. Below, you can enter a minimum and maximum number of how often this hybridization occurs in the molecule to be elucidated. Additionally, an optional number of hydrogens connected to this kind of atom can be prescribed in the H: field. Having entered a number there, you can add the entries to the listbox on the right side of the dialog just by pressing the ADD button. Removal of an entry is easily done by marking it in the listbox and pressing DELETE. One entry only can be selected for deletion.

When using hybridization in combination with macroatoms, a note of caution is appropriate similar to the above on combination of hydrogen distribution and macroatoms: *Hybridization information is used for non-macroatoms only.*

The default value -1 in the H: field stands for an unknown number of hydrogens connected to the atom.

This dialog can only be called after a molecular formula was entered.

3.9 Displaying the result

After a successful generation, the saved isomers can be displayed by pressing the RESULT button in the MOLGEN project window (see Figure 3.2). A 2D placement of the structures is computed now, and several structures are displayed in the Result window.

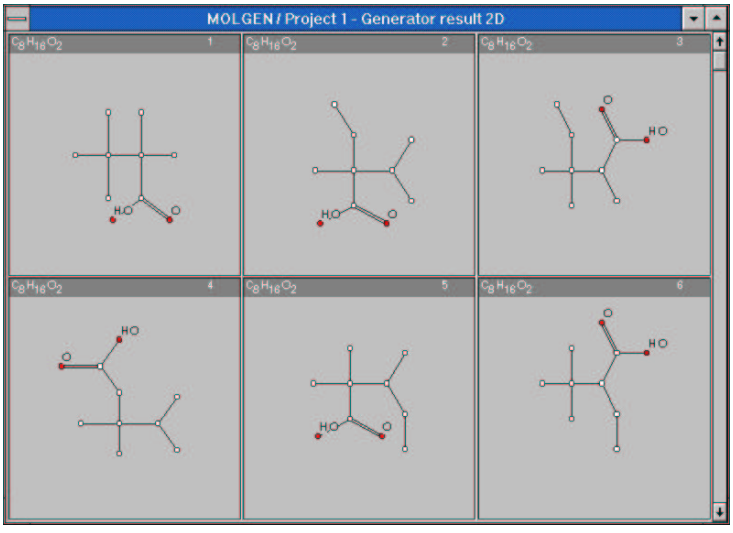


Figure 3.22 2D Result window

Since in most cases there are more than 6 structures, you can scroll through them using the scrollbar at the right.

Now you should regard every single structure as *one* object of the solution. You have many opportunities to manipulate these objects. Clicking on one of the structures, a popup-menu will occur, similar to the one in Figure 3.12.

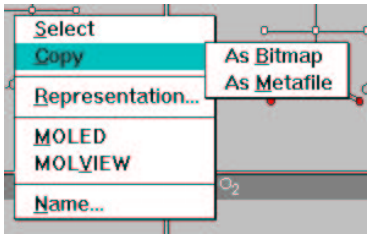


Figure 3.23 Molecule result popup

The COPY option should be well known from Subsection 3.4.1 and the NAME... option will open the same dialog as described there (see Figure 3.13). Every single structure can be named now.

Clicking on MOLED will extract the structure out of the set of solutions into a temporary file. Now, you can manipulate and save this structure using MOLED (see section 3.12). You might think about taking a certain part of a solution as a new substructure to decrease the number of computed isomers. MOLVIEW , as mentioned before, will compute a 3D representation of this single structure.

The REPRESENTATION... option opens the following dialog:

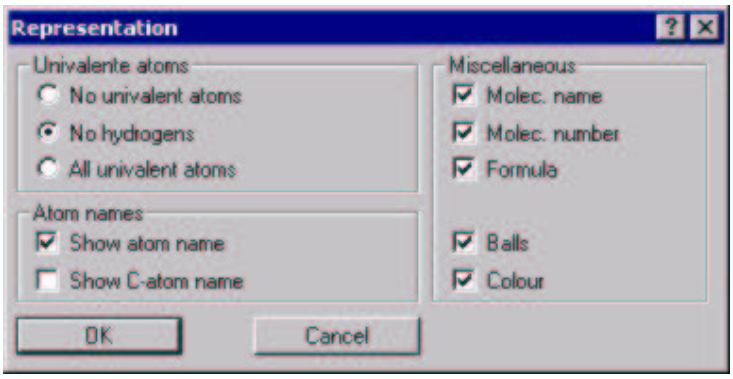


Figure 3.24 Molecule representation dialog

As you see, some options for drawing the isomer can be changed to make the structure look the way you like it. This dialog is also offered when clicking on the substructure in the preview window Figure 3.11. Note that changing the style here will change the representation of a single molecule only and not that of all structures. This can also be done and is described later.

The SELECT option in the popup menu will change the background color of the single molecule. Its number is registered now and you can use it to select a few structures for printing or exporting into MDL MOLFILE format.

These were actions to be taken on a single structure. Apart from this, it is also possible to manipulate the whole set of solutions. To do this, click on 2D Generator, which is available in the MOLGEN project window whenever the 2D result window is active (see Figure 3.25).

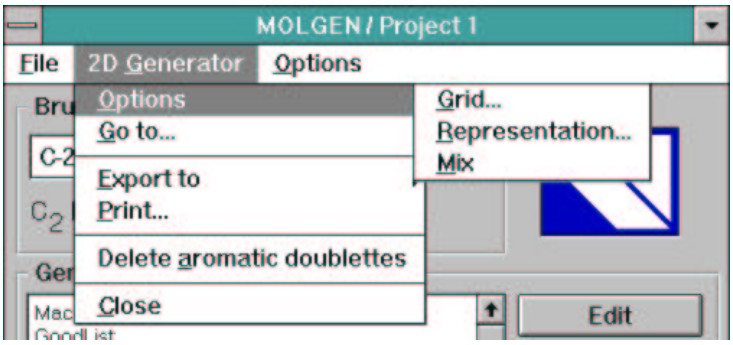


Figure 3.25 Result window menu options

Selecting OPTIONS/GRID... from the menu enables you to change the number of structures that are displayed in the result window.

Figure 3.26 Grid window

The size suitable for you depends on the resolution of your screen. High numbers in rows and columns may slow down the screen build up.

The OPTIONS/REPRESENTATION... dialog is known from before, but this time it affects the whole set of structures displayed.

Checking the OPTIONS/MIX entry mixes all structures, to give you an overview over the different types of structures generated. This may be useful since the generation algorithm produces similar structures close to each other (from the mathematical point of view). Choosing this option again reverses the mixing.

The dialog opened with the GO TO... option lets you jump directly to a desired structure without scrolling through the window.

Figure 3.27 Select a particular structure

3.9.1 Export to MDL MOLFILE

The EXPORT TO MOLFILE FORMAT... entry of Figure 3.25 offers you the possibility to export selected structures from the solution generated by MOLGEN into a file having the MDL MOLFILE format.

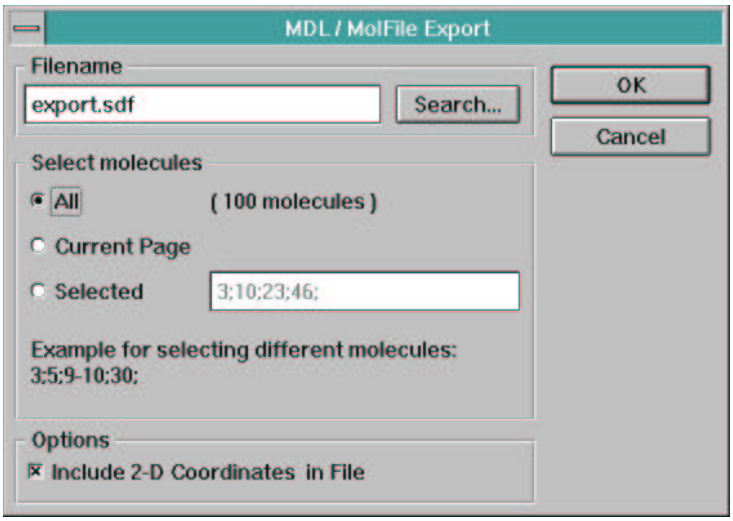


Figure 3.28 Export to Molfile

In the FILENAME section you can enter the name of the file to which structures the are saved. If you do not find a name you can select one by using the SEARCH... button, which will open a standard system SAVE AS... dialog to select the appropriate filename.

You can also easily select the structures to be saved inside the SELECT STRUCTURES section. Behind the ALL radio-button, the total number of structures saved on the disk is displayed. Remember that *all of these* are stored, if this option is chosen. The CURRENT PAGE option saves only the structures currently displayed in your result window. Their number depends on the grid you have chosen. If you have selected single structures in the result window, their number will be displayed in the edit field behind the SELECTED option. As described in the window, you can either enter the numbers of single structures, separated by a semicolon, or a certain interval where the lower and upper limit have to be separated by a " - " character.

In the Options Section it is possible to prevent the computation of the 2D coordinates by disabling INCLUDE 2-D COORDINATES IN FILE. This will save some time during the storing process and it is useful if you do not need the placement for further data manipulation.

3.9.2 Print structures

The PRINT... menu entry of Figure 3.25 opens a dialog similar to the one described before.

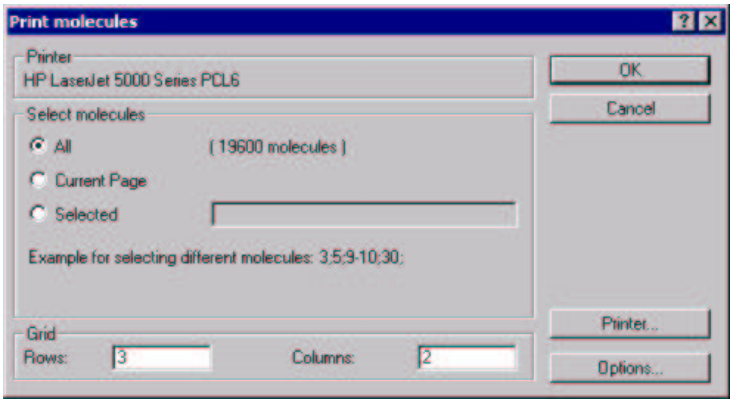


Figure 3.29 Printing structures

The Printer section shows the name of the current default printer. In order to change it, use the PRINTER... button that leads you to a system–specific dialog for selecting a default printer. The OPTIONS... button is used for changing printer-dependent options such as the resolution.

The SELECT STRUCTURES section is used for selecting exactly the structures you want to print. The ALL/CURRENT PAGE/SELECTED options have the same function as described in Subsection 3.9.1. In addition, it is possible to enter a new GRID for printing. By default, the grid shown on the screen is used, but often printers provide a higher resolution than displays, you can change the grid here. ROWS and COLUMNS are used with respect to the smaller and larger side of the paper, so if you enter more columns than rows, their values are exchanged automatically unless the landscape mode is active.

3.9.3 Eliminate aromatic duplicates

The DELETE AROMATIC DOUBLETTES option provides another tool for restricting the number of candidates. To understand its function, you have to understand how MOLGEN constructs all possible isomers according to your input. It uses the connection information between atoms, and a bond degree can only be 0,1,2,3,... . So MOLGEN cannot distinguish between aromatic and "normal" bonds. So MOLGEN constructs two or more solutions where the chemist perceives only one. To solve this problem, we can screen the solution set and eliminate identical (with respect to aromaticity) structures. Choosing this option will make a simple window appear showing you the progress of this procedure. You can cancel the process at any time by pressing the CANCEL button. After the elimination, a status window will appear showing how many aromatic duplicates were deleted.

3.10 Working with MOLGEN projects

3.10.1 Using projects

MOLGEN projects may be saved. To do this, choose FILE/SAVE from the main menu of MOLGEN .

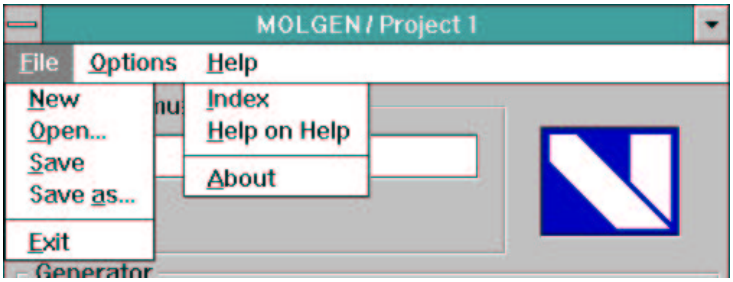


Figure 3.30 MOLGEN main menu options

On saving, the generator and the expander input only are saved. The result is not saved since it is quickly reproduced but would need a lot of disk space.

You can LOAD a saved project at any time by choosing the OPEN... option. NEW will discard all input done so far and start with a blank input. SAVE AS... will save a project by another filename. By default, every MOLGEN project saved gets the extension .mpr. You may change this default setting by entering another extension in the filesystem dialog.

EXIT will, of course, leave MOLGEN .

Choosing HELP/ABOUT opens the following information window about MOLGEN . You can read your license number here and find contact information.

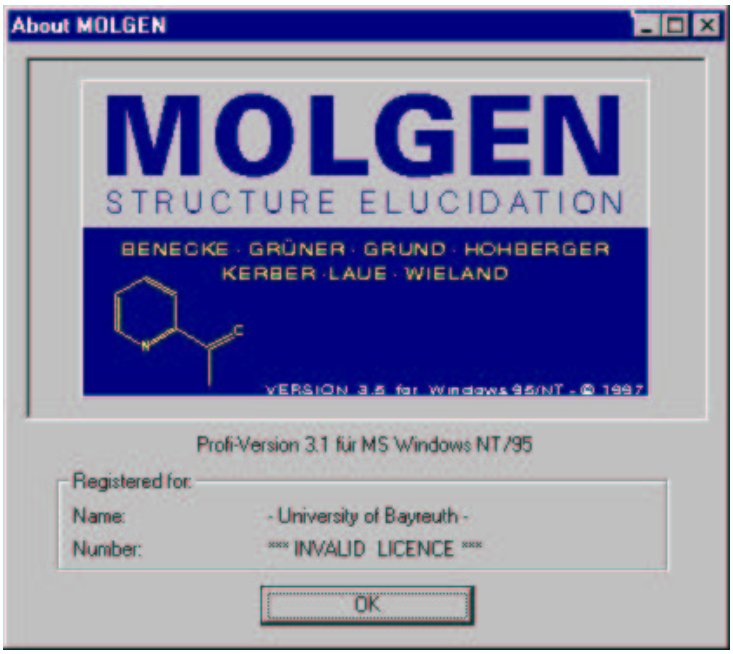


Figure 3.31 Information about MOLGEN

3.10.2 Default settings

Choosing the OPTIONS / SETTINGS... menu entry of Figure 3.30 will open the following window.

