MOLGEN–QSPR
User Guide

Software for Computation and Application of
Quantitative Structure – Property Relationships

J. Braun, A. Kerber, R. Laue, M. Meringer, C. Rücker,

Bayreuth, München, Freiburg,

June 10, 2009

Contents

Introduction

The software package MOLGEN–QSPR provides methods for the study of quantitative structure–property relationships (QSPRs) and the prediction of property values for compounds in virtual combinatorial libraries. Figure 1 shows a simplified flowchart of QSPR search and application.

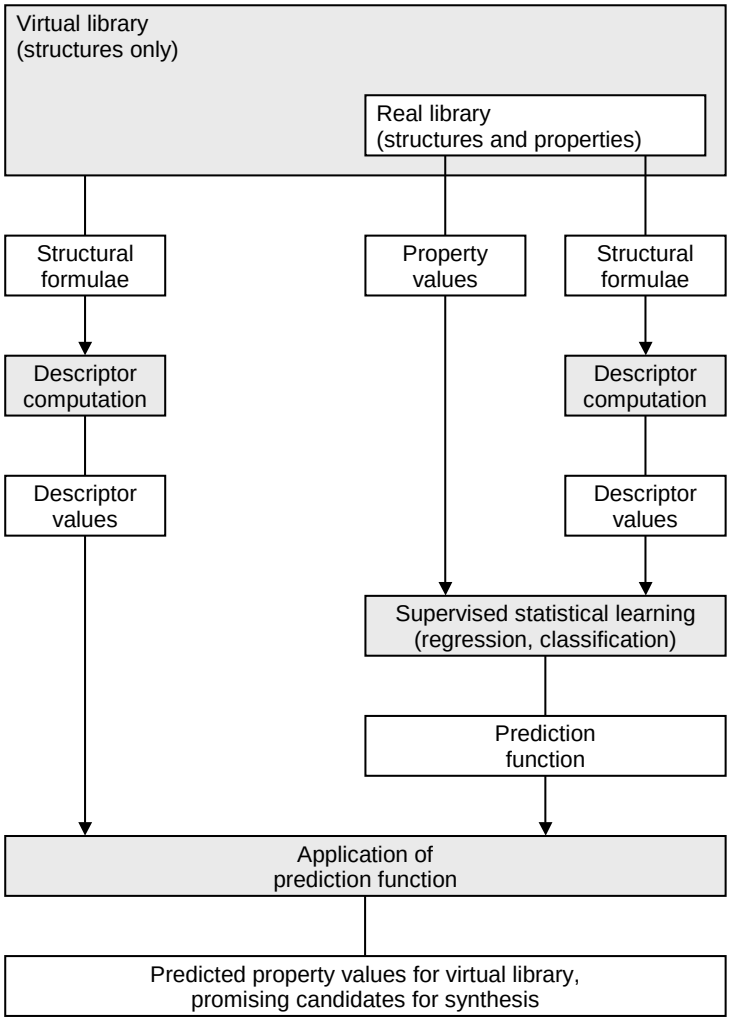


Figure 1: Flowchart of QSPR search and application

The input of MOLGEN–QSPR is a set of chemical compounds given as molecular graphs together with values for a continuous target variable representing the physico-chemical property under consideration. In the following *tutorial* we will treat the boiling points of decanes as an example.

The QSPR search consists of four principal steps:

- structure preprocessing,
- descriptor computation,
- regression analysis and validation,
- prediction of unknown property values.

All these steps can be performed with MOLGEN–QSPR.

Structure preprocessing includes addition of H atoms, which are typically suppressed in electronic representations of molecular graphs, identification of aromatic bonds, which are often coded as alternating single and double bonds, and computation of a 3D layout using a force field model. The latter is necessary if geometrical descriptors are to be applied.

Molecular descriptors are used in order to map molecular structures onto real numbers. Currently MOLGEN–QSPR provides about 700 built–in descriptors of various types, among them arithmetical, topological and geometrical indices. Furthermore, substructure and fragment counts can be used as molecular descriptors.

Once the descriptor values are calculated, methods of supervised statistical learning are applied in order to find prediction functions that fit the target variable well. There are several methods available covering linear regression, artificial neural networks, support vector machines, regression trees and nearest neighbors regression.

Finally, if a good QSPR is found, it can be applied for property prediction for all members of a virtual combinatorial library. Such libraries can be constructed using MOLGEN's structure generators.

MOLGEN–QSPR's features such as structure generation, structure canonization and removal of duplicates, numerous descriptors of various types, descriptor transformation, its ability to plot each variable (including residuals and predictions) *vs* each other variable, its variety of statistical learning methods, and its ability to provide predictions for complete sets of compounds render MOLGEN–QSPR unique among similar programs.

Chapter 1

First steps

1.1 System Requirements

MOLGEN–QSPR is available for *MS Windows 95/98/NT4.0/Me/ 2000/XP/Vista*.

1.1.1 Hardware

In order to use MOLGEN–QSPR the following hardware requirements have to be fulfilled:

- IBM–compatible PC (80486 or higher).
- CD–ROM drive for installation.
- At least 10 MB RAM and the same amount of free disc space. The space needed depends of course on the problem, i.e. on the number of structural formulas to be processed.

1.1.2 Software

Some of the algorithms included in MOLGEN–QSPR call routines provided by the software package for statistical computing *R 2.8.1* or higher. This software can be downloaded free of charge at `http://cran.r-project.org/`. In order to be able to access sophisticated regression methods, additionally the following R packages need to be installed: *tree* (regression trees), *e1071* (support vector machines) and *pls* (partial least squares).

1.2 Installation

MOLGEN–QSPR consists of one executable and does not require any DLLs or anything else. Therefore you can start it already from the CD-ROM. However, it is useful to copy the program and the sample files on your hard disc. Proceed as follows:

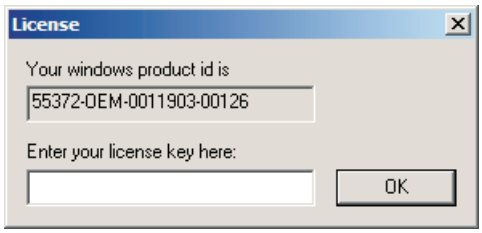


Figure 1.1: *License* dialogue

1. Insert the MOLGEN–QSPR installation CD–ROM into your CD–ROM drive.
2. Copy the complete folder MOLGEN–QSPR into the *Programs* directory of your hard disc drive. This is located for instance at `C:\ Program Files`.
3. Optionally create shortcuts to your desktop or your start menu.

1.3 Activation

After you first start MOLGEN–QSPR the *License* dialogue (Figure 1.1) will be displayed. Please send your windows product id to

`molgen@molgen.de`

You will receive a license key for activation.

1.4 Demo

For evaluation purposes a free demo license can be ordered. In case you received such a demo version, no license key will be required. The demo license offers full functionality for calculating QSPRs. However, import functions are limited: Only the input files *DecanesReal.sdf* and *DecanesReal.txt* delivered with the demo version can be imported. Structure generators are not accessible in the demo version.

Chapter 2

Tutorial

This part of the MOLGEN–QSPR *User Guide* gives a brief description of all you need to know for your first QSPR calculations. It is described step by step, beginning with data input, followed by descriptor calculation, regression analysis, and finally property prediction.

2.1 Data Input

2.1.1 Importing Structural Formulas

There are several possibilities to import electronically stored chemical structures. For our first example we import a library of 50 decanes stored as *MDL SDfile* on the MOLGEN–QSPR CD.

1. Click on *File|Import...* to get to the *Import File* dialogue.
2. Select *SDfiles (∗.sdf)* in the *Filetype* combo box.
3. Click on *DecanesReal.sdf* in order to select the desired *SDfile*.
4. Use the *Open* button to open the selected file.

The 50 decanes (the real library) will now be displayed as *Molecule* document on the screen (Figure 2.1).

There are various functions and controls available to modify the layout of structures, for instance

- *View|Hydrogens* to display hydrogen atoms,
- *View|Symbols* to display element symbols,

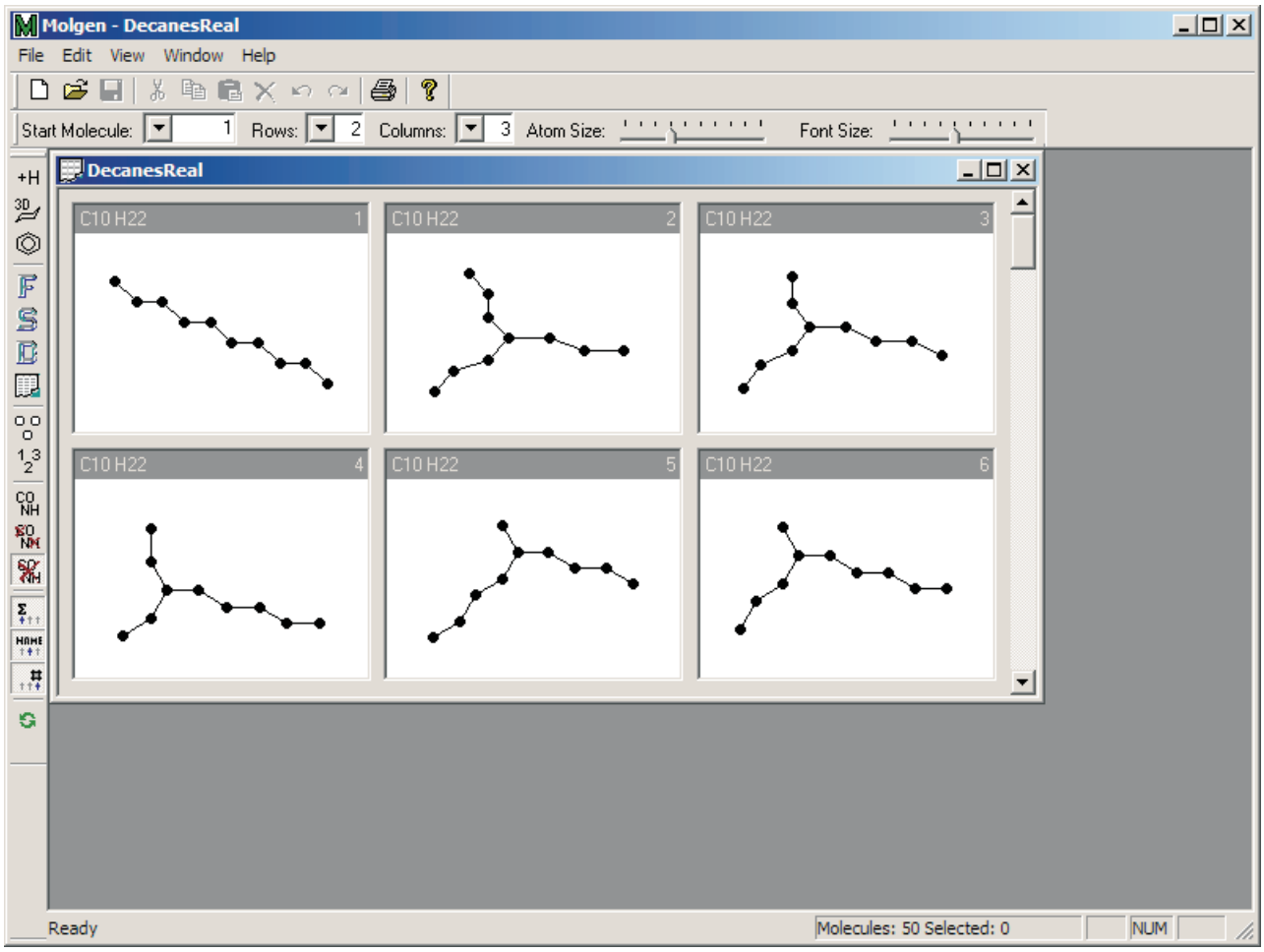


Figure 2.1: *Molecule* document containing 50 decanes

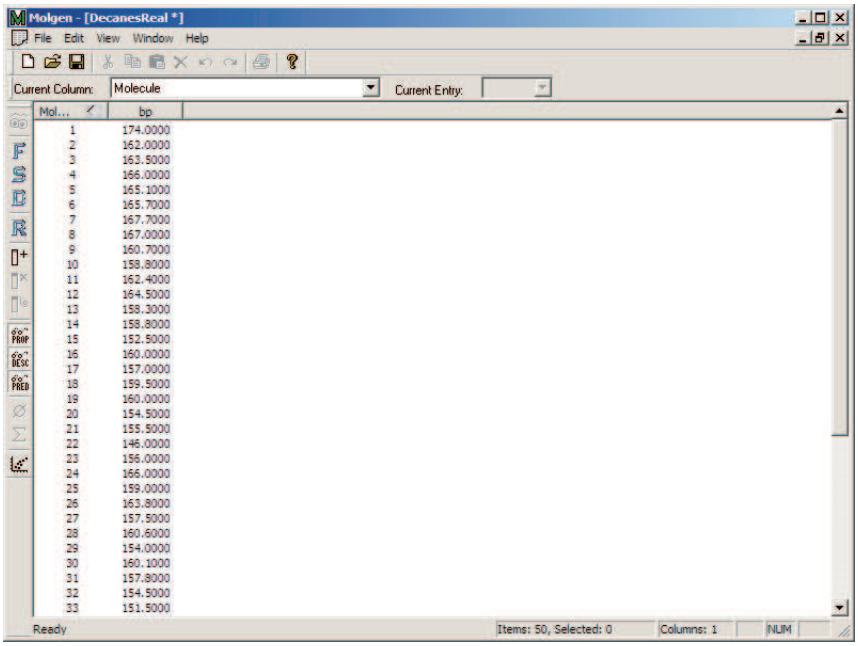


Figure 2.2: *Molecular Descriptors* document containing 50 boiling points

- *Start Molecule* combo box and the scrollbar to navigate through the library,
- *Rows* and *Columns* combo boxes to change the grid, etc.

2.1.2 Importing Property Values

The next step in a QSPR study is to supply property values for the structures. In this example property values are stored in a *tabulator separated ascii table*. Such a file is structured in the following way: The first line contains column heads, the following lines contain data for compounds, one line for each compound. The first column contains the compound name, the following column(s) contain(s) property values. Columns are separated by tabulators. Such a file is already prepared with boiling points of the structures above. Use the following steps to import the property file:

1. Click on *File|Import...* to open the *Import File* dialogue.
2. Select *Ascii Table (tabulator separated) (*.txt)* in the *Filetype* combo box.
3. Click on *DecanesReal.txt* in order to select the desired file.
4. Use the *Open* button to open the selected file.

The boiling points of the real library will now be displayed on the screen (Figure 2.2).

The status bar shows that there are 50 rows and one column in this file (the structure names are not counted as column). Again, there are various functions available to change the layout of the table and to retrieve additional information about the data, for instance

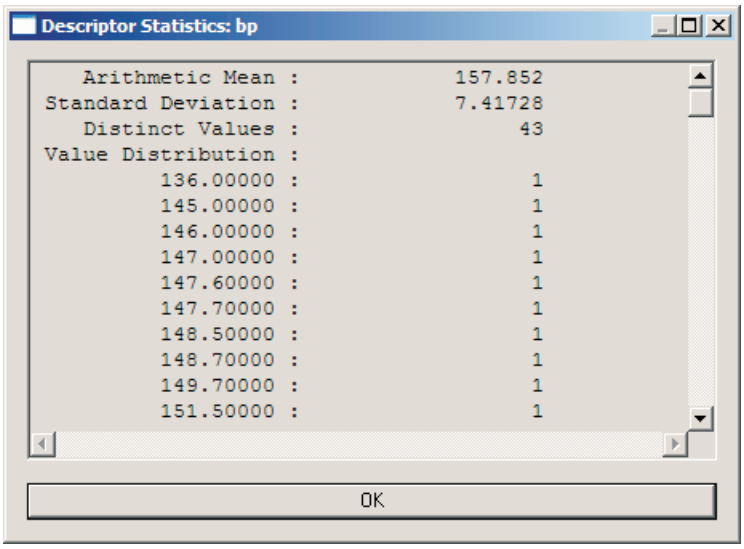


Figure 2.3: *Descriptor Statistics* dialogue

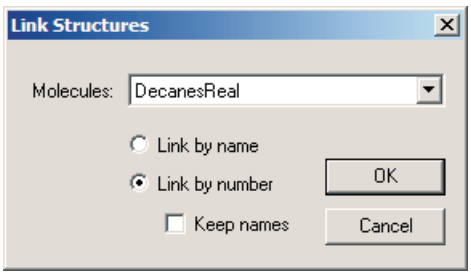


Figure 2.4: *Link Structures* dialogue

- Click on a column head to sort rows by ascending/descending values and to simultaneously make this particular column the current column.
- The *Current Column* combo box offers a way to change the current column without sorting rows.
- The current column is always marked by one of the symbols , or .
- Use *View|Statistics* to display some fundamental statistical values of the current column such as arithmetic mean or standard deviation (Figure 2.3).

2.1.3 Linking Structures and Property Values

The property values are not yet linked to the structures from the *Molecule* document. Therefore use *File|Link Structures* (Figure 2.4).

Use the *Molecules* combo box to select the structures and *Link by number*. By clicking *OK* the structures will be linked to the table with the property values. It can be useful to save this document with *File|Save* (Figure 2.5).

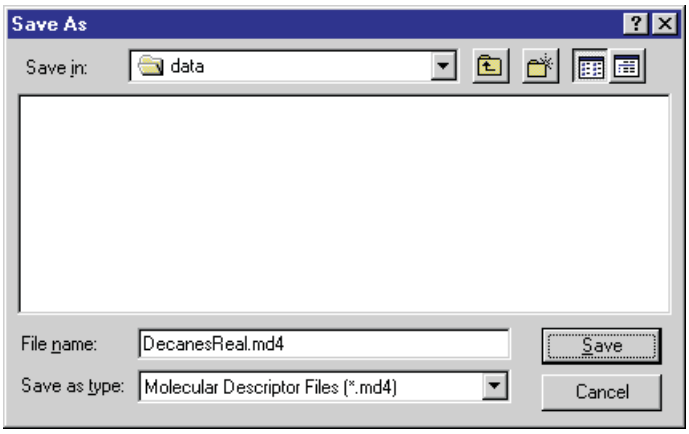


Figure 2.5: *File Save* dialogue

A *Molecular Descriptors File* (extension *.md4*) is created. At this moment it contains molecular structures together with property values, later it will also contain descriptor values and other data. If the initially imported sdf file provided compound names, these are now displayable using *View|Names*.

2.1.4 Alternatives for Data Input

Of course there exist various alternatives to supply data for QSPR studies, and MOLGEN–QSPR offers several other ways for data import. Among these are

- Edit structures with the built–in structure editor MOLED, use *File|New|Moled* to draw a molecular structure as a molfile.
- Import structures from several *MDL Molfiles*, use *File|New|Molecules* and then *File|Append*.
- Import structures and property values from *CODESSA input files*. Use *File|Import* and select an *.inp* file.
- Add and edit property values within an existing *Molecular Descriptors* document, see Subsections 2.2.2 and 2.2.3.

2.2 Displaying and Editing Data

Before starting the molecular descriptors calculation we will have a closer look at some functionality of the *Molecular Descriptors* document.

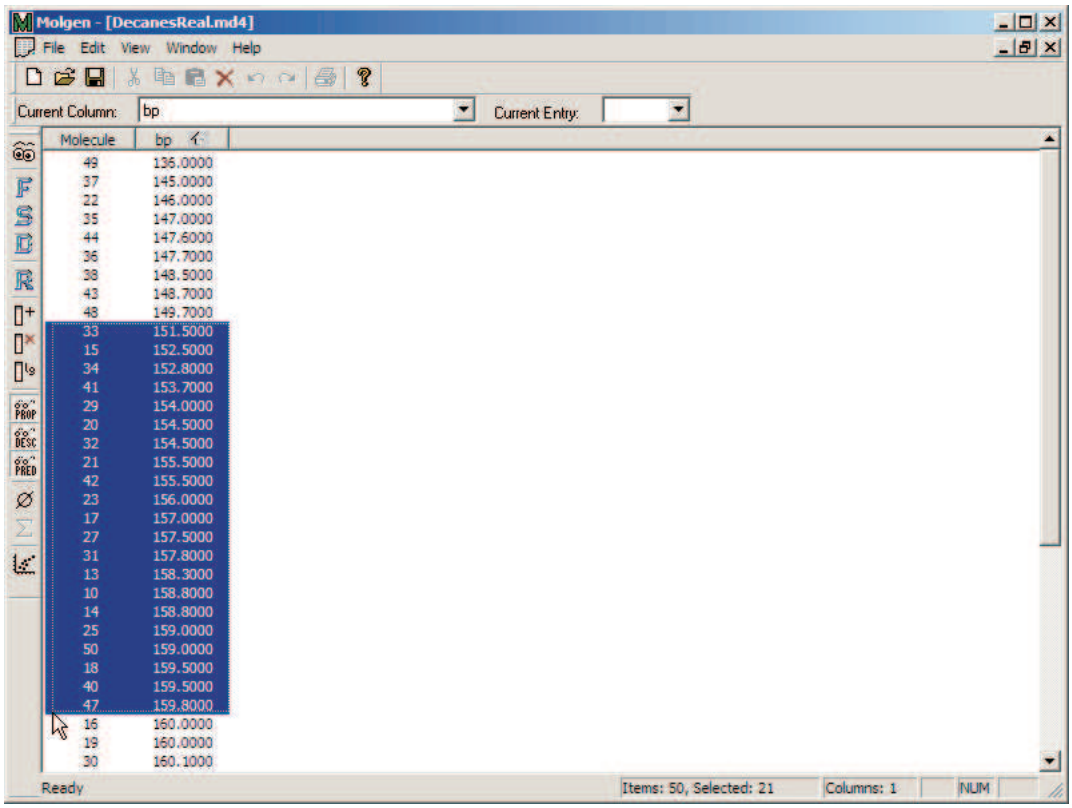


Figure 2.6: Selection of rows with bps between 150 and 160 °C

2.2.1 Displaying Structural Formulas

As already mentioned, rows can be sorted by property values. If we want to have a look at the decanes of our real library with bps above 150°C and below 160°C we have to conduct the following steps:

1. Click on the bp column head to sort rows by ascending bps.
2. Use the left mouse button to select all rows with bps between 150 and 160 (Figure 2.6).
3. *File|Pass Values* will cause the values of the current column to appear as names in a new *Molecule* document containing the selected structures.
4. Use *File|Molecules* to create the said new *Molecule* document of selected structures (Figure 2.7).