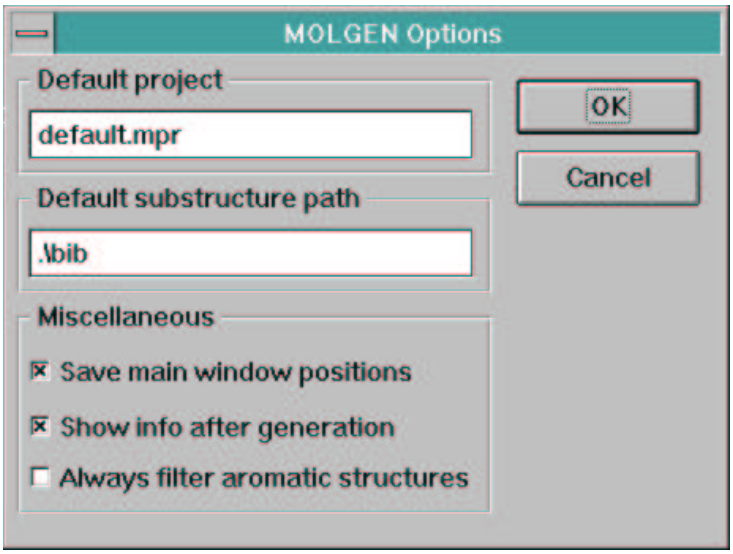


Figure 3.32 MOLGEN default settings

The DEFAULT PROJECT field offers an easy way to use your own default project every time you start MOLGEN or create a new project. Just fill in all desired options in the generator and expander input window and even the molecular formula and save these settings under any project name you want. If you do not want to save a certain molecular formula, please ignore the message that the molecular formula is incomplete. The remainder of the input will be saved correctly.

Of course, this is an easy way to provide e.g. a *permanent badlist*, containing substructures that do not make sense. Enter all these structures in the badlist of the generator and save the project as described before.Then enter the filename in this edit field and you will automatically get all the structures as a part of the badlist whenever you start MOLGEN .

The DEFAULT SUBSTRUCTURE PATH is the default path the substructures are searched at. You can also enter a path accessible via network here.

SAVE MAIN WINDOW POSITIONS enabled will save the position of the windows on your screen. So you do not always have to move them to the desired position. Additionally, the grid used for displaying results will also be saved.

SHOW INFO AFTER GENERATION is provided for those who do not want the INFO window to pop up automatically after a successful generation or expansion. However, it is always shown when an error occurred or when you abort the generation by pressing the CANCEL button.

If you enable the ALWAYS FILTER AROMATIC STRUCTURES checkbox, after each generator or expander run, structures are filtered by a special algorithm to eliminate aromatic duplicates. For further information see Subsection 3.9.3.

3.11 Error handling

Wrong input or other mistakes in usage will produce errors that the program reports to you. Continuing without repairing the error is often impossible. Therefore MOLGEN error messages include hints how the error can be corrected (for details see Appendix).

3.11.1 Receiving error messages

- If an error occurs or you interrupt the generation (expansion) by pressing the CANCEL button, you will receive the info window. There, you can find a detailed description of the error(s).
- If an error is *ignorable*, you may continue pressing the CONTINUE button; the result, however, may be incomplete. The Continue button is only shown if the errors made are ignorable.
- Accept the info window by pressing OK or press UNDO to get back to the previous input before starting the generator or expander. All input done before the latest generation will be cleared.

If you enter, for instance, a molecular formula that lacks some atoms or includes a wrong frequency, e.g. $C_8H_{15}O_2$, the message will be:

> *No isomers do exist with that molecular formula.*

3.11.2 Displaying the current error messages

You can always recall the current error messages produced by the generator or the expander by pressing the INFO button in either the generator or the expander section of the MOLGEN project window (see Figure 3.2 and Figure 3.16).

3.12 The structure editor MOLED

Probably not every substructure you need is included in the MOLGEN package. So you may create new substructures based on available ones, based on a generator result, or from scratch. This is done with the aid of MOLED .

3.12.1 How to start MOLED

MOLED can, of course, be started like MOLGEN , by clicking on the MOLED icon on your desktop. Every single substructure is saved in one file. So you can load and edit existing substructures easily. Apart from this, several entry points exist inside MOLGEN from where you can start MOLED . These points were described in previous sections, see Figures 3.9, 3.12, 3.23.

Having started MOLED , you obtain the following window. If you ran in from MOLGEN , a structure may already be loaded.

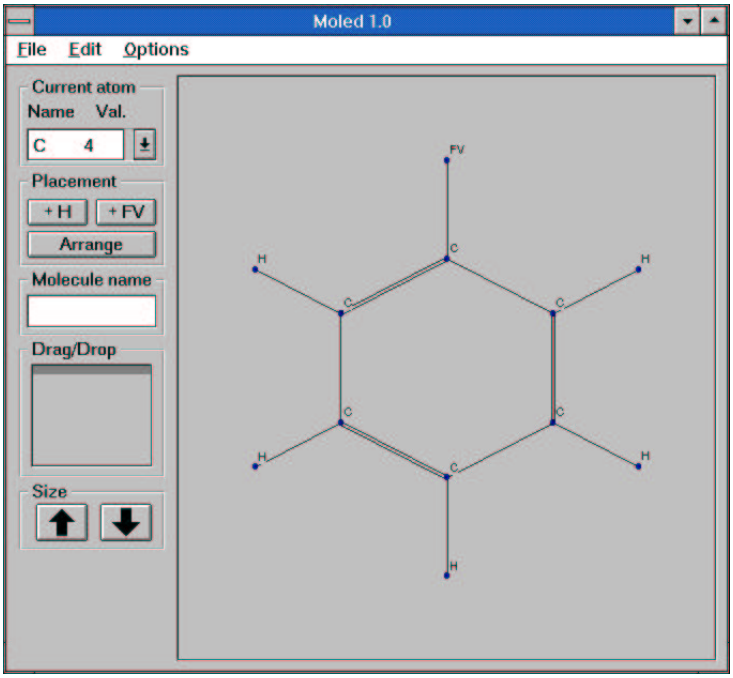


Figure 3.33 MOLED main window

The MOLED project window offers many possibilities to edit a structure.

3.12.2 Choosing an atom type

To choose an atom you want to draw, use the CURRENT ATOM section on the left side of the window.

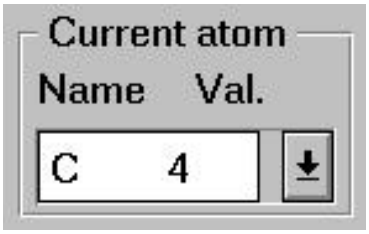


Figure 3.34 Current atom

There are several types of atoms preset in MOLED and by default, carbon is used. To choose an atom, click on the button on the right of the edit field. A listbox will occur and you can easily choose an atom type. If the required atom type is not included in the list, you can insert it by using the current atom field. Just write the atom symbol and its valence instead of those listed as current. Now you can use this atom and it is automatically added to the list of atoms.

3.12.3 Drawing a structure

To edit a structure, keep in mind the following use of the mouse. The left mouse button adds an atom to the structure. Now a *rubber band* follows the cursor and when you click the left button again somewhere else on the input window, a second atom of the current type is drawn, connected to the first one. At any time during the drawing, the current atom type can be changed.

You can make the rubber band disappear by clicking on its origin with the left button. Setting another atom now will not connect the previous to the new one. Moving around with the mouse will show you three special types of cursors.

- The first one is shown when the cursor is above an atom. In this mode, you can perform several actions on that particular atom.

- Clicking the left mouse button will start a new rubber band from this atom or connect it with the last one you clicked on.

- Pressing the left button, holding it down and moving around will move the atom upon release of the button. While moving, the cursor will change into a second special type.

- The last possible action at that point is pressing the right mouse button above an atom. This will remove the atom including all “H” and “FV” atoms connected to it.

- The third special type of cursor will be displayed above a bond between two atoms. If you click the left mouse button, the bond degree will increase. Clicking on a single bond for example will make it a double bond, and so on. Of course, the maximum bond degree between two atoms depends on the valences of these atoms. Clicking the right button will decrease the bond degree.

3.12.4 Arranging and completing structures

The buttons in the PLACEMENT Section of the MOLED window offer an easy way to complete your molecule or to draw it with an automatic placement.

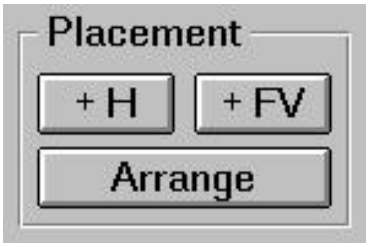


Figure 3.35 Structure placement

Pressing the +H button will add H atoms to all atoms where valences are not bonded. Pressing the +FV button will add free valences wherever possible. The ARRANGE button will redraw the complete structure using an automatic placement algorithm.

3.12.5 Changing the display size

Figure 3.36 Structure size

You can enlarge your structure by pressing the UP button and reduce it by pressing the DOWN button.

3.12.6 Naming the structure

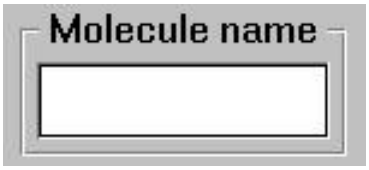


Figure 3.37 Structure internal name

Entering a name in the MOLECULE NAME section gives the structure an internal name saved with the structure. This is not the name of the file in which the structure is saved.

3.12.7 Copying the structure to the clipboard

Choosing the EDIT / COPY option on the main menu offers you two possibilities to copy the structure to the clipboard.

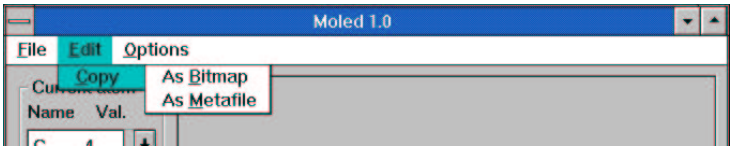


Figure 3.38 MOLED edit menu entries

Choosing the AS METAFILE option provides more opportunities to manipulate the picture in a vector graphics program such as CorelDraw.

Since some *Windows* applications do not understand this format, it is also possible to copy the structure AS BITMAP to use it for documentation purposes.

3.12.8 Display options

Using the OPTIONS menu entry provides several options helping you to edit the structure.

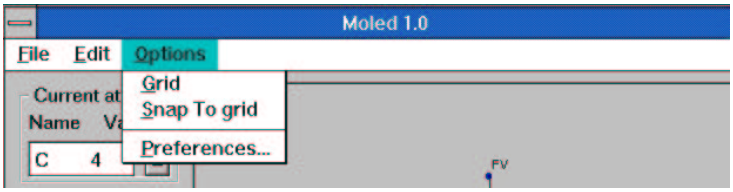


Figure 3.39 MOLED options menu entries

Using grids

GRID and SNAP TO GRID are checkable menu entries. Choosing the GRID option will add a grid to the edit window allowing to draw molecules "nicer". Atoms are not automatically set on grid points. To do this, enable the SNAP TO GRID feature. The GRID option need not be enabled in order to use the snap to the grid feature.

Changing MOLED preferences

Clicking on the PREFERENCES entry opens the following dialog:

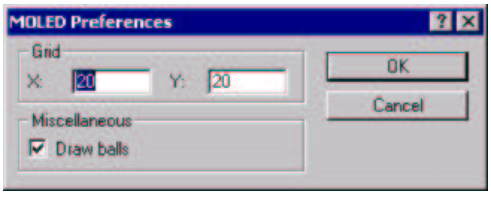


Figure 3.40 MOLED preferences

Changing the values in the GRID fields changes the amount of dots displayed in the current edit window if you enabled the GRID or SNAP TO GRID options.

The MISCELLANEOUS section allows to change the way the atoms are displayed in the window. If you do not like the balls representing each atom, you can disable them with the DRAW BALLS checkbox.

3.12.9 Printing a structure

To print a structure, select the FILE / PRINT... option from the main menu. The following dialog will appear:

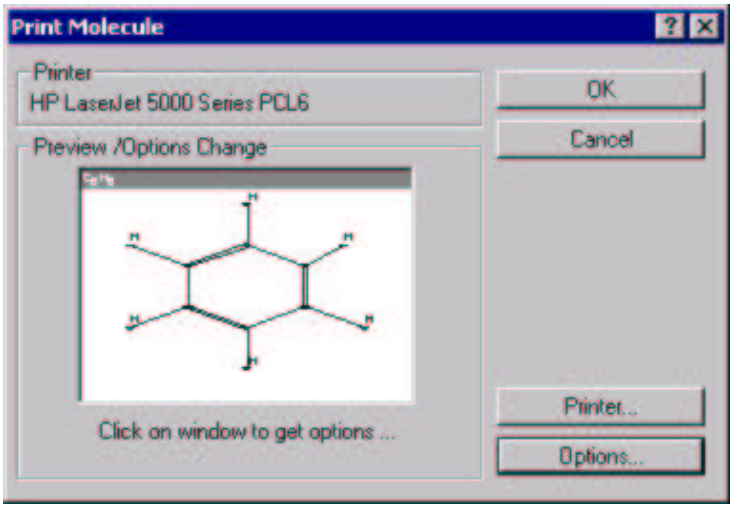


Figure 3.41 MOLED print dialog

The PRINTER section shows the name of the current system printer. The printer can be changed by using the PRINTER... button. Printer-dependent options can be entered using the OPTIONS... dialog.

The PREVIEW window inside the dialog shows the structure the way it will be printed. To change its look, click inside the window and the following popup menu will appear.

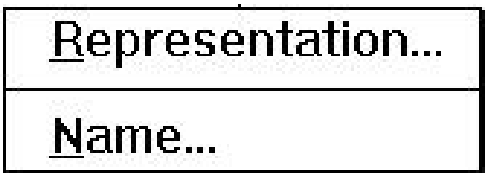


Figure 3.42 Printing options

The REPRESENTATION... entry opens a dialog known from MOLGEN (see Figure 3.12). It can be used to change several options for displaying and printing the structure.

Using the NAME... entry opens another dialog where you can enter a new name for the structure. This name is *not* saved with the structure (compare Figure 3.43). It is only used for the printed structure.

Figure 3.43 Structure name for printing

3.12.10 How to drag and drop a structure

The MOLED main window has an extra window that provides Drag and Drop possibilities for every structure.

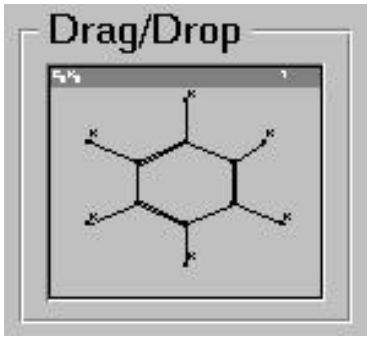


Figure 3.44 Drag & Drop window

Dragging a structure from MOLED to MOLGEN , for example into the Goodlist, is possible after saving the structure. If you click on the DRAG/DROP window and try to move the cursor away while holding the button down, a dialog will be displayed giving you the opportunity to save the structure now. In addition, the current structure must be connected and all valences must be used, either for bonds or as explicit free valences (FV).

If all these conditions are met, the structure can be dragged out of the DRAG/DROP window. Since normal filenames are used, you can also drop a file you have chosen from the file manager over this window. If the file is a valid MOLED substructure, it is inserted as if loaded using the EDIT/OPEN dialog.