2.2.2 Editing Property Values

Often it is necessary to edit some property values after data import. To do so proceed as follows:

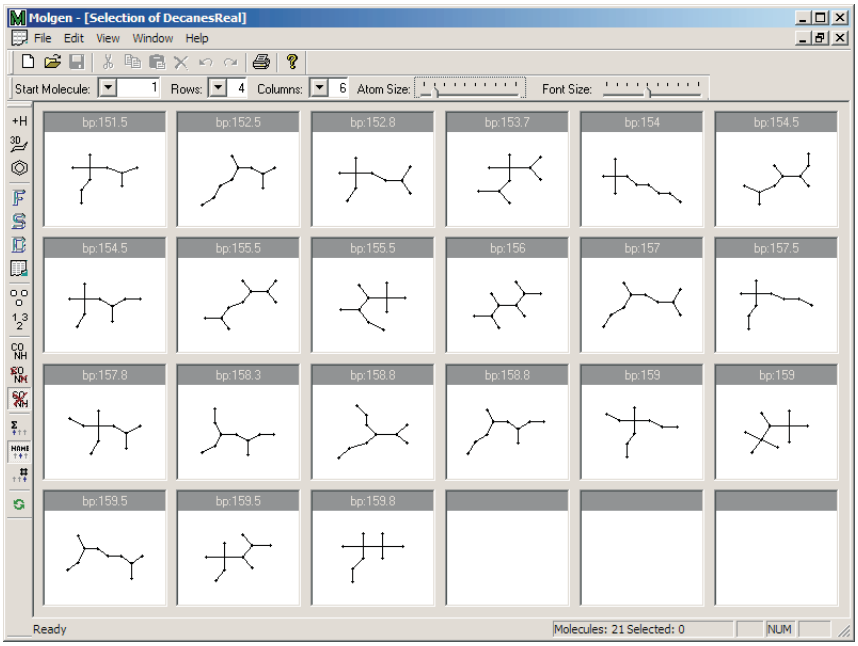


Figure 2.7: Structures with bps between 150 and 160 °C

Figure 2.8: Editing property values using the *Current Entry* combo box

1. Select the property column you want to edit by clicking the column head or using the *Current Column* combo box.

2. Select the row of the property value you want to edit. The *Current Entry* combo box becomes activated and the selected property value appears (Figure 2.8).

3. Edit the property value in the *Current Entry* combo box. The value is immediately transferred to its place in the *Molecular Descriptors* document.

2.2.3 Further Edit Operations

There are some further operations available to modify a *Molecular Descriptors* document. Selected row(s) can be deleted using *Edit|Delete*. To delete a column make it the current column, then click *Edit|Delete Column*. To delete several columns simultaneously, check them on the *Regression Setup Variables* page (see Section 2.5.1), click *OK* and then *Edit|Delete Columns*. A new column is added by *Edit|Add Column*.

2.3 Descriptor Calculation

For calculation of QSPRs we need values of molecular descriptors as input for statistical learning procedures. MOLGEN–QSPR offers three types of molecular descriptors: *Indices*, *substructure counts* and *fragment counts*.

2.3.1 Calculating Indices

Having the *Molecular Descriptor* document selected as active window

1. use *File|Indices* to obtain the *Molecular Descriptors* dialogue (Figure 2.9).

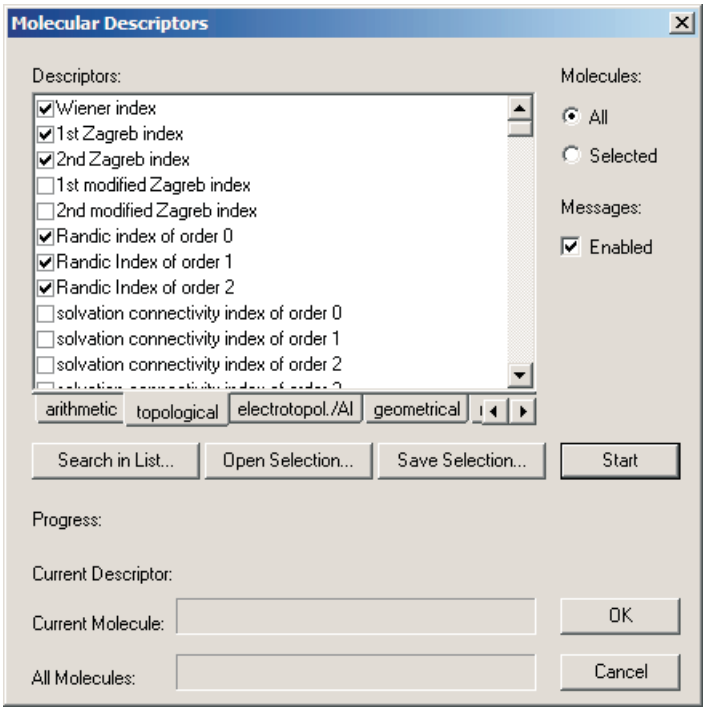
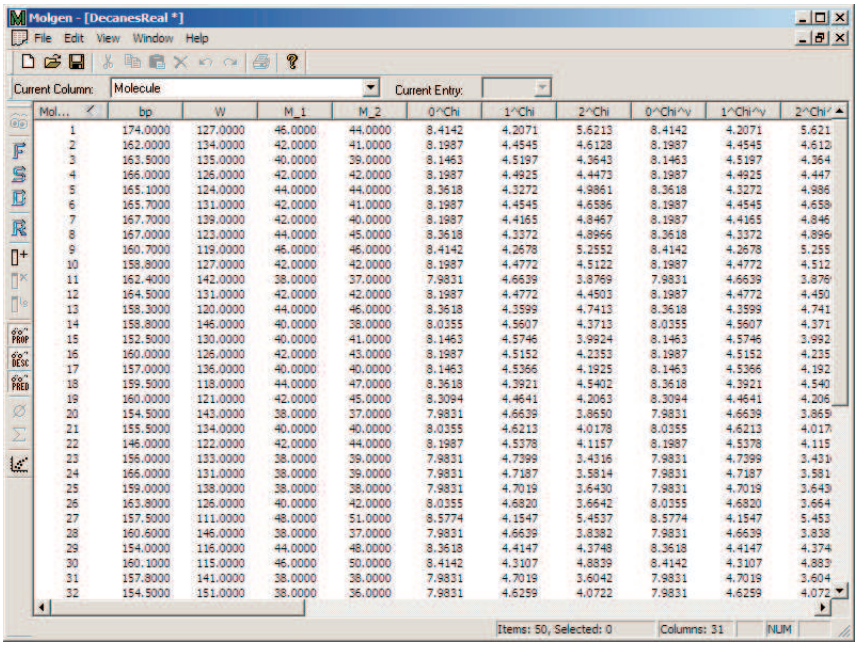


Figure 2.9: *Molecular Descriptors* dialogue

2. Activate check boxes in the *Descriptors* field to select descriptors to be calculated. Click the tabulator fields to switch between various categories of indices:

   - arithmetic indices,
   - topological indices,
   - electrotopological indices,
   - geometrical indices,
   - miscellaneous indices and
   - overall indices.



Mol...	bp	W	M_1	M_2	0^Chi	1^Chi	2^Chi	0^Chi^v	1^Chi^v	2^Chi^
1	174.0000	127.0000	46.0000	44.0000	8.4142	4.2071	5.6213	8.4142	4.2071	5.621
2	162.0000	134.0000	42.0000	41.0000	8.1987	4.4545	4.6128	8.1987	4.4545	4.612
3	163.5000	135.0000	40.0000	39.0000	8.1463	4.5197	4.3643	8.1463	4.5197	4.364
4	166.0000	126.0000	42.0000	42.0000	8.1987	4.4925	4.4473	8.1987	4.4925	4.447
5	165.1000	124.0000	44.0000	44.0000	8.3618	4.3272	4.9861	8.3618	4.3272	4.986
6	165.7000	131.0000	42.0000	41.0000	8.1987	4.4545	4.6586	8.1987	4.4545	4.658
7	167.7000	139.0000	42.0000	40.0000	8.1987	4.4165	4.8467	8.1987	4.4165	4.846
8	167.0000	123.0000	44.0000	45.0000	8.3618	4.3372	4.8966	8.3618	4.3372	4.896
9	160.7000	119.0000	46.0000	46.0000	8.4142	4.2678	5.2552	8.4142	4.2678	5.255
10	158.8000	127.0000	42.0000	42.0000	8.1987	4.4772	4.5122	8.1987	4.4772	4.512
11	162.4000	142.0000	38.0000	37.0000	7.9831	4.6639	3.8769	7.9831	4.6639	3.876
12	164.5000	131.0000	42.0000	42.0000	8.1987	4.4772	4.4503	8.1987	4.4772	4.450
13	158.3000	120.0000	44.0000	46.0000	8.3618	4.3599	4.7413	8.3618	4.3599	4.741
14	158.8000	146.0000	40.0000	38.0000	8.0355	4.5607	4.3713	8.0355	4.5607	4.371
15	152.5000	130.0000	40.0000	41.0000	8.1463	4.5746	3.9924	8.1463	4.5746	3.992
16	160.0000	126.0000	42.0000	43.0000	8.1987	4.5152	4.2353	8.1987	4.5152	4.235
17	157.0000	136.0000	40.0000	40.0000	8.1463	4.5366	4.1925	8.1463	4.5366	4.192
18	159.5000	118.0000	44.0000	47.0000	8.3618	4.3921	4.5402	8.3618	4.3921	4.540
19	160.0000	121.0000	42.0000	45.0000	8.3094	4.4641	4.2063	8.3094	4.4641	4.206
20	154.5000	143.0000	38.0000	37.0000	7.9831	4.6639	3.8650	7.9831	4.6639	3.865
21	155.5000	134.0000	40.0000	40.0000	8.0355	4.6213	4.0178	8.0355	4.6213	4.017
22	146.0000	122.0000	42.0000	44.0000	8.1987	4.5378	4.1157	8.1987	4.5378	4.115
23	156.0000	133.0000	38.0000	39.0000	7.9831	4.7399	3.4316	7.9831	4.7399	3.431
24	166.0000	131.0000	38.0000	39.0000	7.9831	4.7187	3.5814	7.9831	4.7187	3.581
25	159.0000	138.0000	38.0000	38.0000	7.9831	4.7019	3.6430	7.9831	4.7019	3.643
26	163.8000	126.0000	40.0000	42.0000	8.0355	4.6820	3.6642	8.0355	4.6820	3.664
27	157.5000	111.0000	48.0000	51.0000	8.5774	4.1547	5.4537	8.5774	4.1547	5.453
28	160.6000	146.0000	38.0000	37.0000	7.9831	4.6639	3.8382	7.9831	4.6639	3.838
29	154.0000	116.0000	44.0000	48.0000	8.3618	4.4147	4.3748	8.3618	4.4147	4.374
30	160.1000	115.0000	46.0000	50.0000	8.4142	4.3107	4.8839	8.4142	4.3107	4.883
31	157.8000	141.0000	38.0000	38.0000	7.9831	4.7019	3.6042	7.9831	4.7019	3.604
32	154.5000	151.0000	38.0000	36.0000	7.9831	4.6259	4.0722	7.9831	4.6259	4.072

Figure 2.10: *Molecular Descriptors* document with descriptor values

On the right there are radio buttons that determine whether descriptors should be calculated for all molecules in the *Molecular Descriptors* document or for selected molecules only. Using the *Messages* check box error messages can be disabled. There are further buttons for searching indices by their name, saving descriptor selections and opening previously saved selections.