If you are currently editing another structure, you will be asked whether to save it or not before using the dropped structure.

3.13 3D placement and stereoisomerism using MOLVIEW

MOLVIEW is able to calculate and display a 3D representation of a structure. For the spatial placement a potential energy is computed from the bond lengths, bond angles, torsion angles and non-bonded distances, and it is minimized numerically. But please note

that the computation of 3D-coordinates is time-consuming. The time required increases exponentially with the number of atoms.

3.13.1 How to start MOLVIEW

MOLVIEW can of course be started as MOLGEN , by clicking on the MOLVIEW icon on your desktop. You can then load, optimize and display existing 3D structures. Apart from this, several entry points exist inside MOLGEN from where you can start MOLVIEW . These points are mentioned in the previous sections when describing Figures 3.12 and 3.23.

After calling MOLVIEW , the following window will appear. If you started it from MOLGEN , the current structure might be optimized before you can access and modify it, but the current shape of the molecule is permanently displayed during the optimization process:

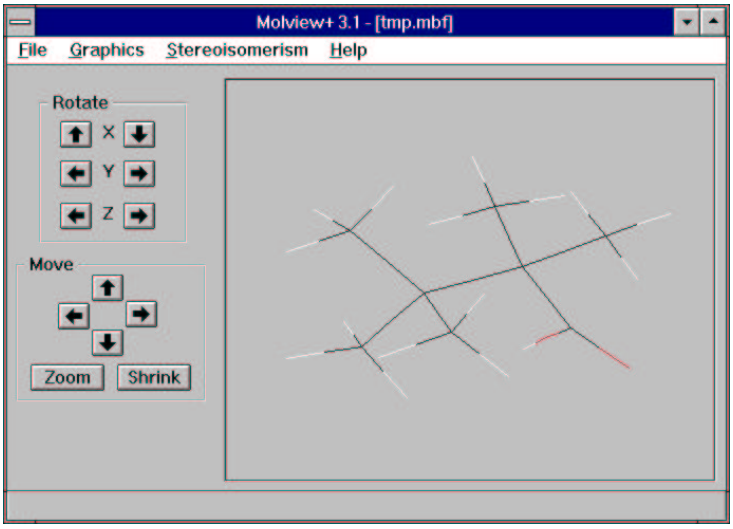


Figure 3.45 MOLVIEW main window

3.13.2 Using MOLVIEW

The MOLVIEW window (see Figure 3.45) is separated into two major areas. The largest space is reserved for the 3D window for displaying a structure. On the left hand side, there are several buttons for rotating and moving the structure. At the top of the window there is the menu bar for changing display parameters and for calling further functions.

3.13.3 Rotating and moving a molecule

The following features are available:

- On the left-hand side there is a field named ROTATE. With its six buttons the structure can be rotated about the axes X,Y,Z.

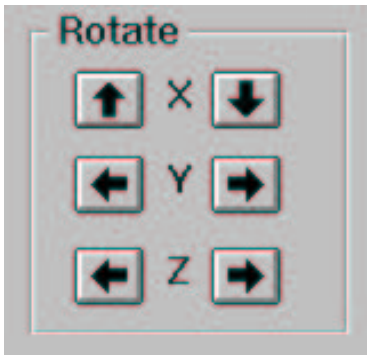


Figure 3.46 Rotation buttons

- Moving the molecule along the axes is also possible. This is done by the six buttons in the MOVE field. The buttons ZOOM and SHRINK perform a motion along the z-axis, thus acting like an enlargement and a reduction, respectively.

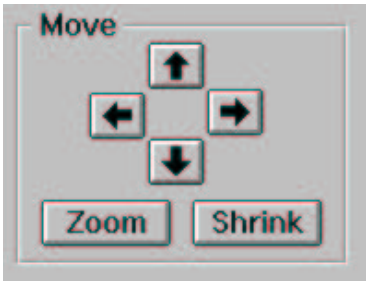


Figure 3.47 Move and size buttons

To ensure a smooth display in rotation and motions approximately 10 to 15 pictures per second must be computed and drawn.

3.13.4 Changing the drawing mode

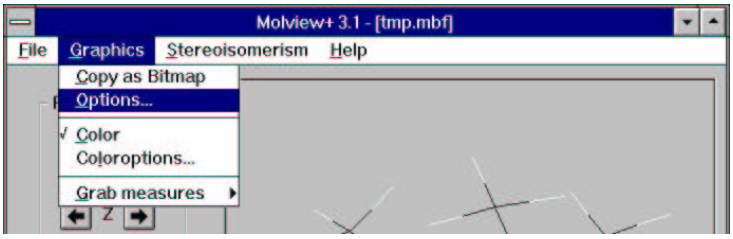


Figure 3.48 MOLVIEW main menu graphics entry

If you select the menu item GRAPHICS and click at OPTIONS... the following dialog window pops up for changing the display mode settings.

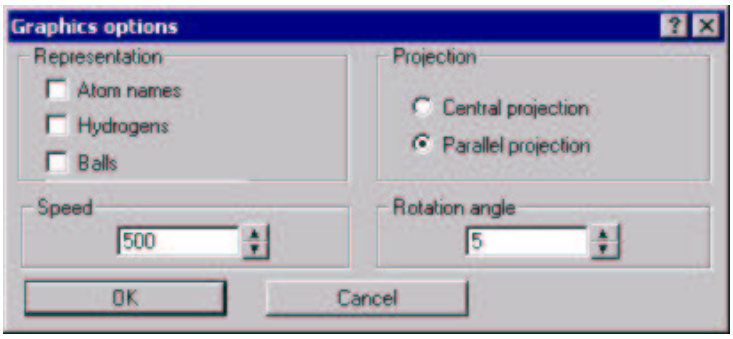


Figure 3.49 Molecule graphics options

The REPRESENTATION section allows to switch display features on and off:

- The ATOM NAMES button switches on and off the display of element symbols.
- The HYDROGENS button turns on and off the display of hydrogen atoms.
- The BALLS button decides whether or not atoms are drawn as balls.

These features are enabled immediately, not just after closing the box.

In the PROJECTION section the display in PARALLEL PROJECTION and in CENTRAL PROJECTION can be interchanged. In central projection distances are decreased for high z-coordinates, so as to simulate a perspective view.

Finally there is the option to change the SPEED of rotating and moving. The spin buttons and keyboard input in the numerical field (see Subsection 3.3.2) alter the speed in the range from 1 (minimum) to 500 (maximum). This value may be decreased if the operating system does not support the refresh cycles of MOLVIEW . Another way to manipulate the speed of the display is to change the ROTATION ANGLE. The displayed value stands for the degree of the angle that is used for rotations. This means, with the rotation angle set to 5 and a click on the X button, the molecule will be rotated by an angle of 5° about the x-axis. Setting the angle to 180° allows to switch between the front and the rear view by a single mouse click.

Leaving the window by OK keeps all the changes; selecting CANCEL will discard them. All selections are saved for later sessions.

3.13.5 Adjusting colors

The entry COLOR in the GRAPHICS menu (see Figure 3.48) is checkable. Disabling it makes all lines and balls appear in the same (default) color.

The background color as well as various colors for representing the elements as lines or balls may be adjusted. A corresponding dialog window can be obtained by the COLORS... item.

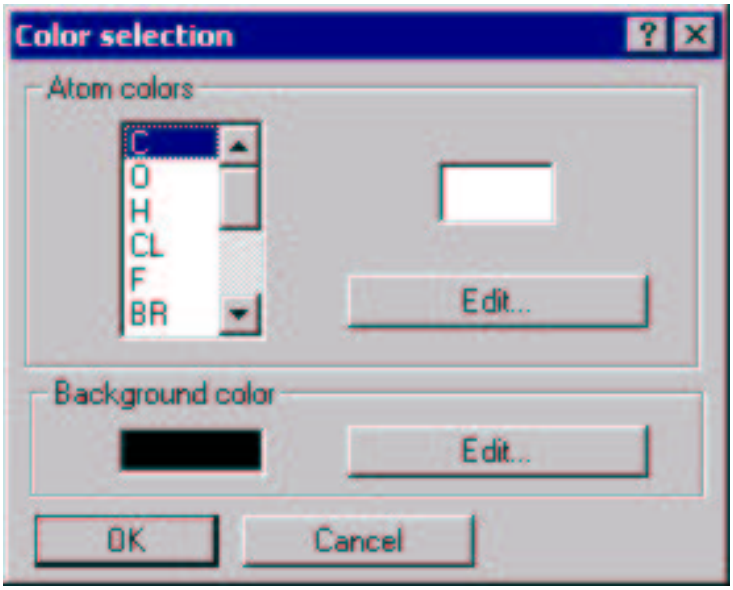


Figure 3.50 Color option dialog

The listbox on the left side contains all elements known to the program. If you select one, its display color will be shown in the adjacent window. To modify the color press the EDIT... button or double-click on the element in the listbox. For modification of the background color click EDIT... in the corresponding section.

Now the following *Windows* standard dialog appears. Please refer to the *Windows* manual for its usage.

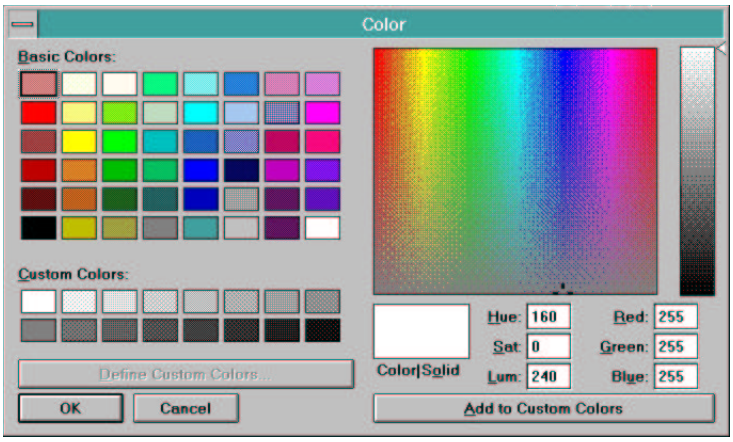


Figure 3.51 Standard color dialog

Your input will be saved on OK for later MOLVIEW sessions.

3.13.6 Copy as bitmap

This menu item allows the current drawing in the display window to be copied to the clipboard for printing and documentation, just as in the MOLGEN preview popup (see Subsection 3.4.1).

3.13.7 File operations

The basic file operations are accessible via the FILE menu:

Opening an existing file

Choosing one of the OPEN items the *Windows* standard open dialog appears for selecting a file name. After confirming with OK the desired file will be loaded and displayed. Only files in the MOLGEN binary format (MBF) can be read. Two alternatives are offered:

- OPEN FOR DISPLAY... The coordinates are taken from the file. The optimizer is only invoked if the file does not contain 3D coordinates.
- OPEN FOR OPTIMIZATION... Here the file is read, but the coordinates are ignored; the optimizer is started anyway. This method takes much more time, so it is only recommended if a new conformation calculation is desired.

Saving the current structure

You may store the current structure by SAVE AS.... A file name is selected in the usual *Windows* dialog window. Rotations performed since loading are also taken into account. As of stereoisomers, the current isomer is saved.

There are several file formats that can be used for saving:

- *The* MOLGEN *binary format (MBF)*: This is the default format which is also the only one that can be read.
- The *Brookhaven protein database (PDB) format*: This ASCII-format can be handled by a number of chemistry software programs. Since MOLVIEW covers no information about proteins, all atoms are saved as heteroatoms.
- *The MDL MolFile format*: In contrast to the generator (see section 4.9.1) MOLVIEW writes only one single structure to the file, containing the coordinates of all non-hydrogen atoms.
- *The MolPic MP3-format*: This is an easy ASCII format that is used by the educational software MolPic which is able to display 3D structures in a larger variety than possible in MOLVIEW . Moreover this format is so easy to read that it provides a good starting point if you want to convert the output to your own inhouse format.

Except in the case of the MBF format all export filters pop up a dialog asking for scaling of coordinates.

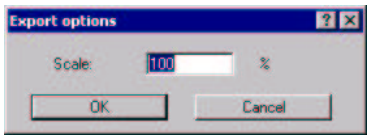


Figure 3.52 Scaling dialog

Coordinates will be enlarged or reduced by the given percentage before they are written to a file. This simplifies the display of small structures in programs that are designed to handle very large ones such as RASMOL. This option does not affect the display in MOLVIEW .

Reoptimizing the current placement

The placement computed by the optimizer might not fulfill your expectations. The item OPTIMIZE allows to update the CURRENT COORDINATES by another call of the energy calculation. Either to re-use the current coordinates which will just refine the current conformation, or select RANDOM COORDINATES to compute a new conformer from scratch.

MOLVIEW shows the complete structure in its current coordinates during the entire optimization process. So the progress is visualized, and you may decide whether the current spatial arrangement fulfills your needs and stop the process.

Loading and fitting a second structure

For comparison and evaluation a second structure can be added. Note that this second file must already contain coordinates.

After calling ADD SECOND... from the FILE menu and selecting a filename from the corresponding dialog the second structure will be displayed together with the first. For converting the file-internal coordinates the scaling of the first one is taken.

Now some menu items are no longer accessible. Menu operations such as stereoisomer calculations etc. now only refer to the first molecule, while the other one remains unchanged. Rotations and translations, however, act on both, as do the graphics settings (see Subsection 3.13.4).

For a better comparison of the two molecules they are displayed in uniform colors, green and red. To abandon the second structure, click on REMOVE SECOND in the FILE menu.

The menu item FILE - FIT SECOND which is now enabled starts an algorithm for fitting the second structure to the first one. Each atom is compared to its counterpart and Euclidean distances between corresponding atoms are minimized. The root mean square of the deviations is displayed.

Note that this function is still weak in that the structures are compared only according to their initial numbering, i.e. atom 1 with atom 1, atom 2 with atom 2 etc., and distinct conformational properties are neglected. For two conformers in particular the differences of the conformations are well visualized by this operation.

Drag & Drop function

MOLVIEW provides Drag & Drop functionality. You may drag a structure in MBF file format out of MOLED or the file manager and drop it in the MOLVIEW display window. An optimization will only be invoked if necessary.

3.13.8 Geometric information

The displayed molecule is stored internally in 3D coordinates. So distances, angles etc. are easily computed. A user interface is provided by the menu item GRAPHICS - GRAB MEASURES. The submenu determines the measure type: DISTANCE, ANGLE or TORSION ANGLE. After selecting one of these, click on an atom, it will be marked by a circle.

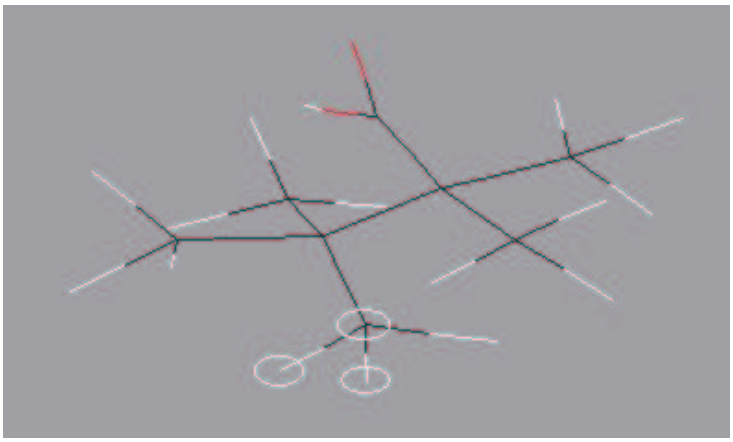


Figure 3.53 Selected atoms

For a distance select 2 atoms, for an angle 3 and for a torsion angle 4 atoms are required. After marking the last one, an info window with the result appears automatically.

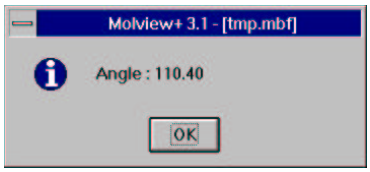


Figure 3.54 Geometrical info

Distances are given in Å (10^{-10} m), angles and torsion angles in degrees.

3.13.9 The stereoisomer generator

Apart from molecular formula and connectivity there are further ways to distinguish molecules depending on the spatial locations of their atoms. The decisive notion is that of *configuration* which reflects the spatial arrangement of atoms. Molecules interconverted by intramolecular rotations will not be distinguished. MOLVIEW is able to calculate for a given structure *all configurational isomers* and their spatial realizations. (The notion of stereoisomerism is not used unambiguously in the literature. In the following stereoisomerism always means configurational isomerism.)

Supported effects

Basically the following effects are taken into account:

- *Stereocenters of four-coordination*: If a tetrahedral molecule contains a tetravalent atom with four different ligands, two non-superimposable spatial arrangements are possible which are mirror images of each other and are called *enantiomers*. An object nonidentical to its mirror image is called *chiral*.

  If there are more than one such so-called "chiral centers" present in the molecule, chiral and/or achiral stereoisomers may result depending on the symmetry of the molecule.

- *pi-Diastereoisomerism*: A structure with one double bond carrying different ligands on both ends exists as two diastereomers forms (not mirror images).

  If there are more than one such double bonds present (isolated or conjugated), each of them independently gives rise to this type of isomerism.

- *Exo-methylene and spiro compounds*: Chiral exo-methylenecycloalkanes and chiral spiro compounds are also recognized.

Construction of stereoisomers

Please note:

> *Configurational isomers are generated exclusively. Stereoisomers that differ by rotations around single bonds, so-called conformational isomers (conformers), are not taken into consideration. Information about the 3D coordinates of the molecule is not used in stereoisomer generation.*

However, for displaying stereoisomers 3D realizations are constructed geometrically. These are based on reference coordinates computed by the optimizer. An example is given by the following four stereoisomers of 1,2,3,4-tetramethylcyclobutane.

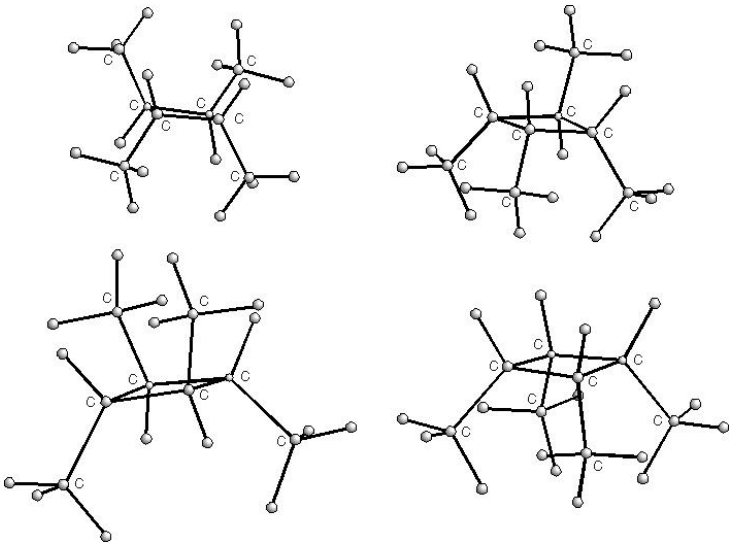


Figure 3.55 1,2,3,4-Tetramethylcyclobutane

There are limits of the geometrical construction as well:

> *If a molecule contains stereocenters belonging to more than one ring, where the rings have two or more atoms in common, or if it contains a stereo-relevant double bond in a ring, no exact construction will be possible. In these cases approximate placements are computed showing the configurations, but including some "strange" bond distances or angles.*

Such a construction for cyclohexene is depicted in the next Figure:

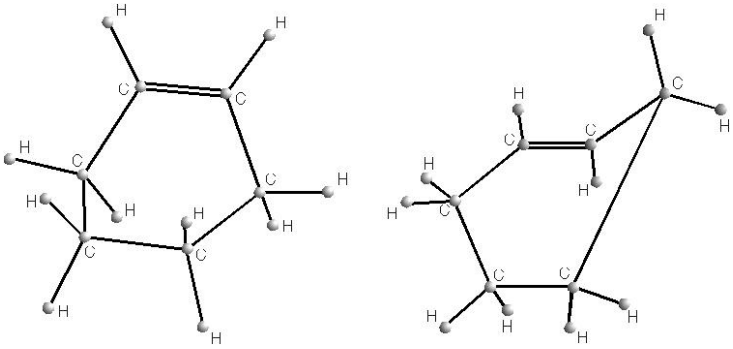


Figure 3.56 Cyclohexene

The usage is very simple.

Generating all stereoisomers

- Call MOLVIEW with the placement of a structure in whose stereoisomers you are interested.
- Select the menu items STEREOISOMERISM and CALCULATE ISOMERS.

A message is given if the structure can occur in one stereoisomeric form only. Otherwise the number of generated isomers is given. If no exact construction is possible (see above), a warning will be displayed. On the left side of the screen a new field STEREOISOMERS is displayed. The number shown there is the number of the isomer currently displayed. You may switch to the previous and to the following isomer using the spinbuttons aside of the field.

The display of stereoisomers can also be varied.

Changing the display of stereoisomers

In the STEREOISOMERISM menu there are two ways of changing the display:

- After the generation, stereocenters are marked by red balls. This feature may be switched off and on by HIGHLIGHT CENTERS.
- Besides the geometrical display also the output in terms of R/S–nomenclature after Cahn/Ingold/Prelog and — for double bonds — in cis/trans-classification is available. It may be activated by R/S-DESCRIPTORS.

If an isomer has not been constructed correctly, you may re-apply the energy optimization. In general, the generated configuration is kept, only the distortions are fixed.

After reoptimizing via FILE - OPTIMIZE you will get the MOLVIEW window again for the optimized stereoisomer. Please note that the new coordinates are the reference for *all* constructions; a previously undistorted isomer might be distorted now.

The constructed or optimized placement of the current isomer may be saved by FILE - SAVE AS... .

The introduction to MOLGEN is finished now. In further chapters you will find more detailed information about special cases and the background of the methods used in this program.

Chapter 4

Strategies and peculiarities

4.1 The substructure concept

After getting accustomed to using the functions of MOLGEN we now want to have a closer look at the methods of the program. Some basic strategies will be presented that are useful for working with MOLGEN . This chapter is intended to enable you to use the program efficiently. The main aim is to make you familiar with the interplay of structure generator and structure expander, the corresponding options and special features, as well as some tricks. We will illustrate the explanations by examples, where we shall only mention input data and results. Necessary steps in between may be looked up in Chapter 3.

4.1.1 Why not expand immediately?

Perhaps you wondered why we separated generator and expander in the structure calculation with macroatoms in MOLGEN . It would have been possible to run both steps simultaneously by default (which is optionally possible, as we saw). The following idea, however, pleads for the concept implemented here: Due to mathematical and program technical reasons, expansion of macroatoms is the most time and memory consuming part in the computation of isomers. Therefore we have included the possibility to decrease the number of candidates before expansion.

In particular, during expansion all isomers are kept in the computer's main memory because a test on identity has to be performed – this establishes a natural limit anyway.

If the generator produces connectivity isomers using macroatoms, the result is a set of candidates of reduced molecular formula consisting of the macroatoms represented by single nodes of the molecular graph and the remaining atoms. Quite often already at this stage candidates can be canceled, for example, if the remaining molecule part comprises substructures or ring sizes to be excluded. This strategy is especially feasible since a

macroatom as a whole need not be bonded to other atoms (or macroatoms) in a strictly determined manner. However, the candidates will have a different structure after the expansion of macroatoms. More details are found in Subsection 4.1.3. We illustrate this by giving three examples:

Example 1

Consider the molecular formula $C_7H_6O_2$ together with the well-known macroatom COOH that has a simple structure and only one free valence. So the appearance of the reduced candidates comes quite close to the desired structures.

- *Generate with the use of* COOH, the output should consist of 685 structures. (For the sake of completeness we mention that there is a total of 122,391 isomers.)

This time we are interested in results without rings of size 3 or 4 only. Since COOH is univalent, it cannot occur within a ring. Additionally, we know that the ring of maximal size can be formed by the remaining six carbon atoms only. Consequently, we restrict the ring size to 5 to 6.

- *Enter ring sizes* 5 to 6 at Cycle Sizes. - The generation will then provide only 96 resulting structures.

Taking a look at these solutions the large number of triple bonds is suspicious, which we now also want to exclude. Here we use the fact that COOH is *singly* connected to the rest of the molecule.

- Enter at Bond degree the maximal bond degree 2. Another run of the generator yields 38 solutions.

If, for example, a macroatom bears three free valences, it is not as easy to make use of a restriction of that kind for the generator, because a triple bond may "disappear" *after* expanding. For verification we show you the numbers 1, 12 and 33 out of the 38 solutions. If everything is correct, the following structures should appear in your Result Window at the given numbers:

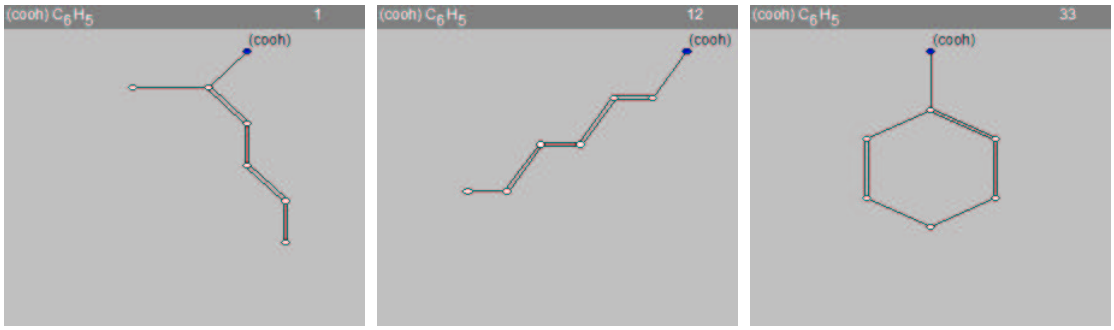


Figure 4.1 Three isomers from Example 1

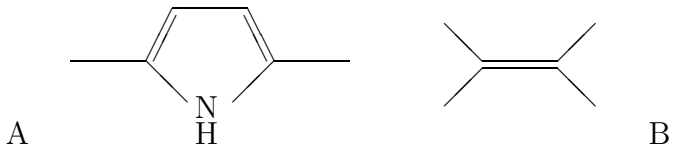
After expanding COOH there are, of course, still 38 structures.

To obtain the same result in this example, you could have expanded COOH in the 685 initial solutions *immediately* and entered the described restrictions *in the input menu of the expander* — try this for verification! In the latter case the additional restrictions are only considered in the expanded molecules. The difference in computation time here is certainly small, such restrictions may, however, be remarkably efficient if large macroatoms are used.

This is a simple standard case that you often might come across. For univalent macroatoms this procedure is excellent. The next two problems are somewhat more complicated.

Example 2

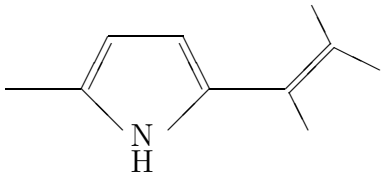
We study the formula $C_8H_{11}NO$ and choose the macroatoms



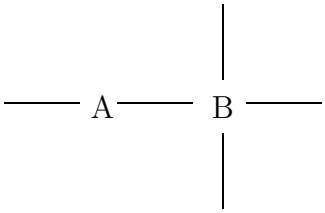
We will indicate them in short as A and B.

- *Create generator results using the given macroatoms*, there should be 51 solutions.

Among the structures we will allow only those that do *not* bear a vinyl group at the 5-membered ring, i.e. the following substructure shall *not* occur among the final structures:



Since A stands for the 5-membered ring and B for a vinyl group, in the generator results the occurrence of the following substructure A-B can already be excluded:



This substructure contains atom names not known to MOLGEN *. If you want to build them by the molecule editor, you will have to introduce the user-defined atoms A with valence 2 and B with valence 4. (For details see Subsection 3.12.2.)*

- Create generator results using the macroatoms, and A-B in the badlist — there are now only 31 solutions.

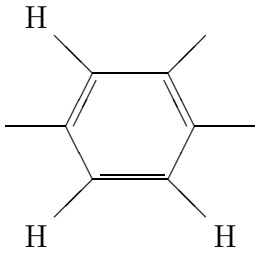
After expansion and testing for isomorphism you will receive 45 structures. In this case, you could also have used the badlist *after* expanding, but A-B then of course in *expanded* form, i.e. as shown in the last but one drawing.

Please note that in this case no *further* fragments among the 45 solutions can occur that consist of a 5-membered ring of type A and a C-C double bond of type B. That is at least not obvious, since a C-C double bond could, for example, be built by the remaining atoms which do not belong to A or B. In general, that is for other molecular formulae, the expanded badlist structure should be entered once more for expansion to cover all possibilities. This shall, however, not worry you:

> The badlist in the generator has just the purpose to keep the number of solutions to be expanded as low as possible and it is often an efficient **pre-filter**. The fewer expanded solutions the faster they may then be **post-filtered**.

Example 3

Finally we consider the formula C_8H_7NO and select a benzene ring with three free valences as macroatom.



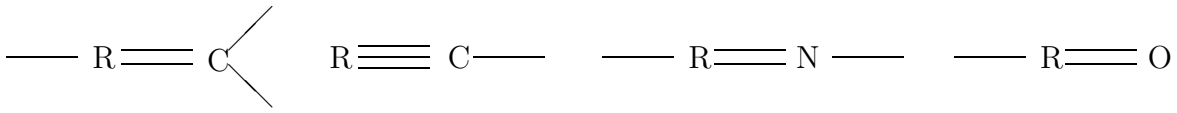
We will denote this macroatom by R. This structure may easily be created by the molecule editor. We are interested in all possibilities to attach the benzene ring to the remaining atoms in a way that no further rings arise.

- *Generate all solutions using* R — the result consists of 333 structures.