3. Click on the *Start* button to start descriptor calculation.

4. When the calculation is finished click *OK* to return to the *Molecular Descriptors* document. After descriptor calculation, descriptor values will appear in additional columns (Figure 2.10).

2.3.2 Calculating Substructure Counts

A second type of molecular descriptors are *substructure counts*. A substructure is a part of the hydrogen–suppressed molecular graph. The substructure procedure implemented in MOLGEN–QSPR systematically finds all substructures up to a certain size that occur in a molecular library and counts their occurrences in all molecules in the library. For example, in 2–fluorobutane, $H_3C-CHF-CH_2-CH_3$, the substructures F, $C-F$, $C-C-F$, $C-C-C-F$, and $C-C(-F)-C$ will automatically be retrieved and counted, along with fluorine–free substructures.

Starting from the *Molecular Descriptors* document

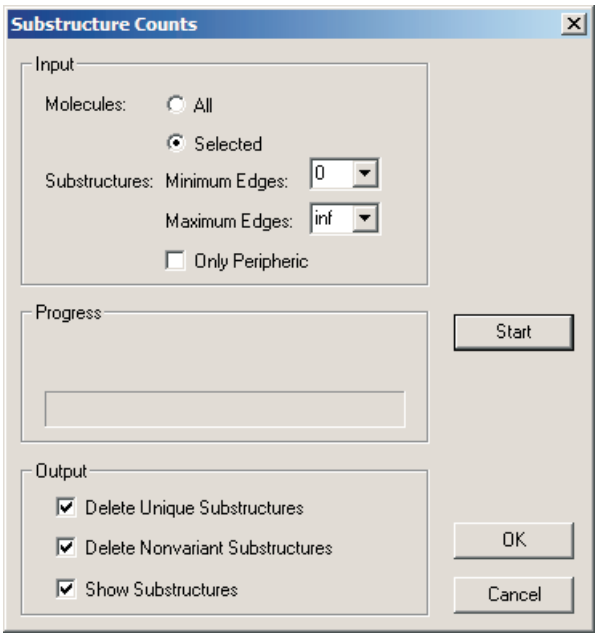


Figure 2.11: *Substructure Counts* dialogue

1. call *File|Substructure Counts* to obtain the *Substructure Counts* dialogue (Figure 2.11).

2. In the *Minimum/Maximum Edges* combo boxes specify the lower and upper number of edges for the substructures to be retrieved.

3. Click the *Start* button to start the calculation.

4. After the calculation is finished you can decide to ignore unique and/or nonvariant substructures by the check boxes in the *Output* field. Activate the *Show Substructures* check box if you want to create a new *Molecule* document with the retrieved substructures.

5. Press *OK* to add the substructure counts to the *Molecular Descriptors* document.

2.3.3 Calculating Fragment Counts

Fragment counts are a third type of molecular descriptors: A fragment is defined by the user. A fragment may contain hydrogen atoms, so it is a part of the hydrogen–containing molecular graph. Thus, in $H_3C-CHF-CH_2-CH_3$, 2–fluorobutane, $H-C-F$, H_3C-CHF etc. are fragments, they will be retrieved and counted only when defined and searched as such.

To calculate fragment counts do the following:

1. Use *File|New|Moled* to edit the fragment of interest (Figure 2.12)

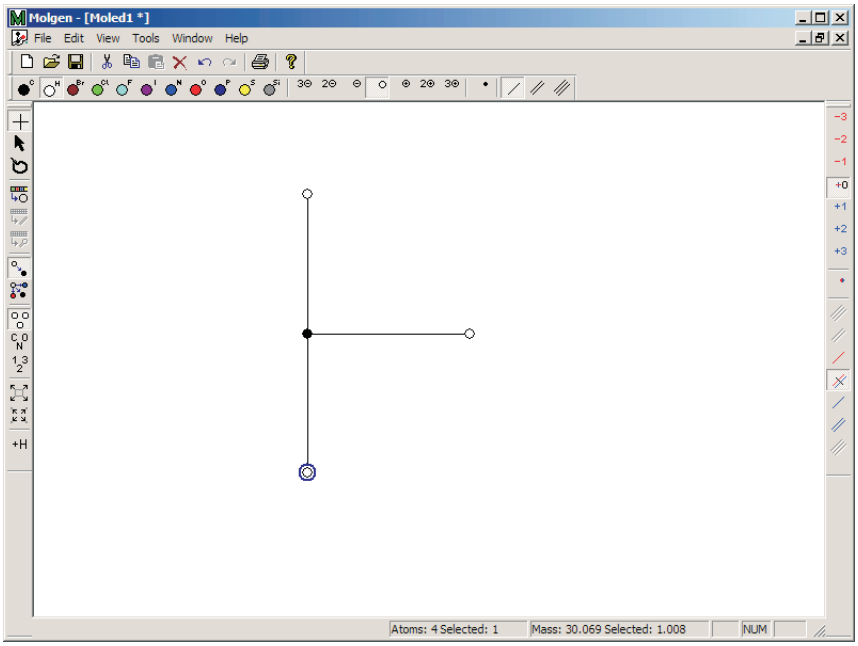


Figure 2.12: *Moled* document

2. Name the fragment by means of *Edit|Properties*. The *Fragment Property* sheet (Figure 2.13) appears. Enter the desired name and press *OK*.

3. Switch back to your *Molecular Descriptors* document using the *Window* submenu or clicking on the *Molecular Descriptors* document's window.

4. Call the *Fragment Counts* dialogue (Figure 2.14) by *File|Fragment Counts*.

5. Add fragments using the *Add* button. In the following dialogue (Figure 2.15) you can select fragments from opened *Moled* documents.

6. Once you have selected one or more fragments start the calculation using the *Start* button.

7. After the calculation is finished you can decide to ignore unique and/or nonvariant fragments by the check boxes in the *Output* field.

8. Press *OK* to add the fragment counts to the *Molecular Descriptors* document.

Our example fragment Methyl counts CH_3 groups, whereas the substructure count for C is the occurrence number of C atoms, i.e. the sum of occurrences of CH_3, CH_2, CH groups, and C atoms without H.