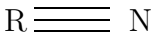
The following argument will lead to the desired structures: If further rings must not occur, the generated solutions already have to be ring-free.

- *Select additionally at* PROPERTIES *trees only.* — The corresponding generation yields 117 solutions.

A superficial consideration would be finished now, but there is still some subtlety to be taken into account: Since the generator produces structures with the artificial atom name R, and this R bears three free valences, it might happen that R is doubly or triply bonded. *After* expansion such bonds may become rings, e.g. a 3-membered ring out of R=C if the vicinal atoms in R are connected to that carbon atom. So we will forbid that R is doubly or triply bonded. This is carried out by creating the following four structures by the molecule editor and including them into the badlist:



These are in fact *all* possibilities for R to be attached to the remaining atoms by a double or triple bond. For example, the case



need not be considered, because an isomer with this substructure would be *disconnected*.

- *So enter these structures in the badlist.* A new run of the generator gives only 44 solutions.

Here numbers 1 and 8 are shown:

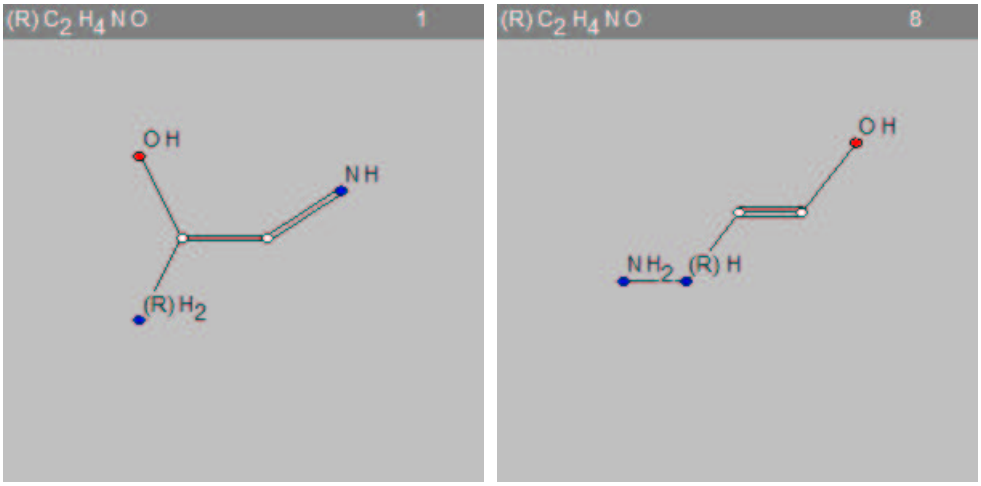


Figure 4.2 Two isomers from Example 3

After expansion of the benzene ring there are 105 isomers that are ring-free except the benzene ring. Of these, no more than 88 will survive elimination of aromatic doublettes. So here restrictions were realized *before* expansion that would have been very hard to fulfil afterwards.

Some further remarks:

1. Overlooking subtleties as just described is no major problem anyway. You always have the possibility to grope your way *step by step*. After first starting the computation of the generator you will encounter more and more badlist-structures among the solutions, that could be used in the generating process already. For a very large set of resulting structures you may also abort the generation in order to save time and to look at the solutions computed so far.

2. Besides this, you should consider the quite simple method to edit the resulting structures from the 2D-display in order to build new badlist entries immediately! (See again Section 3.9.)

These examples show only *some* of the various possibilities of an efficient combination of generator and expander. It is impossible here to deal with all the possibilities, but perhaps you were encouraged to go on experimenting on your own.

4.1.2 Dummy atoms

Perhaps it was difficult to compile the badlist in the last example in Subsection 4.1.1. Therefore there is a feature in MOLGEN that allows to simplify conditions of that kind. If only the *type* of the bond(s) counts and not the *nature* of the atoms, there is the option to define a *variable* atom, a so-called *dummy atom* for this restriction. These atoms get the reserved name "X". (Note that X does not stand for a halogen as sometimes in the chemical literature.)

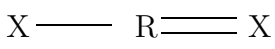
In the last example we wanted to exclude the case that R is doubly or triply bond. Here we did not care about the type of the atom to which R is connected; we only did not want any double or triple bonds to occur. We formulated this by entering *all* possible bonds. With the use of dummy atoms the badlist can be abbreviated like this:

In the first group X stands for an *arbitrary* atom with valence greater than or equal to 3, in the second case for an arbitrary atom with valence greater than or equal to 2. Thus, atom type and valence of a dummy atom are variable.

Please observe the following:

1. Dummy atoms are allowed only in goodlist or badlist structures, but not in the molecular formula or in macroatoms. (MOLGEN will otherwise give an error message.)

2. Dummy atoms must have at least valence 2. For other purposes you can simply use a free valence. Consider the second badlist structure of the example above: The variable single bond was defined by a free valence. The structure



   however, would not yield the desired result.

3. When using substructures with more than one dummy atom you should be aware that each dummy atom may stand for a different atom type, although they all are denoted by X.

4.1.3 Reduced and expanded molecular formulae

In this section a summary of the consequences of the use of macroatoms will be given. This kind of substructure has some special qualities, some of which have already been discussed in chapter 4 and in Subsection 4.1.1.

> *If your input data contain a macroatom, it will be treated like a single atom, i.e. all its atoms are* ***subtracted*** *from the molecular formula and* ***replaced*** *by the symbol of the macroatom. The valence of that "artificial" atom is then equivalent to the number of free valences of the macroatom. This yields a so-called* ***reduced molecular formula****.*

In the example of chapter 2 the use of the macroatom COOH in $C_8H_{16}O_2$ results in the reduced molecular formula

$$C_7H_{15}(COOH)_1.$$

For sake of clarity the macroatom is embraced by parentheses. This is also the way the formula appears in the 2D display of generator results. The substructure COOH has one free valence, so it is treated as an atom of valence 1 by the generator. Note that the naming of the macroatom makes no difference. We could also have labeled the substructure by "M". In this case the reduced molecular formula would have been $C_7H_{15}(M)_1$. The name merely indicates under which *filename* this substructure is saved — you are free to choose it by your requirements. If a macroatom is entered *repeatedly*,

the formula will be analogously reduced *repeatedly*. For instance, define a macroatom: Call it "COC" and require it twice, the formula will be reduced to

$$C_4H_{16}(COC)_2.$$

The macroatom COC has valence 6. In the same way, several *different* macroatoms can be subtracted from a molecular formula. Errors can arise for two reasons:

- Reducing is impossible since the molecular formula does not contain the required number of atoms of a given type. (In the previous example COC cannot occur three times.)
- Reducing is possible, but it is impossible (for principal reasons) that structures with a macroatom of that kind exist. (Try, for example, $C_8H_{18}O_2$ with COOH.)

These errors are detected by the structure generator, and you will receive corresponding messages W15 and W29, which are explained in detail in Subsection 6.1.1. Since you can use input other than macroatoms for the generator, note the following:

> *All goodlist and badlist structures as well as further restrictions, which you declare at the* ***generator*** *in addition to macroatoms, always refer to the* ***reduced*** *molecular formula.*

So in isomers of $C_8H_{16}O_2$ there may occur rings of sizes 8 to 10, but this cannot happen in $C_7H_{15}(COOH)_1$. In this case it is also not possible to require a badlist structure that contains an O atom, since the reduced molecular formula does not contain any oxygen at all. In both cases you have to expand COOH before applying the restrictions. The latter problems are also detected by the generator, which sends corresponding error messages. A useful hint is:

> *The easiest way to get an overview over the reduced formula and the generator results is to start the generator first with the macroatoms alone (i.e. without goodlist or badlist) and to look at the 2D display. Afterwards you may repeat the generator run with reasonable additional restrictions.*

In the examples in Subsection 4.1.1 the intelligent combination of macroatoms and other input for the generator was explained in detail.

The expander undoes the reduction of the molecular formula in the expansion of the macroatoms, i.e. it "expands" the molecular formula. The rule formulated above reads now:

> *All goodlist and badlist structures as well as further restrictions, which you declare at the* ***expander*** *in addition to macroatoms always refer to the* ***expanded*** *molecular formula.*

Also in this case you may approach the final result step by step with the aid of the 2D display.

4.2 Using macroatoms

4.2.1 Macroatoms or goodlist?

Perhaps the *distinction between macroatoms and goodlist* made in MOLGEN is still dubious to you. The following section is therefore devoted to a more precise description of the effects of these substructures.

Example 1

Let us once more consider the formula $C_8H_{16}O_2$ from Chapter 1. There are exactly 39 isomers out of 13,190 that contain a carboxyl group COOH. We carried out the necessary computations with the aid of a macroatom. Now we want to achieve the same result by using a goodlist.

- *Insert* "COOH 1" *in the goodlist in the Generator prescription window and start the generator.*

For quite some time the generator is seemingly silent — nothing happens on the screen. But then we get at last the message that 39 isomers have been constructed. During the delay the generator computed all 13,190 isomers and searched through each of them for a COOH group. The molecules obtained this way are exactly those we got in Chapter 1. So in this case the goodlist makes no sense, the calculation is much more expensive. In general we can state:

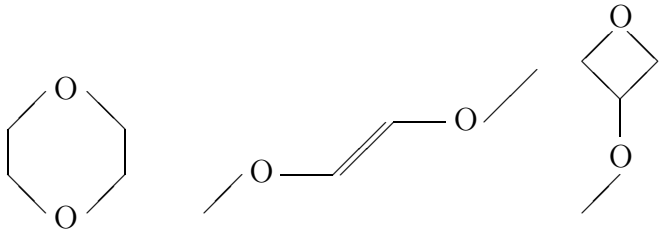
> *Using a substructure as a macroatom only once you will get exactly the same solutions after expansion as if you had put this structure in the goodlist.*

In the latter case there is no expansion necessary, of course.

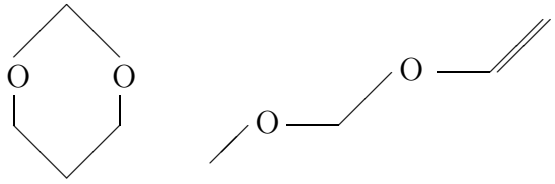
Things get more difficult when substructures are used multiply or when different substructures have to be taken into account. Let us take the molecular formula $C_4H_8O_2$ and COC as macroatom. We want to carry out two calculations with that formula and

- COC *as a macroatom required twice with expansion afterwards*,
- COC *as a goodlist entry required twice.*

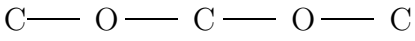
The first run yields the three isomers:



In the second case we get 12 isomers among which we find the 3 from above. Further solutions are for instance:



The difference comes from the fact that *the goodlist filter allows overlap* of substructures. In the case of our two COC groups the possible overlapping is obvious.



Note that no structures containing an oxiran ring are generated here. This is due to CLOSED SUBSTRUCTURES being active (see Section 3.6).

Macroatoms, however, are regarded as *separate structures from the very beginning, they must not overlap.* The general rule reads:

> *Using multiple substructures or different substructures in the goodlist, there are at least the same solutions as computed with the use of macroatoms, and possibly more.*

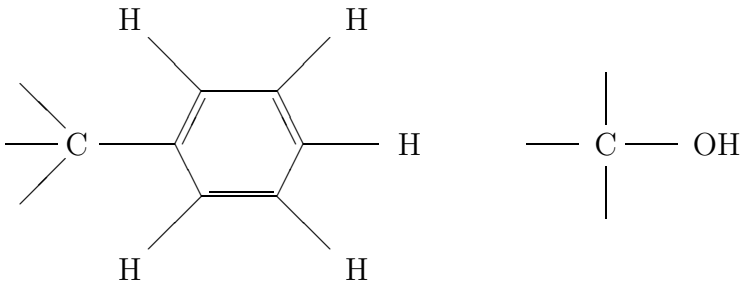
However, previous examples show that there need not always be more solutions; note that the two CH_3 groups cannot overlap. In general overlappings in large substructures may become very complex and hardly manageable.

If you want to use prescribed substructures, you have to make sure which of those occur without overlapping before the computation of isomers (you may enter them confidently as macroatoms) and which do not. Often a good remedy is to delete a few atoms from macroatom structures, if these may also occur in other substructures. To find the best solution in such cases requires some experience and combinatorial skills.

At the end of this section we want to demonstrate how to combine goodlist and macroatoms in a simple case by deleting a single atom.

Example 2

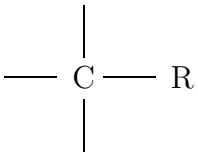
The formula shall be $C_{10}H_{12}O$ and the substructures



These substructures have to occur in the isomers – possibly overlapping. Obviously, they can only overlap in the carbon atom with three free valences. Of course you are free to insert both in the goodlist and start the generator then:

> *The computer would have to search through 3,916,111 isomers for these two structures. It is absolutely capable of doing so, but this is obviously an inefficient procedure.*

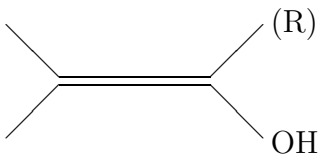
We can take another way: Since both structures can only overlap in the carbon atom with three free valences, we delete this C from the first structure and define the remainder (a benzene ring with one free valence, i.e. a phenyl group) as macroatom R. Besides, we need not throw away the knowledge about the whole structure; let us construct the additional group



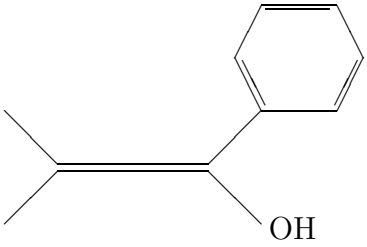
So after entering the molecular formula you have to follow these steps:

- *Input* R *as macroatom.*
- *Enter the* COH *and the* CR *structure in the goodlist.*

The option that both substructures overlap is given by them being goodlist entries. The generator is nevertheless able to compute the desired solutions quickly, due to the macroatom R. The result consists of 39 structures, the first of which shall be displayed:



After expansion there are still 39 isomers and the molecule above is expanded to



4.3 Special use of the expander

4.3.1 The expander as a filter

Up to now we used the expander mainly according to its name, i.e. for expanding macroatoms. There are, however, numerous other features offered by this program. As for the generator, you can use goodlist and badlist as well as restrictions.

> *There is the possibility to screen the isomers computed by the generator. Additionally the results of the expander can be re-entered for a new run as often as desired.*

Having achieved a tolerable number of structure candidates it is usually intended to reduce this set step by step until reaching an appropriate, manageable list of possibilities.

> *Note the following for the expander: If no macroatoms are expanded, no identity test (isomorphism test) is necessary.*

Thus we included the option to switch off that test. This way some problems with memory size also vanish; you can filter arbitrarily large sets of structures. To begin, here is a simple, classical example:

Example 1

We want to search through the isomers of C_6H_6. We use the generator to compute *all* 217 isomers and the expander to filter them several times. First we look for structures with rings of size 5 to 6:

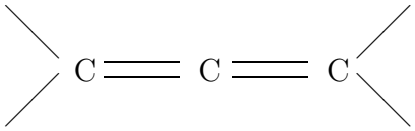
- *Enter in the* Cycle Sizes *section of the expander input window the interval 5 to 6, switch off the isomorphism test and start the expander*, this yields 31 structures.

Now we can *re-filter* these 31 structures, i.e. re-running the generator and the expander with additional restrictions is not necessary; we can simply search through the latest expander result. Next we want to exclude triple bonds.

- *Enter 2 as maximal bond degree at* Bond degree *and enable* Use Expansion Result *in the generator input window. You obtain 12 isomers.*

The message window tells us that only 31 structures were processed. Of course, we could have removed the restriction of ring sizes from above, since it was fulfilled anyway. At this point you also you may ignore the result. This is indicated by the UNDO button in the Info window. Pressing this button means that you return to the previous result. This is particularly important, if e.g. the expander provided 0 solutions due to wrong input, or if it had to stop before finishing because no more memory was available. In such cases you can continue with the old results and new input.

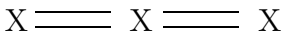
Finally we want to get rid of cumulenes, i.e. double bonds of the following form:



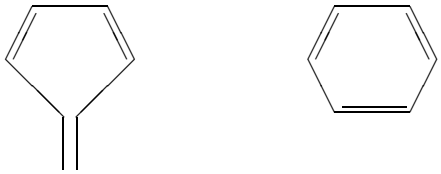
You can proceed as follows:

- *Enter this structure at* SUBSTRUCTURES *in the badlist, now 2 isomers remain.*

Note that we could also have defined as badlist entry a structure with dummy atoms, like this:



The two solutions are:



Did you expect that in the end only one structure, the benzene ring, would be left?

Remarks:

- In this example we also could have achieved the result step by step with the generator. There are often many different ways to reach a desired result with MOLGEN .
- In order not to become confused when filtering it is recommended to look at the result window of the filtered (expanded) structures as well as to use this display to get an overview over the current set of candidates.

Another example will illustrate the option to switch off the isomorphism test in order to manage a large number of structures by the expander.

Example 2

This time we take the formula $C_{11}H_{12}O$ and as macroatom the 1,2,4-trisubstituted benzene ring R as in Example 3 in Subsection 4.1.1. The generator produces 3,390 solutions for this problem, which have to be saved completely for the following computations. Please do the following:

- *Select* GET GENERATOR MACROATOMS *in the expander menu and start the expander*. Depending on the memory configuration of your computer the expander can calculate about 5,000-6,000 solutions until it has to stop due to lack of memory. The Info window also tells us that the reason is the isomorphism test. *Now select* UNDO.
- *Switch off the* isomorphism test and start the computation again. This yields 15,391 solutions that should all be saved. (The expander is now faster than before.)

In this case the set of solutions presumably contains duplicates, i.e. some of the saved isomers might occur *multiply*. Nevertheless we now have the chance to get a complete overview over all isomers by the result window and to sieve them by further restrictions. We perform this by the simple condition that rings must occur, and only rings of size 5 to 6.

- *Enter these conditions* into PROPERTIES.

- *Remove the macroatom R from the expander input in the* SUBSTRUCTURES *section.* (Otherwise at the next run a message appears that R cannot be expanded — since this has already been done.)

- *Select* use expansion result *and run the expander again.* Now MOLGEN reduces the 15,391 structures to 4,258.

Finally we make the expander reject the double entries.

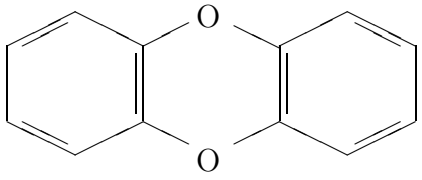
- *Select the isomorphism test* and start the computation again. Do not forget to also select use expansion result. Now, only the isomorphism test will be performed. This run reduces the set of candidates to 2,733 structures. Finally, after removal of aromatic doublettes 2,501 isomers will survive.

So we first reduced the set of solutions by imposing restrictions and carried out the memory intensive isomorphism test at the end. Also in this case this is only *one* possible way to obtain the desired structures. The last two runs, for instance, could have been performed together, sieving out the inappropriate rings *and* testing on isomorphism at the same time. Internally the restrictions are checked first and then the test is carried out, i.e. the only decisive aspect for the memory demand of the expander is the final number of structures. Switching off the isomorphism test is especially reasonable and recommended if it is not known before expansion which filtering conditions shall be used, and if a survey of the various solutions is required.

This was only a small part of the variety of ways how the expander can be used for filtering. You are once more invited to continue on your own.

Another example

At the end of this chapter we would like to demonstrate another interesting example to be processed by MOLGEN . As commonly known, so-called dioxins share the following core



and have molecular formulae of the form

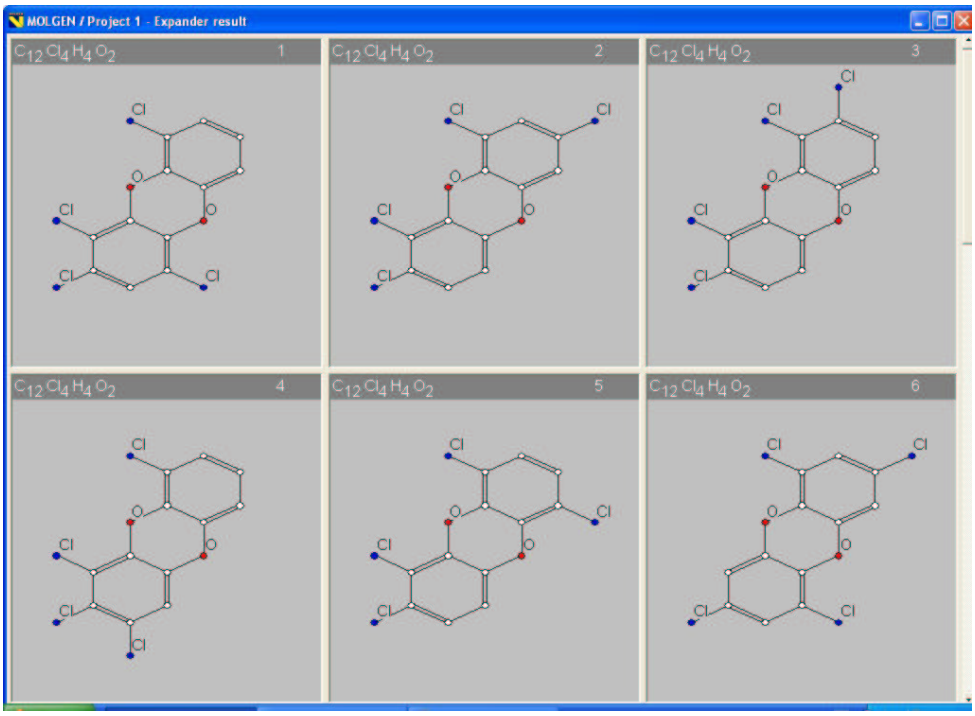
$$C_{12}O_2H_iCl_j.$$

The variables i and j represent numbers between 0 and 8 which must fulfill i+j=8. We now want to generate all dioxins for

$C_{12}O_2H_4Cl_4$.

With the molecular formula alone we shall not get very far, the generator would run for days or weeks or even longer, and it would produce billions of isomers and therefore sooner or later run out of storage space. All carbon and oxygen atoms are, however, located on three 6-membered rings.

This skeleton to be defined as a macroatom should be available in your installation of MOLGEN under the name **dioxin** (click Structures (in the Edit window) and then in Select substructures look up the lists ⟨Dir⟩). The aim of this example is to demonstrate the possibilities to distribute H and Cl atoms over the free skeleton positions. After entering the molecular formula and dioxin 1 as macroatom we start the generator. You may be puzzled that the generator produces only *one* solution. The free valences of the macroatom are indistinguishable to the generator; so an 8-valent atom with hydrogen and chlorine ligands is created, where all positions of the ligands are equivalent. After expansion this is no longer true: It is then, for example, essential to which of the two possible positions in a benzene ring a Cl atom is bonded. With the macroatom dioxin expanded there are now 22 distinct isomers, which describe the essentially different substitution patterns. Here are the numbers 1 to 6.



You may compute the solutions for other cases. Table 1 contains the results.

molecular formula	number of dioxins
$C_{12}O_2Cl_8$	1
$C_{12}O_2H_1Cl_7$	2
$C_{12}O_2H_2Cl_6$	10
$C_{12}O_2H_3Cl_5$	14
$C_{12}O_2H_4Cl_4$	22
$C_{12}O_2H_5Cl_3$	14
$C_{12}O_2H_6Cl_2$	10
$C_{12}O_2H_7Cl_1$	2
$C_{12}O_2H_8$	1
total	76

Table 1. Isomer numbers of dioxins

Chapter 5

Mathematical appendix

5.1 Historical development of the isomerism problem

In 1797 already Alexander von Humboldt stated that chemical compounds may exist with different properties but with the same atomic composition. In the book

> A. von Humboldt: Versuche über die gereizte Muskel- und Nervenfaser, nebst Vermutungen über den chemischen Prozeß des Lebens in der Tier- und Pflanzenwelt, *Rottmann, Leipzig* 1797,

he wrote:

> *Drei Körper a, b und c können aus gleichen Quantitäten Sauerstoff, Wasserstoff, Kohlenstoff, Stickstoff und Metall zusammengesetzt und in ihrer Natur doch unendlich verschieden sein.*

(Three substances a, b and c may consist of the same quantities of oxygen, hydrogen, carbon, nitrogen and metal, but may differ infinitely in their nature.) A quarter of a century later, this statement was verified more or less at the same time by J. L. Gay-Lussac and J. von Liebig, and independently by F. Wöhler — after Gay-Lussac and Liebig had sufficiently developed the methods of analytical chemistry. (It is interesting to see that Humboldt and Gay-Lussac were close friends while von Liebig was a protégé of Humboldt, but it is not known to us whether there was a direct influence of Humboldt concerning the isomerism problem.) They discovered compounds with the same molecular formula but with different properties. Here is a corresponding quotation from a footnote to a paper of Wöhler, by Gay-Lussac:

> *... comme ces deux acides sont très différents, il faudrait pour expliquer leur différence admettre entre leurs éléments un mode de combinaison différent. C'est un objet qui appelle un nouveau examen.*

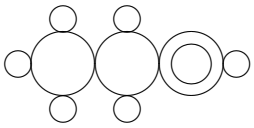
(As these two acids are very different, to explain their difference one has to allow different combinations of their elements. This is a topic that appeals to a new examination.)

It was not before 1830 that Berzelius recognized this fact as a general phenomenon and called it *isomerism*.

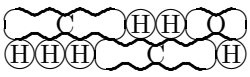
In order to find the reason for this phenomenon, chemists started to visualize molecules. Here are a few prominent versions of the molecule of ethanol C_2H_5OH. First we give Couper's version. It shows *two* oxygen atoms which is due to the fact that at his time it was not clear what the atomic weight of oxygen would be:

$$\begin{array}{ll} C & \left\{ \begin{array}{l} O \cdots OH \\ \quad H_2 \end{array} \right. \\ \vdots & \\ C & \cdots H_3 \end{array}$$

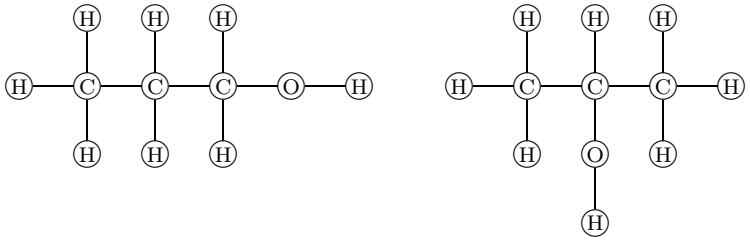
Here is Loschmidt's version:



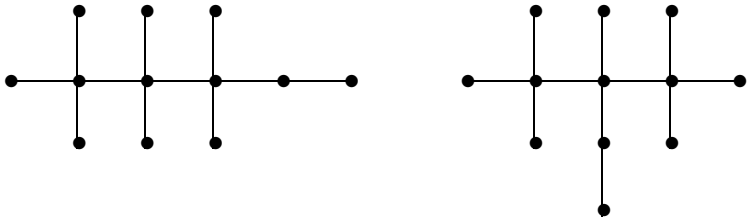
and this is Kekulé's way of drawing this molecule:



Loschmidt's version in fact solved the problem, but the breakthrough came with a method invented by Alexander Crum Brown, who replaced Couper's *dots* and Loschmidt's and Kekulé's *points of contact*, which indicate interatomic bonds, by the much more distinguishing *edges*. Crum Brown showed in 1864 among other things that for propanol there exist in fact *two* ways of connecting three carbon atoms with valence 4, eight hydrogen atoms of valence 1, and one oxygen atom with valence 2 in such a way that an alcohol arises, which means that there is a hydroxyl group, i.e. the oxygen atom has a bond to a hydrogen atom. The two possibilities are shown below:



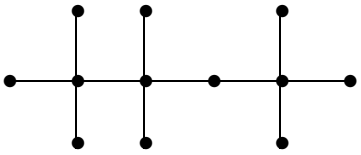
The corresponding alcohols in fact have different boiling points: 97.1 °C and 82.4 °C, respectively. If we omit the letters indicating the atom types, then we obtain the corresponding *molecular graphs*:



The solution of the isomerism problem given by Loschmidt and Crum Brown therefore can be formulated in the following way:

> *To a given molecular formula there may exist different molecules that have this particular molecular formula but different molecular graphs.*

Thus the reason for chemical isomerism was found, and graph theory was born (of course, as in many other cases, this mathematical theory had several origins, another one was the Königsberg bridges' problem, and a third one the theory of electric circuits, for example). But the question remained how the molecular graphs corresponding to a given molecular formula can be obtained, or at least how they can be counted. Chemical compounds that correspond to the same molecular formula (say to C_3H_8O) but have different graphs, i.e. whose atoms are connected in a different way, are called *connectivity* or *constitutional isomers*, and the graphs are called *structural formulae* or *structures*. It is important to note that connectivity isomers corresponding to a given molecular formula can belong to different compound classes. For example, for C_3H_8O there is a third molecular graph, shown in the next figure. It does not represent an alcohol, since there is no hydroxyl group, as the only vertex with valence two is not connected to a vertex of valence one:



In other words: For the *molecular formula* C_3H_8O there exist exactly three connectivity isomers (= three structural formulae), two of which contain a hydroxyl group and thus are alcohols, while the third contains a C-O-C fragment and thus is an ether.

In summary, the problem of determining a molecular structure is the following one (first approximation):

From chemical information, such as a given molecular formula, compound class, spectroscopic data etc., we have to evaluate all connectivity isomers that fulfil these conditions. The connectivity isomers corresponding to a given molecular formula are just the connected multigraphs with the given sequence of valences, the vertices of which are colored with the element symbol given by the molecular formula. It would be nice if we could separate the isomers into compound classes, i.e. we wish to be able to prescribe substructures optionally, for example hydroxyl groups, and we should also be able to forbid substructures.