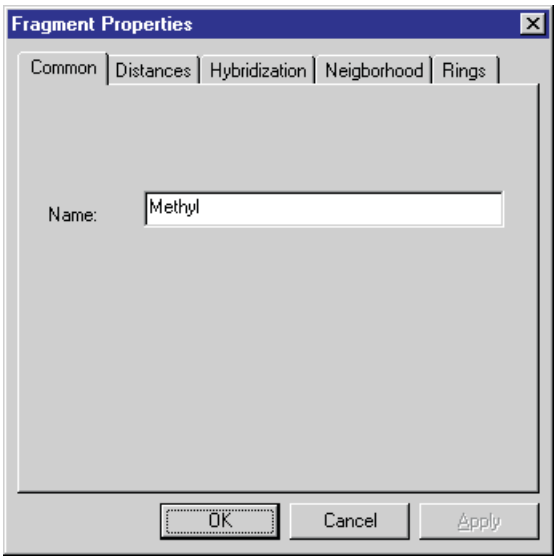


Figure 2.13: *Fragment Properties Common* page

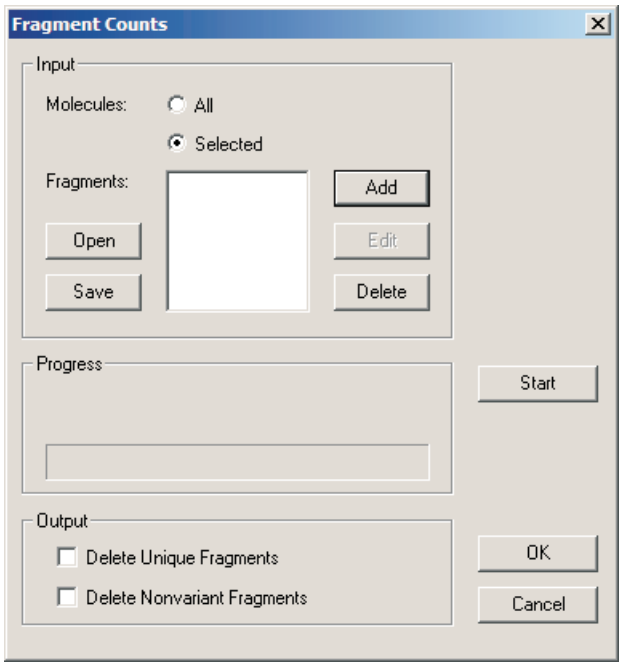


Figure 2.14: *Fragment Counts* dialogue

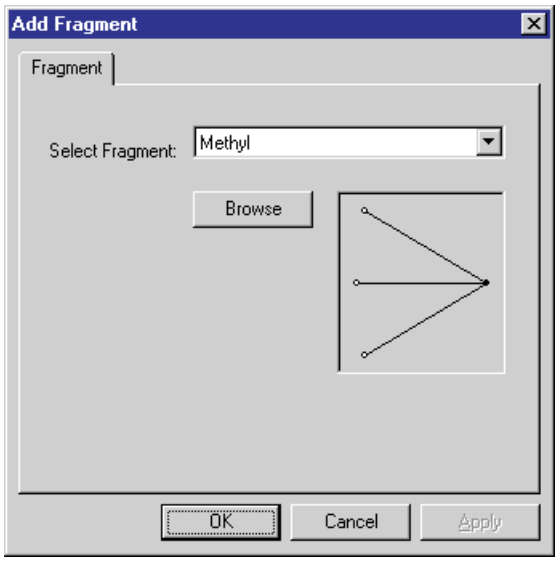


Figure 2.15: *Add Fragment* dialogue

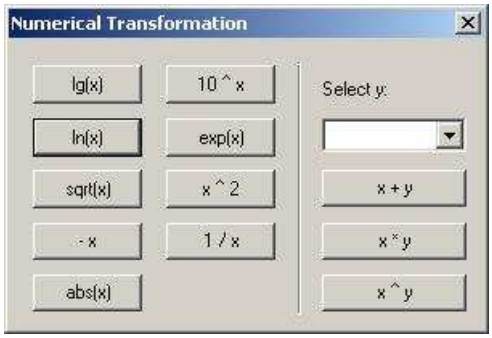


Figure 2.16: *Transform column* dialogue

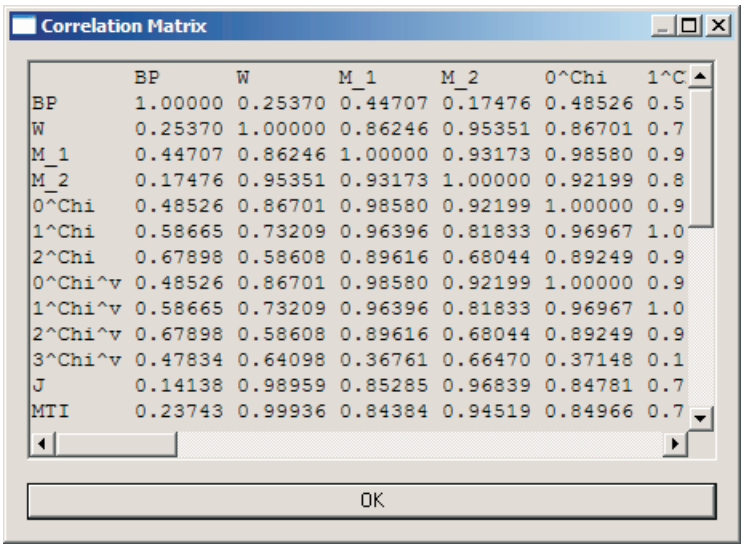
2.3.4 Descriptor Transformation

If you need a somewhat more complex variant of a descriptor already present, such as the reciprocal, square, square root, logarithm, or a sum or product etc. of two descriptors already present, use *Edit*| *Transform Column* (see Figure 2.16).

A transformation chosen here works on the current column.

2.4 Correlation Analysis

In order to select descriptors for a QSPR study it might be useful to initially analyse property–descriptor and descriptor–descriptor correlations.



Correlation Matrix

	BP	W	M_1	M_2	0^Chi	1^C
BP	1.00000	0.25370	0.44707	0.17476	0.48526	0.5
W	0.25370	1.00000	0.86246	0.95351	0.86701	0.7
M_1	0.44707	0.86246	1.00000	0.93173	0.98580	0.9
M_2	0.17476	0.95351	0.93173	1.00000	0.92199	0.8
0^Chi	0.48526	0.86701	0.98580	0.92199	1.00000	0.9
1^Chi	0.58665	0.73209	0.96396	0.81833	0.96967	1.0
2^Chi	0.67898	0.58608	0.89616	0.68044	0.89249	0.9
0^Chi^v	0.48526	0.86701	0.98580	0.92199	1.00000	0.9
1^Chi^v	0.58665	0.73209	0.96396	0.81833	0.96967	1.0
2^Chi^v	0.67898	0.58608	0.89616	0.68044	0.89249	0.9
3^Chi^v	0.47834	0.64098	0.36761	0.66470	0.37148	0.1
J	0.14138	0.98959	0.85285	0.96839	0.84781	0.7
MTI	0.23743	0.99936	0.84384	0.94519	0.84966	0.7

OK

Figure 2.17: *Correlation Matrix* dialogue

2.4.1 Calculating the Correlation Matrix

To obtain the correlation matrix of all variables (properties, descriptors, residuals, predictions) choose *View|Correlations*. A window will appear showing the matrix of absolute correlation coefficients (Figure 2.17).

Often a *Molecular Descriptor* document will contain many columns, say several hundred. In such cases it is advisable to calculate the correlation matrix for a small subtable only. Editing the table is described in Section 2.2.3. In order not to lose data edit a copy of your table rather than the table itself.

Missing values (N/A) will prohibit the correlation matrix calculation, so make sure to exclude a column or row containing missing values (see Section 2.2.3).

For a visualisation of intercorrelations use the scatterplot feature.

2.4.2 Displaying Correlations

Using *View|Scatterplot* you can change the *Molecular Descriptors* document to be displayed as scatterplot (Figure 2.18).

Using the upper left combo boxes select a variable for the x and one for the y axis. Again use the mouse to select and display certain subsets of structures.

Note: You may plot any column in the table (property, descriptor, residual, prediction) *vs* any other column.

To return to the table display use *View|Scatterplot* again.

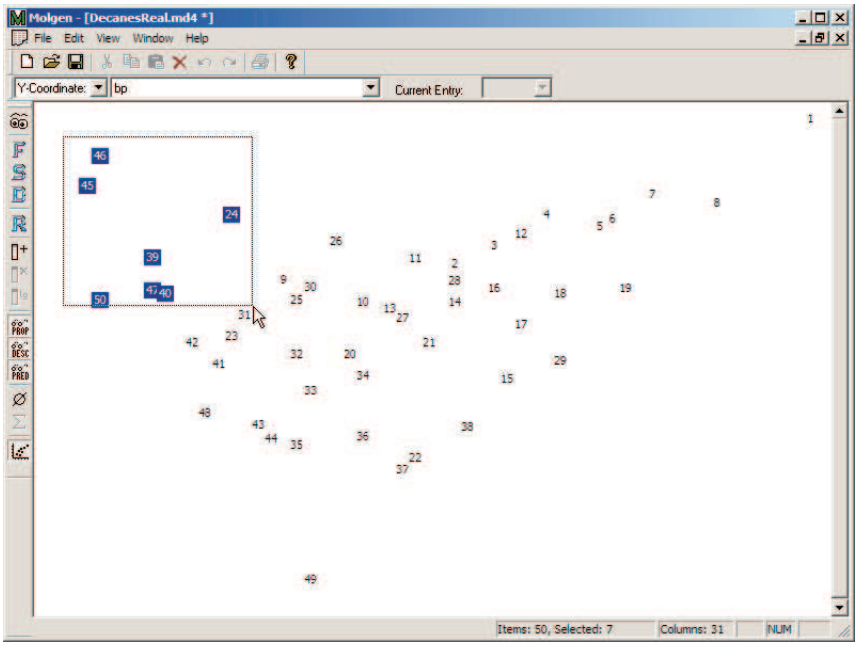


Figure 2.18: *Molecular Descriptors* document displayed as scatterplot

2.5 Regression Analysis

The most important feature of MOLGEN–QSPR is the ability to calculate quantitative structure property relationships. Use *File|Regression* to get to the *Regression* dialogue (Figure 2.19).

Before we start the regression analysis several settings concerning variables, preprocessing and regression method have to be specified. Therefore press the *Setup* button. You receive the *Regression Setup* sheet.

2.5.1 Variable Selection

Click on the *Variables* tabulator field in order to define the dependent and independent variables (Figure 2.20).

The dependent variable is chosen with the *Target Variable* combo box. Independent variables are selected with the check boxes in the *Regressors* field.

2.5.2 Regression Preprocessing

Go to the *Preprocessing* tabulator field in order to define scaling and/or centering methods for the dependent/independent variables (Figure 2.21).

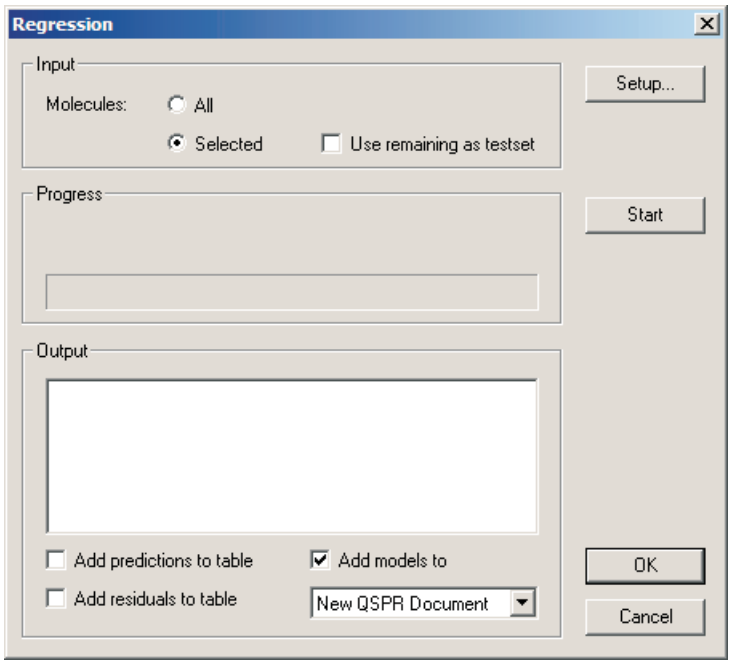


Figure 2.19: *Regression* dialogue

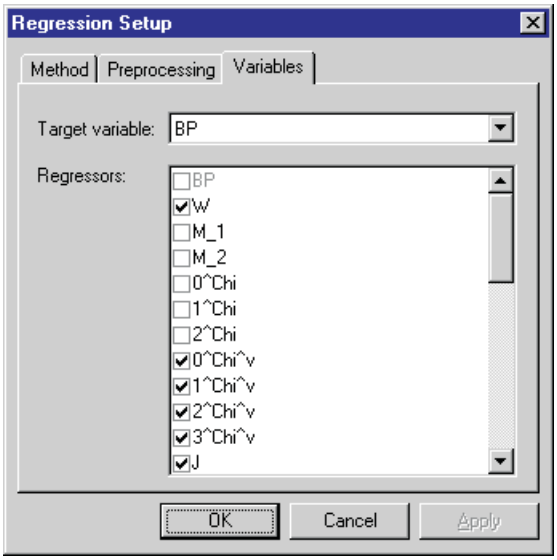


Figure 2.20: *Regression Variables* page

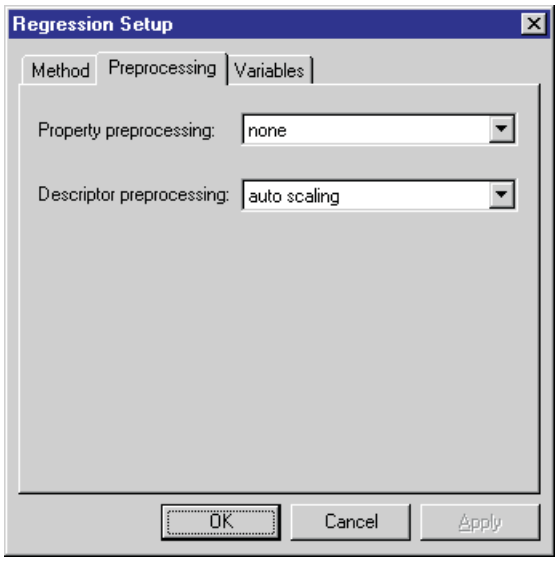


Figure 2.21: *Regression Preprocessing* page

For both kinds of variables there are five types of preprocessing available:

- none,
- centering, the shift of variable values by their arithmetic mean,
- range scaling, transforming the variable values in such a way that they range from 0 to 1,
- auto scaling, transforming the variable values in such a way that they have mean 0 and variance 1,
- normalization which divides the variable values by their euclidean norm, i.e. after transformation they have euclidean norm 1.