This is the central mathematical problem of molecular structure elucidation. Therefore, in the center of program systems devoted to this problem, **a generator for molecular graphs is essential**, which allows to generate in an efficient way the complete set of all molecular graphs that correspond to a given set of conditions.

5.2 3D placing of molecules

Besides the feature of 2D placement mentioned above, a spatial placement is implemented in MOLGEN . It is carried out by minimizing an empirical energy function.

5.2.1 An empirical energy function

To obtain information about the 3D shape of a molecule, we use a mechanical model. In the mathematical formula for the energy all forces that affect the molecule are summed up. The function created this way is called *empirical energy function* since it was derived experimentally, not in a general theoretic way. The potential model used in MOLGEN is a simplified version of the MM2 model by N. L. Allinger. The function consists of the following elements:

- The first aspect taken into account is the length of each bond. The average length of covalent bonds in a molecule can be determined sufficiently exactly by spectroscopic measurements. The deviation from this ideal value appears quadratically in the potential:

$$E_s = 143.88 \cdot \frac{k_s}{2}(|x_i - x_j| - l_{ij}^{tab})^2.$$

  Here x_i and x_j are the vectors of two bonded atoms, l_{ij}^{tab} is the "ideal" length of a bond of the respective type, and k_s is a constant that depends on the two atom types and the bond degree. For a C-C single bond of two sp^3-hybridized carbon atoms, $k_s = 4.4$ and $l_{ij}^{tab} = 1.523$ Å.

- Another important influence comes from each bond angle. Again the deviation from a spectroscopically derived "ideal" value is considered:

$$E_b = 0.043828 \cdot \frac{k_b}{2}(\alpha_{ijk} - \alpha_{ijk}^{tab})^2.$$

The constants have an analogous meaning. For a C-C-C arrangement, where the central C is sp^3-hybridized, there is $k_b = 0.45$ and $\alpha_{ijk}^{tab} = 109.470^0$.

- Moreover, the torsion of four serially bonded atoms contributes to the potential. Here the first three terms of a Fourier series are used,

$$E_\omega = \frac{V_1}{2} \cdot (1 + \cos\omega) + \frac{V_2}{2} \cdot (1 - \cos 2\omega) + \frac{V_3}{2} \cdot (1 + \cos 3\omega).$$

There ω is the torsion angle and V_1, V_2 and V_3 are constants depending on atom types. For a C-C-C-C sequence $V_1 = 0.2, V_2 = 0.27$ and $V_3 = 0.093$. The torsion potentials are calculated and summed for all such sequences of four atoms.

- Atoms that are not covalently bonded also interact. The contribution of their van der Waals interaction to the potential is:

$$E_{vdW} = \sqrt{E_i \cdot E_j} \cdot \begin{cases} 290,000 \cdot \left(\exp\left(-\frac{12.5}{p_{ij}}\right) - 2.25 \cdot p_{ij}^6\right), & \text{if } p_{ij} \leq 3.311; \\ 336.176 \cdot p_{ij}^2, & \text{else.} \end{cases}$$

Here E_i is another constant (an atom's "hardness") and p_{ij} is the ratio of the sum of the van der Waals radii and the interatomic distance,

$$p_{ij} = \frac{R_i + R_j}{|x_i - x_j|}.$$

An sp^3-hybridized carbon atom has $R_i = 1.9$ Å and for an interaction of two such atoms $\sqrt{E_i \cdot E_j} = 0.044$.

The total potential function is therefore the following[1]

$$
\begin{aligned}
E &= \sum_{bond\ i,j} 143.88 \cdot \frac{k_s}{2}(|x_i - x_j| - l_{ij}^{tab})^2 \\
&+ \sum_{angle\ i,j,k} 0.043828 \cdot \frac{k_b}{2}(\alpha_{ijk} - \alpha_{ijk}^{tab})^2 \\
&+ \sum_{torsion\ angle\ i,j,k,l} \left(\frac{V_1}{2} \cdot (1 + \cos\omega) + \frac{V_2}{2} \cdot (1 - \cos 2\omega) + \frac{V_3}{2} \cdot (1 + \cos 3\omega) \right) \\
&+ \sum_{i,j} \sqrt{E_i \cdot E_j} \cdot \begin{cases} 290,000 \cdot \left(\exp\left(-\frac{12.5}{p_{ij}} \right) - 2.25 \cdot p_{ij}^6 \right), & \text{if } p_{ij} \leq 3.311; \\ 336.176 \cdot p_{ij}^2, & \text{else.} \end{cases}
\end{aligned}
$$

In the last term summation is over all pairs of atoms at least three bonds apart. For details see

> N.L. Allinger: MM2. A hydrocarbon force field utilizing V_1 and V_2 torsional terms, *J. Am. Chem. Soc.* **99**, 8127-8134 (1977).
>
> U. Burkert, N.L. Allinger: Molecular Mechanics, *ACS Monograph* **177**, American Chemical Society, Washington, D.C., 1982.

A necessary condition for a conformer is that the empirical potential has a local minimum. So a numerical method is used in order to find such a minimum.

5.2.2 The cg-method

For the minimization of a sufficiently smooth, non-linear function there are several numerical methods. The simplest method is to vary the coordinates in the direction of the negative gradient, i.e. the steepest descent. There is, however, no guarantee to find the minimum by this method after a finite number of steps. The Newton method therefore considers also the matrix of second derivatives. This matrix must be computed completely. If we would use the empirical potential the speed of the program would be decreased drastically. Hence for the solution of our minimization problem the method of conjugate gradients was chosen.

This method was originally developed to solve linear equation systems $Ax = b$. You start with the assumption that the exact solution x is a minimum of

$$F(z) = \frac{1}{2} z^T A z - b^T z + \frac{1}{2} b^T A^{-1} b$$

that is $F(x) = \min(F(z)) = 0$. While the method of steepest descent minimizes in one

[1] Further interactions such as electromagnetic ones are not taken into account.

dimension only, here in step $x_k \to x_{k+1}$, a $(k+1)$-dimensional minimization is carried out:

$$F(x_{k+1}) = \min_{\mu_0,\ldots,\mu_k} F(x_k + \mu_0 r_0 + \ldots + \mu_k r_k), \quad \text{where} \quad r_i = b - Ax_i, \quad \text{for} \quad i \leq k.$$

The application to the minimization of a quadratic function such as

$$f(x) = f(h) + \frac{1}{2}(x-h)^T A(x-h)$$

is now clear, since $Ax = b$ must hold with $b = Ah$ in order to make the gradient ∇f vanish. In the solution of the equations the directions $p_0, p_1, \ldots$ are calculated, so that p_{k+1} is a linear combination of $\nabla f(x_{k+1})$ and $p_i^T A p_j = 0$ for $i \neq j$. Fortunately many of the coefficients vanish, so that only $p_{k+1} = \nabla f(x_{k+1}) + \beta_k p_k$ remains, where $\beta_k = \frac{\nabla f(x_{k+1})^2}{\nabla f(x_k)^2}$. So the algorithm for the minimization reads:

The cg-algorithm

1. Choose a random start vector $x_0 \in \mathbb{R}^{3n}$ and set $g_0 := \nabla f(x_0)$ and $p_0 := -g_0$.

2. If $g_k = 0$: Stop. Else:

3. Determine x_{k+1} by linear minimization from:

$$f(x_{k+1}) = \min_{\lambda \geq 0} f(x_k - \lambda p_k).$$

   Set $g_{k+1} = \nabla f(x_{k+1})$ and $\beta_k = \frac{g_{k+1}^2}{g_k^2}$ and $p_{k+1} = -g_{k+1} + \beta_k p_k$.

4. Go to 2.

If a minimum exists at all, this method provides it in at most n steps. The computed extremum is, however, only local, i.e. the resulting conformations may differ if we use different initial values.

Chapter 6

Appendix

6.1 Errors and limits

6.1.1 Error and event messages

Primarily we want to discuss the messages that appear in the INFO window. Besides these there are a number of other messages and dialog boxes that hopefully explain themselves. The result window always pops up after a generator or expander run. If everything was correct, only the final result and the time needed are displayed. Additionally there are three other kinds of messages:

- **Ignorable error messages** inform you that something in your input data was wrong. Such discrepancies may simply be ignored (e.g. a nonexisting substructure will not be taken into account); so starting the generator (expander) is still possible. Messages of this kind are indicated by the capital letter "**W**" and a serial number.

- **Fatal error messages** also refer to wrong input data. In contrast to the messages described above they cannot be ignored. Starting the generator (expander) will be possible only after correcting the error. Messages of this kind are indicated by the capital letter "**F**" and a serial number.

- **Runtime messages** report events that occurred during a calculation of the generator etc. (e.g. memory failure or interruption). Messages of this kind are indicated by the capital letter "**R**" and a serial number.

In the following you find the complete list of all messages. Some of them should not occur in normal usage. For the sake of completeness and safety, however, we document them. They are marked by an asterisk. If one of these should occur, you may contact the address given in this manual. We tried to explain the somewhat complicated input errors by additional examples.

Messages of the generator

Ignorable error messages:

[**W12**] A substructure you wanted to use is not available in the library. All substructures are saved as files under the given substructure name in the subdirectory **bib** (or other subdirectories installed by yourself). Probably you spelt the name wrongly or the file has been deleted in the meantime. Apart from that there might be problems with your hard disk.

[**W13**] A substructure you wanted to use has either no free valences or another (internal) error. The generator can only process substructures that have at least one free valence, because otherwise the substructure is already a complete molecule. If there is an internal error, you should try to create the substructure file once more.

[**W14**] A macroatom you wanted to use contains the atom symbol X. This symbol is exclusively reserved for dummy atoms (see Subsection 4.1.2) in goodlist or badlist structures.

[**W15**] A macroatom you wanted to use cannot be subtracted from the molecular formula. In the case when a macroatom is given, the generator tries to subtract all atoms of this macroatom from the molecular formula. If a macroatom is prescribed to be contained several times in the molecule in question, the atoms are removed the corresponding number of times. The most frequent reason for this error is that a molecular formula contains not enough or the wrong atoms. For example, from C_6H_6 no OH can be removed and from C_6H_6O no two OH. Another reason may be that the valences in the molecular formula are incompatible with the valences in the macroatom, for instance, 6-valent sulfur in the molecular formula and 4-valent sulfur in the macroatom.

[**W17**] A restriction was entered that does not match with the (possibly reduced) molecular formula. This may be either a senseless ring size (e.g. ring size 2) or an unreasonable maximal bond degree (e.g. 13). You should also be aware of the fact that only atoms with valence greater than 1 are able to build rings. The error can also occur if ring size and maximal bond degree did match with the original molecular formula, but this molecular formula was reduced by macroatoms in the meantime. If you prescribe, for instance, a carbon 6-membered ring R in C_8H_{10}, the formula will be reduced to $C_2H_{10}R_1$. Therefore the resulting structures can primarily contain 3-membered rings at most. Only after expansion, ring sizes between 3 and 8 are possible again.

[**W28**] A goodlist or badlist structure cannot be compared with the resulting structures calculated by the generator. This happens if the atom types in the molecular formula do not match with all the types in the substructure, i.e. each atom type of the substructure must occur in the molecular formula. E.g. no OH group occurs in the isomers of

C_6H_6. Moreover, also the valences must match. Sulfur of valence 6, for example, in the substructure does not fit to 4-valent sulfur in the molecular formula. The error may also occur if the substructure matched with the original molecular formula, but does not match with the formula reduced by macroatoms. If you choose, for example, a C-O-structure as macroatom CO in $C_6ClH_{11}O$, the formula will be reduced to $C_5ClH_{11}(CO)_1$. Using a O-Cl-structure in the badlist will then provoke this message, since the resulting structures of the generator primarily contain no oxygen. Only after expansion of CO, the group O-Cl can be used in the badlist. For a better overview you may first generate a few structures and see which goodlist or badlist entries are reasonable.

[**W29**] A macroatom which you wanted to use cannot be subtracted from the molecular formula, since the formula reduced by this macroatom would be invalid. This error occurs if a macroatom is incompatible with the molecular formula. A simple example will explain this: Generate all 25 isomers of C_6H_{12}. In each molecule there is either a ring or a double bond but no triple bond (this is impossible for mathematical reasons). Now try to prescribe a C-C triple bond as a bivalent macroatom. The generator will reduce the formula and recognize that there cannot be any solutions, i.e. that the reduced formula is no longer valid.

Fatal error messages

[**F2***] The internal input file cannot be read correctly. In every run the generator reads from the file gen.inp. This file may have been changed, be corrupt or missing.

[**F3***] The internal output file cannot be created. This message appears if creation (writing) of an output file is impossible.

[**F6**] The molecular formula you want to use consists of too many atoms. MOLGEN is capable of handling structures with up to 100 atoms.

[**F7**] The molecular formula you want to use is not correct, i.e. there cannot exist any isomers with that formula.

[**F8**] The generator was started without entering a molecular formula. This message may also come from an internal error, e.g. if the input-file was not saved correctly.

[**F10**] (see **F8**)

[**F18**] The molecular formula consists of two univalent atoms. This case is not treated by the generator, since it is trivial.

[**F31***] The molecular formula you want to use comprises atoms the valence of which is too large. You can use atom valences up to 12 (see Subsection 6.1.2).

Event messages

[**R19***] The capacity of the generation procedure is exceeded. You got a rare special case. It is a so-called stack overflow.

[**R20***] The capacity of the generation procedure is exceeded. You got a rare special case. It is a so-called dimension overflow.

[**R25**] This message indicates abortion of the generation.

[**R26**] Occasionally the main memory's capacity is exhausted in a generator run. This usually happens in cases with a huge number of solutions. (For more details see Subsection 4.3.1).

[**R27**] This message indicates that the generator reached the given limit. In this case the generator solutions are nearly always incomplete. At the STOP AT field in the EDIT window you can set the number of solutions after which the generator shall stop.

Messages of the expander

Ignorable error messages

[**W7**] A substructure you wanted to use is not available in the library. All substructures are saved as files under the given substructure name in the subdirectory **bib** (or other subdirectories installed by yourself). Probably you spelt the name incorrectly or the file has been deleted in the meantime. Apart from that there might be problems with your hard disk.

[**W8**] A substructure you wanted to use has either no free valences or another (internal) error happened. The generator can only process substructures that have at least one free valence — because otherwise the substructure is already a complete molecule. If there is an internal error, you should try to create the substructure file once more.

[**W9**] A macroatom you wanted to use contains the atom symbol X. This symbol is exclusively reserved for dummy-atoms (see Subsection 4.1.2) in goodlist or badlist structures.

[**W10**] A macroatom you want to expand contains valences that do not match with the remaining valences of the result structures. For example, if there is a 6-valent sulfur in the molecular formula but a 4-valent one in the macroatom.

[**W11**] A macroatom you want to expand cannot be expanded. The frequent reason is that this macroatom has been expanded previously and, in a new run, still belongs to the input data. In this case you can ignore this message without any problem and select CONTINUE. It may also happen that the number of free valences of this macroatom does

not match with the number of free valences of the macroatoms that shall be expanded. You can control this by taking a look at the recent generator or expander result.