All these preprocessings are linear transformations. As such, they do not influence least squares regression and regression trees. However, for neural networks, support vector machines and nearest neighbor regression, variable preprocessing may have an important impact on model quality.

If such a transformation is applied, it is automatically reversed in a final step.

2.5.3 Regression Method

Clicking on the *Method* tabulator field you obtain a page for setting up the regression method (Figure 2.22).

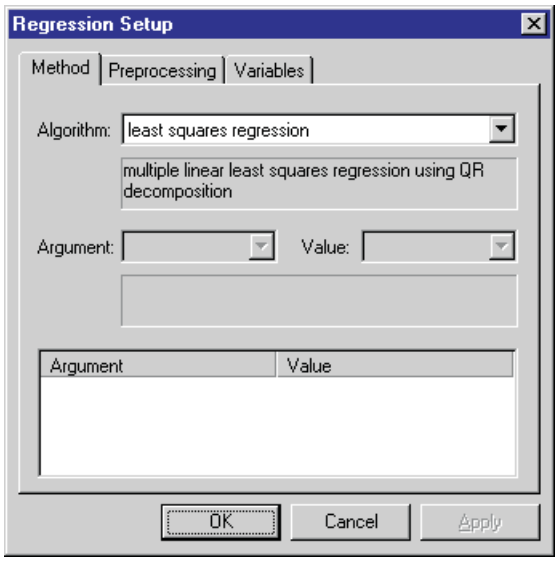


Figure 2.22: *Regression Method* page

Use the *Algorithm* combo box in order to select the regression algorithm to be applied. There are various algorithms available, among them

- least squares regression,
- regression trees,
- neural networks,
- support vector machines and
- nearest neighbor regression.

Note: In order to use regression trees, neural networks, or support vector machines, the statistics software *R* must be installed (cf. Section 1.1.2).

For the (ordinary) least squares regression no further arguments are required. Often you will use the best subset regression (Figure 2.23).

Using the *Argument* and *Value* combo boxes parameters for the regression algorithm can be defined. A short description of the algorithm and the argument is displayed.

2.5.4 Starting the QSPR Calculation

After regression setup is completed close the *Regression Setup* sheet with *OK* and start the regression algorithm by clicking the *Start* button. After a while the regression analysis will be finished and results will be displayed in the *Output* field (Figure 2.24).

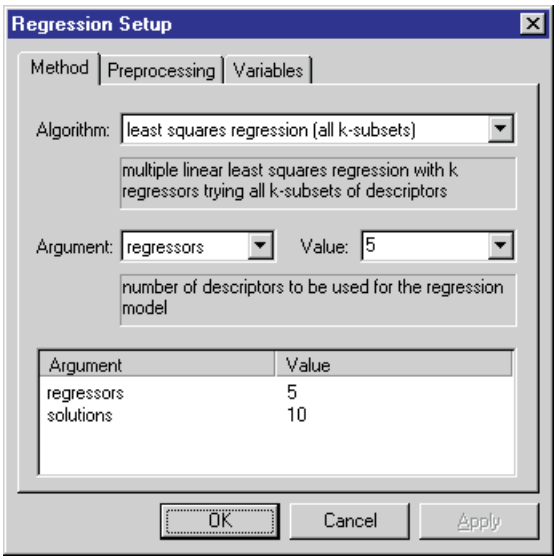


Figure 2.23: *Regression Method* page for best subset regression

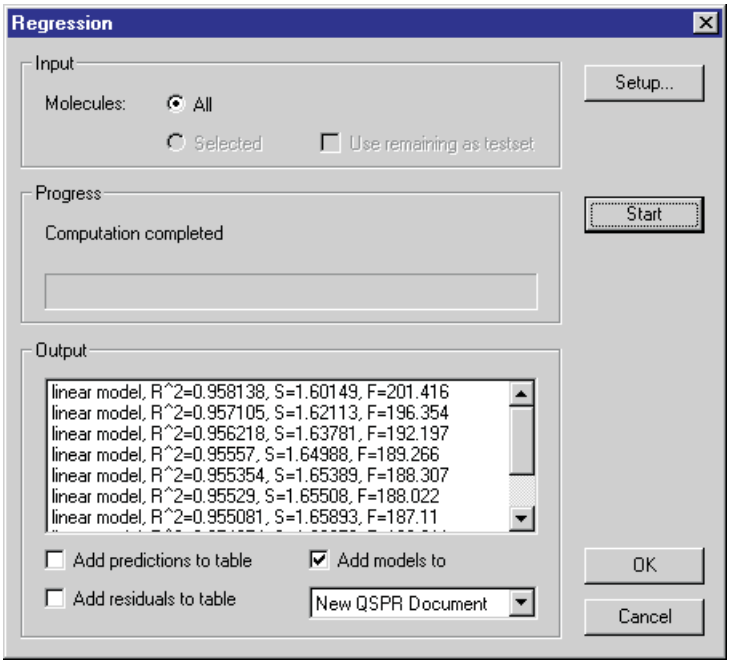


Figure 2.24: *Regression* dialogue with results in the *Output* field

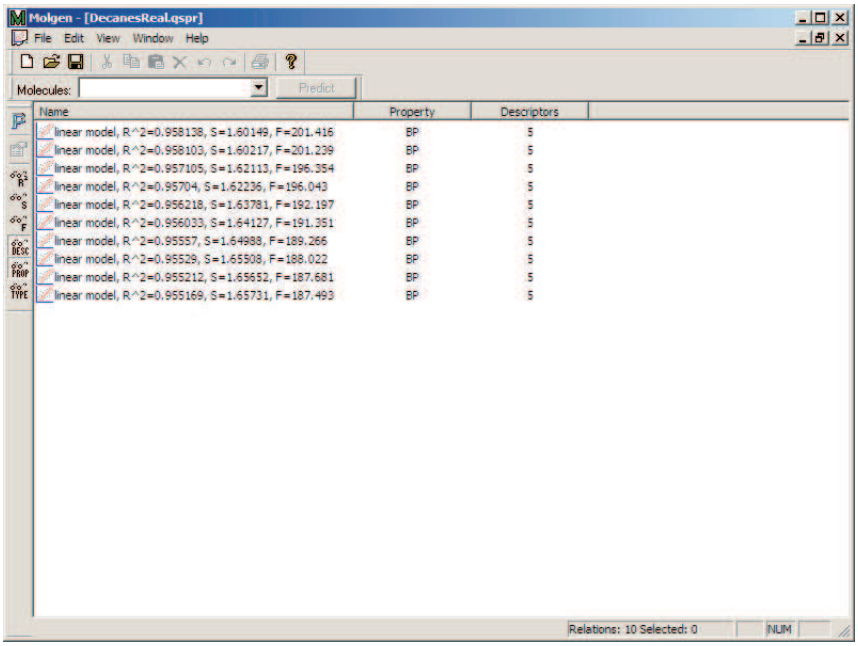


Figure 2.25: *QSPR* document

In the *Output* field you see the best QSPRs calculated (one in each row). Doubleclick on a certain QSPR to obtain further details on the selected QSPR. Use the *Add Predictons/Residuals* check boxes to add values calculated by the QSPR and/or residuals as new column(s) to the *Molecular Descriptors* document. If the *Add Models* check box is activated, QSPRs are added to a new or an existing *QSPR* document specified by the lower combo box.

2.6 Displaying and Saving QSPRs

If you decided to add models to a new QSPR document, the screen could look as shown in Figure 2.25.

In a *QSPR* document different types of QSPRs for different properties using different descriptors and algorithms can be stored. Use *File|Save As* in order to save the *QSPR* document (extension *.qspr*). With the *View* submenu you can add/hide columns with certain characteristics of the QSPRs such as

- model type,
- property name,
- number of descriptors,
- degrees of freedom,

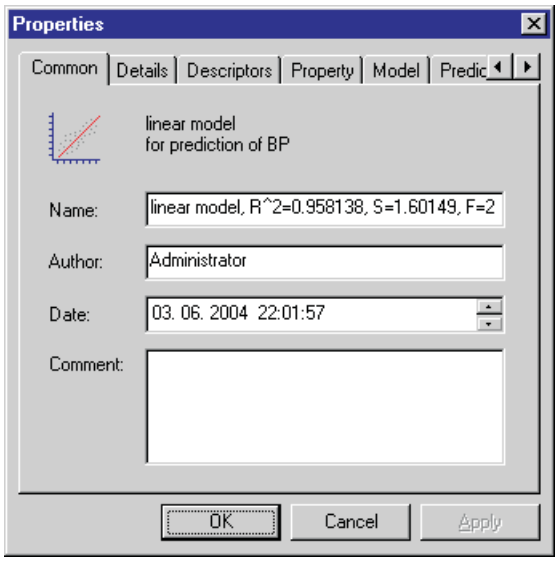


Figure 2.26: *QSPR Common* page

- number of observations,
- R squared,
- standard error,
- Fisher's F value,
- residual sum of squares,
- mean squared residual,
- mean absolute residual,
- maximum absolute residual etc.

Doubleclick on a certain QSPR to get the QSPR's property sheet (Figures 2.26–2.32).

2.6.1 QSPR Common Properties

On the *Common* page you are given the information shown in Figure 2.26. This information can be edited and stored using the *OK* button.

2.6.2 QSPR Details

Statistical details are supplied on the *Details* page (Figure 2.27).

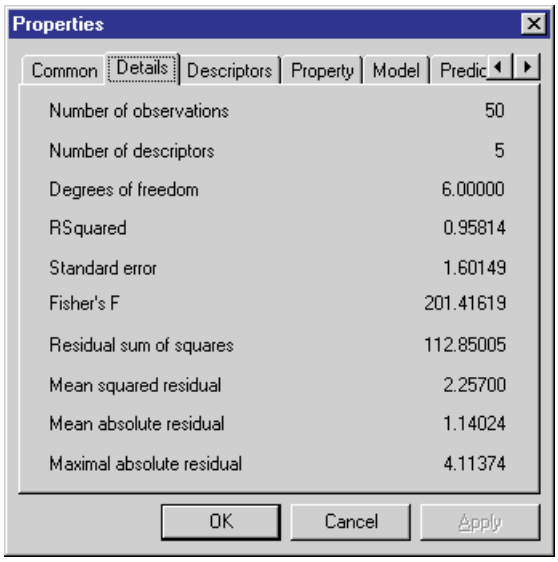


Figure 2.27: *QSPR Details* page

2.6.3 QSPR Descriptors

Names and types of descriptors as well as preprocessing transformations can be seen on the *Descriptors* page (Figure 2.28).

2.6.4 QSPR Property

The property investigated by the QSPR is noted on the *Property* page (Figure 2.29).

2.6.5 QSPR Model

The specification of the prediction function is provided on the *Model* page (Figure 2.30).

2.6.6 QSPR Predictions

The *Prediction* page offers a table of residuals, experimental and calculated values (Figure 2.31).

Note: Use the left mouse button and *Copy* in order to copy the complete table to the clipboard.

2.6.7 QSPR Plot

The *Plot* page shows a plot of experimental *vs* calculated values (Figure 2.32).

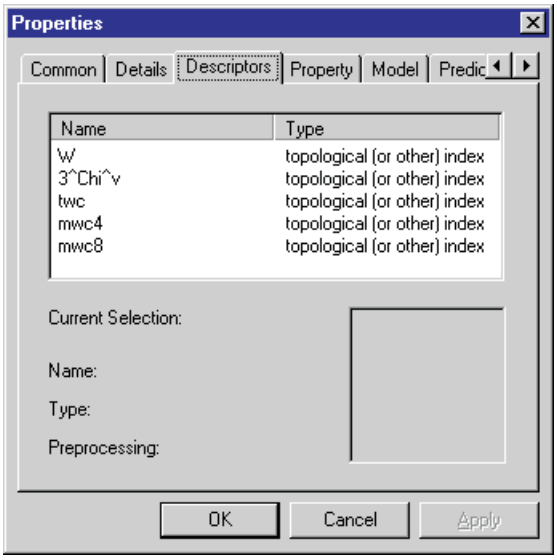


Figure 2.28: *QSPR Descriptors* page

Figure 2.29: *QSPR Property* page

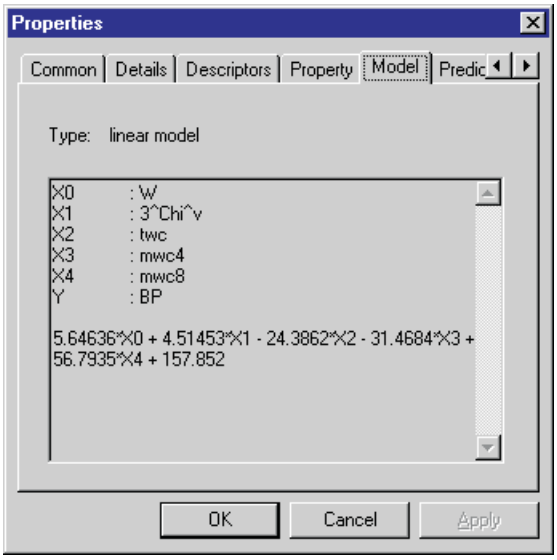


Figure 2.30: *QSPR Model* page

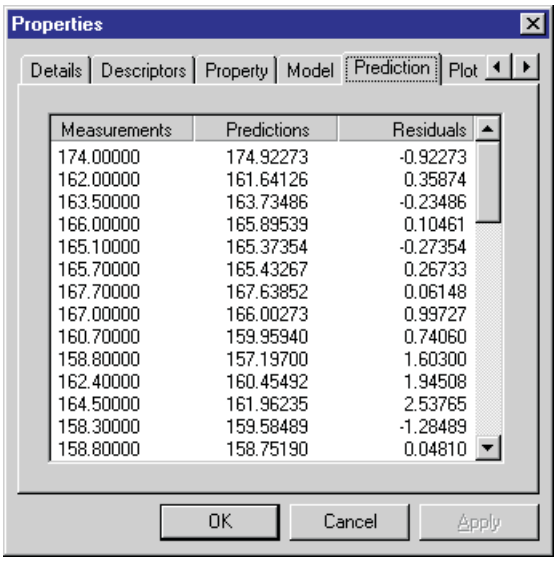


Figure 2.31: *QSPR Prediction* page

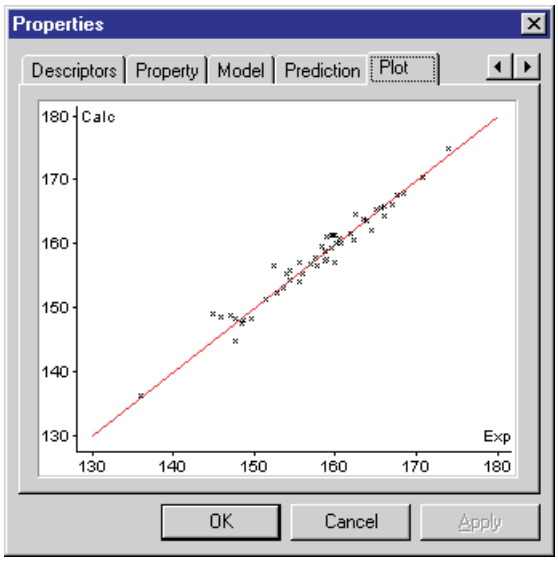


Figure 2.32: *QSPR Plot* page

2.7 Validation

2.7.1 LOO Crossvalidation

As a first validation step for our best QSPR equation let us perform a leave-one-out crossvalidation. Open a .md4 and the corresponding .qspr document containing at least one model, switch to the .md4 document and click *Crossvalidation* in the *View* menu. A page similar to the QSPR Details page will be displayed showing inter alia the values of R^2_{cv} and S_{cv}, see Figure 2.33. Missing values (N/A) will prohibit the crossvalidation calculation, so make sure to exclude rows/columns containing missing values (see Section 2.2.3).

As a necessary (but not sufficient) condition for a valid QSPR equation, the crossvalidation results (R^2_{cv}, S_{cv}, plot) should be only moderately worse than the original ones, compare Figures 2.33 and 2.34 to Figures 2.27 and 2.32, respectively.

2.7.2 Further Validation

As a rule, a particular QSPR model needs further validation before it can be considered reliable. Since various validation methods are in use or recommended by various authors, no corresponding procedures are installed as black boxes in MOLGEN–QSPR . There are, however, a number of features that may be helpful in validation, such as

- Random column,
- Random selection,

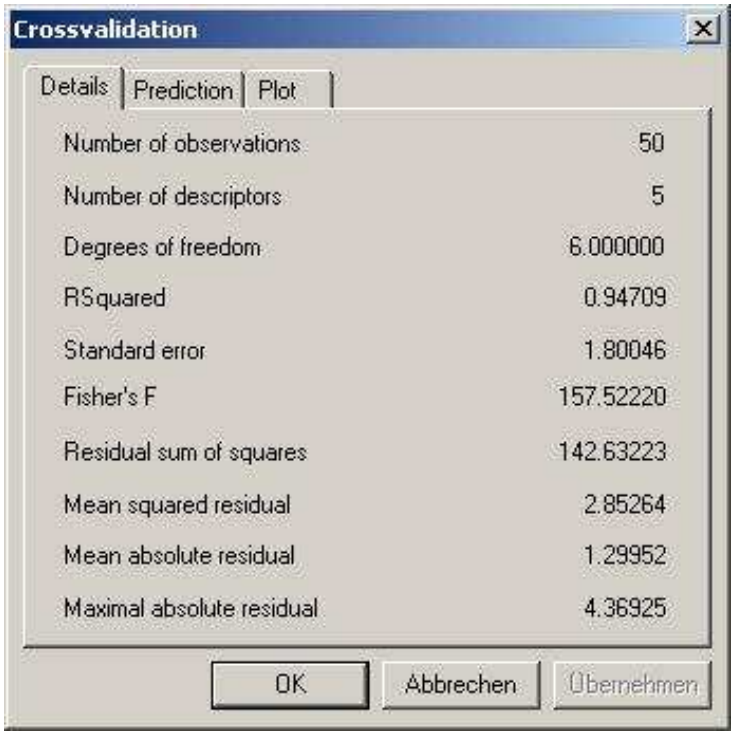


Figure 2.33: *Leave-one-out Crossvalidation Details* page

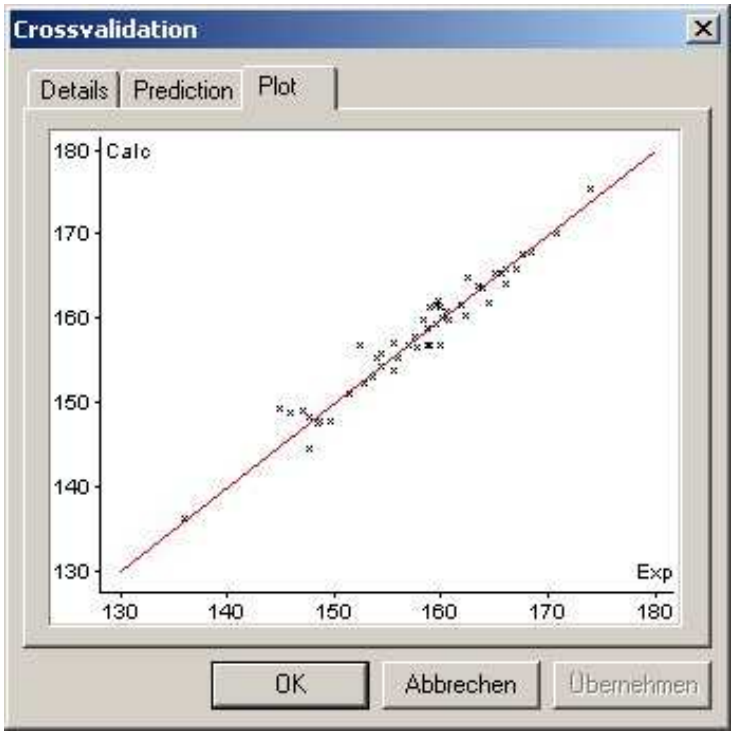


Figure 2.34: *Leave-one-out Crossvalidation Plot* page

- Invert selection,
- Learning set / Test set partition.

2.8 Property Prediction

Let us now apply our best QSPR to predict the boiling points of all those decanes not included in our real library.

2.8.1 Generating a Virtual Library

Therefore we generate all decanes, i.e. structural formulas to the molecular formula $C_{10}H_{22}$.

1. Create a new *Molgen* document using *File|New|Molgen*.
2. Use *Edit|Add|Formula* to call the *Add Molecular Formula* sheet.
3. Enter $C_{10}H_{22}$ in the *Formula* field.
4. Click *OK* to add the molecular formula to the *Molgen* document.
5. Use *File|Save As* to save the *Molgen* document with name *Decanes.mgp*.
6. Start structure generation using *Start* in the *Generator* field.
7. After a moment the computation will be completed resulting in 75 constitutional isomers.
8. Select *File|Open Output* to display the generated structures.

Note: Often virtual libraries cannot be described as isomers of a molecular formula. Rather, particularly in combinatorial chemistry virtual libraries are specified by reactants and reactions. Such libraries can be generated using the reaction-based structure generator MOLGEN–COMB.

2.8.2 Comparing Real and Virtual Library

Now having generated all decanes we want to identify those not included in our real library of 50 decanes with known boiling points. Starting from the *Molecule* document *Decanes.mb4* click *File|Compare* to get to the *Compare Molecule Files* dialogue (Figure 2.35).