[**W13**] A restriction was entered that does not match with the (possibly reduced) molecular formula. This may be either a senseless ring size (e.g. ring size 2) or an unreasonable maximal bond degree (e.g. 13). You should also be aware of the fact that only atoms with valence greater than 1 are able to build rings and that the ring sizes and all other restrictions entered for the expander refer to the expanded isomers.

[**W21**] A goodlist or badlist structure is incompatible with the resulting structures calculated by the generator. That happens if the atom types in the molecular formula do not match with all the types in the substructure, i.e. each atom type of the substructure must occur in the molecular formula. E.g. no OH group occurs in the isomers of C_6H_6. Moreover, also the valences must match. Sulfur of valence 6, for example, in the substructure does not fit to 4-valent sulfur in the molecular formula. For a better overview you may first generate a few structures and see which goodlist or badlist entries are reasonable.

Fatal error messages

[**F1**] The internal file with the structures to be expanded cannot be opened. A frequent reason is that no structures have been constructed yet. Otherwise this message must be due to an internal error. For every calculation the expander reads all structures from the generator output file if the structures coming from the generator are to be expanded, or from a temporary file if the latest expander result is to be used. Perhaps these files are missing or there are problems with your hard disk.

[**F2***] The internal file with the structures to be expanded cannot be read correctly, i.e. it has a wrong format. For every calculation the expander reads all structures from internal files. Perhaps these files are missing or there are some problems with your hard disk.

[**F3***] The internal input file cannot be read correctly. For every calculation the expander reads an internal file. This file may have been changed, be damaged or missing.

[**F4***] The internal output file cannot be created. In every run the generator creates a file *.eou. This message appears if the creation (writing) of this file is impossible.

[**F5***] The molecular formula of the structures cannot be identified by the input data (For possible reasons, see **F2***).

[**F17**] The structures the expander has to compute consist of too many atoms. MOLGEN is capable of handling structures with up to 100 atoms (see Subsection 6.1.2). You should also note that the expanded structures may consist of more than 100 atoms if macroatoms are used.

Event messages

[**R18**] This message indicates abortion of the expansion.

[**R19**] The capacity of the main memory did not suffice. This is usually the case if many structures have to be kept in memory due to the isomorphism test. In this situation you may switch the test off. For details see Example 2 in Subsection 4.3.1.

[**R20**] This message indicates that the expander reached the given limit. In this case the expander solutions are nearly always incomplete. At STOP AT in the EDIT window you can set the number of solutions after which the expander shall stop.

6.1.2 Program Limits

When dealing with the problem of isomorphism one inevitably faces system limits due to the huge variability of structures. A natural limit of the program is always the calculation time needed to construct all isomers. We are now going to describe the most important program limits of MOLGEN .

Molecular formulae

The following limits regarding the molecular formula must be considered:

- The maximum **total number** of atoms in a structure is **100**.
- The **maximal valence** of an atom (or a macroatom) is **12**.

Disk space

The amount of free disk space of your computer determines how many structures can be stored. As an example, we discuss the 13,190 isomers of $C_8H_{16}O_2$: If all isomers are stored by the generator, the disk space needed is approximately 740 kB. For most applications, however, it will be sufficient to format and inspect approximately 1,000 isomers at a time. For the examination of large examples it is recommended first to run the generator or expander *without* storing the solutions. Afterwards, selected intervals of resulting structures can be written on disk. Therefore we included the option to save and also to format the solutions *in intervals* (see Subsection 3.3.2).

Main memory

- The **maximum number of structures** that can be handled by the 2D display is $2^{15} - 1 = 32767$.

This amount is surely sufficient for visual examination of results. However, if you need to generate more than this amount of structures in one run, and want to prepare them for automated processing through other software systems, do not hesitate to contact us. We provide a free tool that converts MOLGEN output to MDL SDfile format without any limitations.

As commonly known, the size of the main memory is limited. We tried to use the memory carefully wherever possible. It may, however, happen — sometimes with the generator, more frequently with the expander — that the memory space does not suffice. The generator runs into problems only in large examples with a huge number of solutions. (Unfortunately, in such situations there is no remedy.) Nevertheless, we hope you will hardly be afflicted by such trouble. The memory requirement of the expander is primarily determined by the isomorphism check, if macroatoms have to be expanded. How to manage difficulties in such cases is described in Subsection 4.3.1.

Chapter 7

Literature about MOLGEN

In this final chapter we give references to research papers, diploma theses and dissertations written during development and the implementation of the various MOLGEN versions. Many of these papers may be downloaded free of charge from the MOLGEN homepage

http://www.molgen.de

7.1 Structure generation

In the first place, MOLGEN is a generator of molecular structures. This is a difficult mathematical problem, the solution of which is a mixture of algebraic and combinatorial methods. Pioneering work was done in the DENDRAL project. Our idea was to provide an implementation that runs on a PC. The result — due to the development of new algorithms and a very careful analysis of suitable data structures — is one of the (two?) fastest generators of connectivity isomers, corresponding to a given molecular formula and (optional) further conditions like forbidden and prescribed substructures, given maximal ring size etc.. Details can be found in the following publications:

R. Gugisch, C. Rücker: Unified Generation of Conformations, Conformers, and Stereoisomers: A Discrete Mathematics-Based Approach. *MATCH Commun. Math. Comput. Chem. 61, 117-148, 2009.*
A. Kerber, R. Laue, M. Meringer, C. Rücker: Molecules in Silico: A Graph Description of Chemical Reactions. *J. Chem. Inf. Model. 47, 805-817, 2007.*
R. Laue, T. Grüner, M. Meringer, A. Kerber: Constrained Generation of Molecular Graphs. *DIMACS Series in Discrete Mathematics And Theoretical Computer Science 69, 319-332, 2005.*
C. Rücker, R. Gugisch, A. Kerber: Manual Construction and Mathematics- and Computer-Aided Counting of Stereoisomers. The Example of Inositols. *J. Chem. Inf.*

Comput. Sci. 44, 1654-1665, 2004.

J. Braun, R. Gugisch, A. Kerber, R. Laue, M. Meringer, C. Rücker: MOLGEN-CID - A Canonizer for Molecules and Graphs Accessible through the Internet. *J. Chem. Inf. Comput. Sci. 44, 542-548, 2004.*

R. Gugisch, A. Kerber, R. Laue, M. Meringer, C. Rücker: Kombinatorische Chemie, eine Herausforderung für Mathematik und Informatik. *Spektrum 1/02, Universität Bayreuth, 64-67, 2002.*

R. Gugisch, A. Kerber, R. Laue, M. Meringer, J. Weidinger: MOLGEN-COMB, a Software Package for Combinatorial Chemistry. *MATCH Commun. Math. Comput. Chem. 41, 189-203, 2000.*

T. Grüner, A. Kerber, R. Laue, M. Meringer: Mathematics for Combinatorial Chemistry. *Scientific Computing in Chemical Engineering II, 74-81, Springer-Verlag 1999.*

A. Kerber, R. Laue, T. Grüner, M. Meringer: MOLGEN 4.0. *MATCH Commun. Math. Comput. Chem. 37, 205-208, 1998.*

A. Kerber, R. Laue, T. Wieland: Erkennung, Beschreibung und Visualisierung molekularer Strukturen, *Proceedings of the BMBF Statusseminar Application-oriented joint projects in the field of mathematics.*

T. Wieland, A. Kerber, R. Laue: Principles of the Generation of Constitutional and Configurational Isomers, *J. Chem. Inf. Comput. Sci. 36, 413-419, 1996.*

C. Benecke, R. Grund, R. Hohberger, A. Kerber, R. Laue, T. Wieland: MOLGEN+, a Generator of Connectivity Isomers and Stereoisomers for Molecular Structure Elucidation, *Anal. Chim. Acta 314, 141-147, 1995.*

T. Wieland: Erzeugung, Abzählung und Konstruktion von Stereoisomeren, *MATCH Commun. Math. Comput. Chem. 31, 153-203, 1994.*

T. Wieland: Computerunterstützte Berechnung von Stereoisomeren. *Master's thesis, University of Bayreuth, 1994.*

7.2 Structure elucidation

The program system MOLGEN-MS is devoted to computer aided molecular structure elucidation. MOLGEN-MS is mainly adapted to low resolution electron impact mass spectra but also includes tools which allow to process high resolution data and results from atomic analysis. Even information gained from IR or NMR interpretation can be added. MOLGEN-MS arose from the idea to provide a database independent tool for molecular structure elucidation in both chemical industry, research and education.

The module that matches structural formulas and mass spectra is also available as console

application MOLGEN-MSF . MOLGEN-MS is available for Windows 9x / NT 4.0 / 2000 / XP / Vista. For details see the homepage of MOLGEN and the following references:

A. Kerber, M. Meringer, C. Rücker: CASE via MS: Ranking Structure Candidates by Mass Spectra. *Croatica Chemica Acta 79, 449-464, 2006.*

J. Meiler, M. Meringer: Ranking MOLGEN Structure Proposals by 13C NMR Chemical Shift Prediction with ANALYZE. *MATCH Commun. Math. Comput. Chem. 45, 85-108, 2002.*

A. Kerber, R. Laue, M. Meringer and K. Varmuza: MOLGEN-MS: Evaluation of Low Resolution Electron Impact Mass Spectra with MS Classification and Exhaustive Structure Generation. *Advances in Mass Spectrometry 15, 939-940, Wiley 2001.*

T. Grüner, A. Kerber, R. Laue, M. Meringer, K. Varmuza, W. Werther: MASSMOL. *MATCH Commun. Math. Comput. Chem. 38, 173-180, 1998.*

T. Grüner, A. Kerber, R. Laue, M. Liepelt, M. Meringer, K. Varmuza, W. Werther: Bestimmung von Summenformeln aus Massenspektren durch Erkennung überlagerter Isotopenmuster. *MATCH Commun. Math. Comput. Chem. 37, 163-177, 1998.*

7.3 QSAR/QSPR

MOLGEN-QSPR provides several tools for the application in combinatorial chemistry. It allows in particular to *construct* virtual combinatorial libraries. The input of this structure generator is a mathematical description of reactions and reactants. Using a canonical form, it is able to *compare combinatorial libraries*, in particular for testing whether a given real library is a subset of a constructed virtual library.

In order to predict physical, chemical or biological properties for the virtual libraries, various molecular descriptors are implemented that serve as input for regression analysis. At the moment there are 708 arithmetical, topological and geometrical descriptors included in our software.

Regression analysis correlates molecular descriptors with measured properties of the real library. Regression methods are provided by the statistics package "R" which is accessed directly from MOLGEN 's graphical user interface. So far multilinear regression, regression trees, neural networks and support vector machines are available in order to suggest promising candidate structures for the target property.

MOLGEN-QSPR is available for Windows 9x / NT 4.0 / 2000 / XP. Here are several relevant publications:

C. Rücker, G. Rücker, M. Meringer: y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model. 47, 2345-2357, 2007.*

C. Rücker, M. Scarsi, M. Meringer: 2D QSAR of PPARγ Agonist Binding and Transactivation. *Bioorg. Med. Chem. 14, 5178-5195, 2006.*

C. Rücker, M. Meringer, A. Kerber: QSPR Using MOLGEN-QSPR: The Challenge of Fluoroalkane Boiling Points. *J. Chem. Inf. Model. 45, 74-80, 2005.*

J. Braun, A. Kerber, M. Meringer, C. Rücker: Similarity of Molecular Descriptors: The Equivalence of Zagreb Indices and Walk Counts. *MATCH Commun. Math. Comput. Chem. 54, 163-176, 2005.*

C. Rücker, M. Meringer, A. Kerber: QSPR Using MOLGEN-QSPR: The Example of Haloalkane Boiling Points. *J. Chem. Inf. Comput. Sci. 44, 2070-2076, 2004.*

A. Kerber, R. Laue, M. Meringer, C. Rücker: MOLGEN-QSPR, a Software Package for the Study of Quantitative Structure Property Relationships. *MATCH Commun. Math. Comput. Chem. 51, 187-204, 2004.*

T. Wieland: The Use of Structure Generators in Predictive Pharmacology and Toxicology, *Arzneim.-Forsch./Drug Res., 46 (I), 223-227, 1996.*

7.4 Mixed and miscellaneous

R. Gugisch, A. Kerber, R.Laue, M. Meringer, C. Rücker: History and Progress of the Generation of Structural Formulae in Chemistry and its Applications. *MATCH Commun. Math. Comput. Chem. 58, 239-280, 2007.*

A. Kerber, R. Laue, M. Meringer, C. Rücker: Molecules in Silico: Potential versus Known Organic Compounds. *MATCH Commun. Math. Comput. Chem. 54, 301-312, 2005.*

A. Kerber, R. Laue, M. Meringer, C. Rücker: Molecules in Silico: The Generation of Structural Formulae and Its Applications. *J. Comput. Chem. Jpn. 3, 85-96, 2004.*

A. Kerber, R. Laue, M. Meringer: An Application of the Structure Generator MOLGEN to Patents in Chemistry. *MATCH Commun. Math. Comput. Chem. 47, 169-172, 2003.*

C. Rücker, J. Braun: UNIMOLIS - A Computer-Aided Course on Molecular Symmetry and Isomerism. *MATCH Commun. Math. Comput. Chem. 47, 173-174, 2003.*

C. Rücker, G. Rücker, M. Meringer: Exploring the Limits of Graph Invariant- and Spectrum-Based Discrimination of (Sub)structures. *J. Chem. Inf. Comput. Sci. 42, 640-650, 2002.*

C. Rücker, M. Meringer: How Many Organic Compounds are Graph-Theoretically

Nonplanar? *MATCH Commun. Math. Comput. Chem. 45, 153-172, 2002.*
T. Wieland: Mathematical Simulations in Combinatorial Chemistry, *MATCH Commun. Math. Comput. Chem. 34, 179-206, 1996.*
C. Benecke, R. Grund, A. Kerber, R. Laue, T. Wieland: Chemical Education via MOLGEN, *J. Chem. Educ. 72, 403-406, 1995.*

7.5 Mathematical Methods

Most important was, of course, the development of mathematical structures and algorithms, as well as the careful selection of suitable and efficient data structures. They are described, for example, in the following theses:

D. Moser: , *diploma thesis, University of Bayreuth, 1987.*
R. Grund: Konstruktion molekularer Graphen mit gegebenen Hybridisierungen und überlappungsfreien Fragmenten, *doctoral thesis, University of Bayreuth, 1994.*
Th. Wieland: Konstruktionsalgorithmen bei molekularen Graphen und deren Anwendung, *doctoral thesis, University of Bayreuth, 1996.*
Th. Grüner: Strategien zur Konstruktion diskreter Strukturen, *doctoral thesis, University of Bayreuth, 1998.*
M. Meringer: Mathematische Modelle für die kombinatorische Chemie und die molekulare Strukturaufklärung, *doctoral thesis, University of Bayreuth, 2004.*
R. Gugisch: Konstruktion von Isomorphieklassen orientierter Matroide, *doctoral thesis, University of Bayreuth, 2005.*