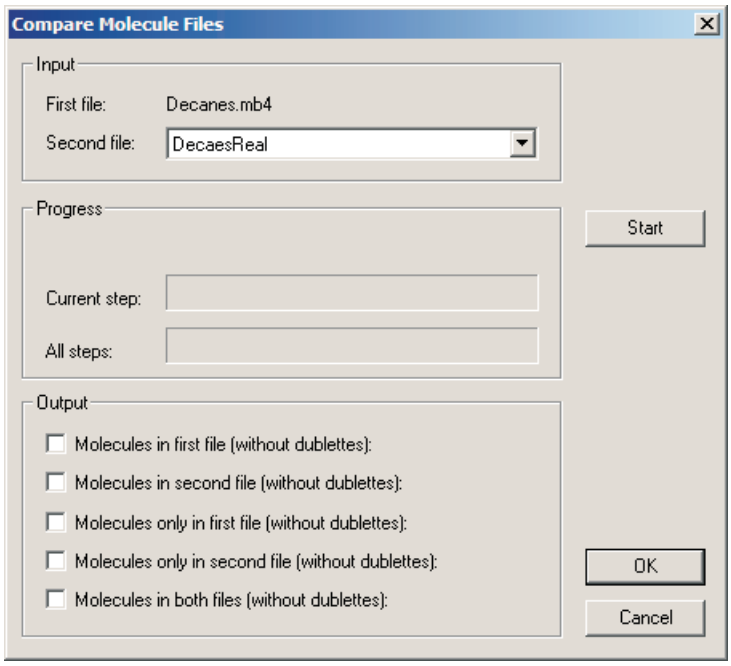


Figure 2.35: *Compare Molecule Files* dialogue

Select *DecanesReal* in the *Second File* combo box and click *Start* to start the comparison of the two *Molecule* documents. The program will answer in the *Output* field (Figure 2.36). As we are interested in structures occuring only in *Decanes* and not in *DecanesReal*,

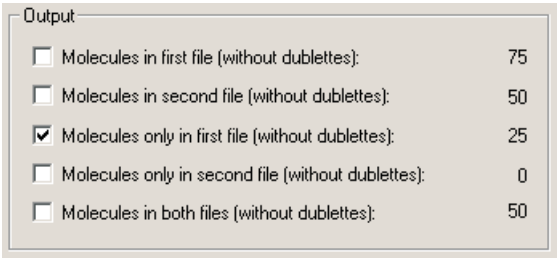


Figure 2.36: *Compare Molecule Files* output

we activate the corresponding check box. After pressing *OK* a new *Molecule* document appears, named *Decanes–DecanesReal* and containing the 25 decanes not included in *DecanesReal*.

2.8.3 Applying QSPRs for Prediction

In order to predict property values we have to switch back to the *QSPR* document. Now select the QSPRs you want to use for prediction. On *File|Prediction* the *Prediction* dialogue appears (Figure 2.37).

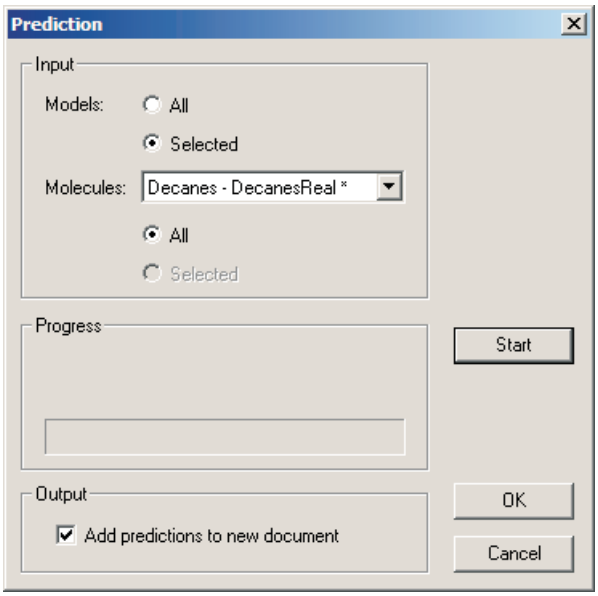


Figure 2.37: *Prediction* dialogue

Select *Decanes–DecanesReal* in the *Molecules* combo box and click the *Start* button. After the computation is finished press *OK*, and the 25 predicted property values will appear in a new *Molecular Descriptors* document, see Figure 2.38.

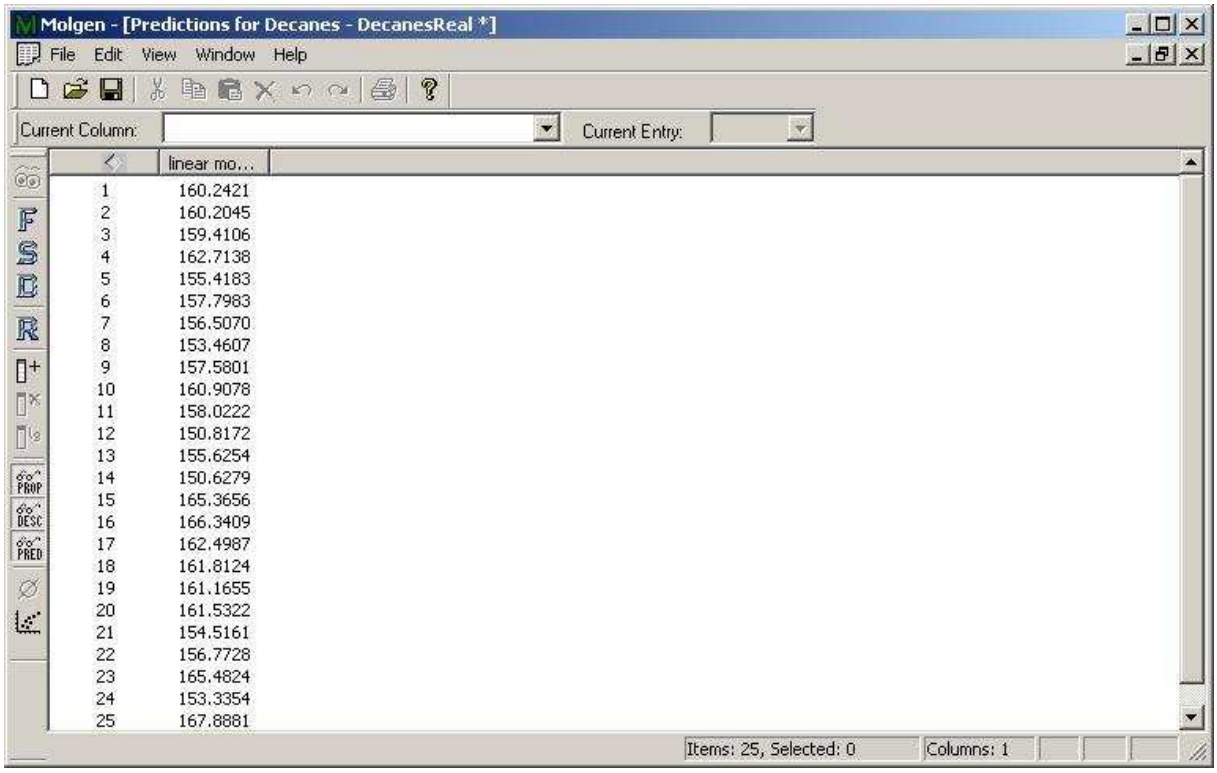


Figure 2.38: *Prediction Result* page

Chapter 3

The Molecular Descriptors

3.1 Arithmetic Indices

$A,\ A\,(incl.\,H)$	number of atoms, number of atoms (incl. H atoms)
$N_H,\ rel.\,N_H$	number of H atoms, relative number of H atoms
$N_C,\ rel.\,N_C$	number of C atoms, relative number of C atoms
$N_O,\ rel.\,N_O$	number of O atoms, relative number of O atoms
$N_N,\ rel.\,N_N$	number of N atoms, relative number of N atoms
$N_S,\ rel.\,N_S$	number of S atoms, relative number of S atoms
$N_F,\ rel.\,N_F$	number of F atoms, relative number of F atoms
$N_{Cl},\ rel.\,N_{Cl}$	number of Cl atoms, relative number of Cl atoms
$N_{Br},\ rel.\,N_{Br}$	number of Br atoms, relative number of Br atoms
$N_I,\ rel.\,N_I$	number of I atoms, relative number of I atoms
$N_P,\ rel.\,N_P$	number of P atoms, relative number of P atoms
$B,\ B\,(incl.\,H)$	number of bonds, number of bonds (incl. H atoms)
$loc.\,B$	number of localized bonding electron pairs
$loc.\,B\,(incl.\,H)$	number of localized bonding electron pairs (incl. H atoms)
$n-,\ rel.\,n-$	number of single bonds, relative number of single bonds
$n-\ (incl.\,H)$	number of single bonds (incl. H atoms)
$rel.\,n-\ (incl.\,H)$	relative number of single bonds (incl. H atoms)
$n=,\ rel.\,n=$	number of double bonds, relative number of double bonds
$rel.\,n=\ (incl.\,H)$	relative number of double bonds (incl. H atoms)
$n\#,\ rel.\,n\#$	number of triple bonds, relative number of triple bonds
$rel.\,n\#\,(incl.\ H)$	relative number of triple bonds (incl. H atoms)
$n_{aroma},\ rel.\,n_{aroma}$	number of aromatic bonds, relative number of aromatic bonds
$rel.\,n_{aroma}\,(incl.\,H)$	relative number of aromatic bonds (incl. H atoms)
C	cyclomatic number

$MW,\ mean\ AW$	molecular weight, mean atomic weight
$MW\,(incl.\,H)$	molecular weight (incl. H atoms)
$mean\ AW\,(incl.\,H)$	mean atomic weight (incl. H atoms)
cha	total charge
rad	number of radical centers
HBD	number of hydrogen bond donors
HBA	number of hydrogen bond acceptors
$N\_charged$	number of charged atoms
$mass\_exact,\ mass\_int$	Monoisotopic mass (exact and integer)

3.2 Topological Indices

W	Wiener index
$M_1,\ M_2$	1st, 2nd Zagreb index
${}^mM_1,\ {}^mM_2$	1st, 2nd modified Zagreb index
${}^0\chi,\ {}^1\chi,\ {}^2\chi$	Randic indices of orders 0,1,2
${}^0\chi^s,\ {}^1\chi^s,\ {}^2\chi^s,\ {}^3\chi^s$	solvation connectivity indices of orders 0,1,2,3
${}^3\chi^s_c$	solvation connectivity index for clusters
${}^0\chi^v,\ {}^1\chi^v,\ {}^2\chi^v,\ {}^3\chi^v$	Kier and Hall (valence connectivity) indices of orders 0,1,2,3
${}^1\kappa,\ {}^2\kappa,\ {}^3\kappa$	Kier shape indices 1,2,3
$\Phi_{\bar{\alpha}}$	Kier molecular flexibility index non–alpha–modified
${}^1\kappa_\alpha,\ {}^2\kappa_\alpha,\ {}^3\kappa_\alpha$	Kier alpha–modified shape indices 1,2,3
Φ	Kier molecular flexibility index
F	Platt number
N_{GS}	Gordon–Scantlebury index
$J,\ J_{unsat}$	Balaban index, unsaturated Balaban index
MTI	Schultz molecular topological index
MTI'	MTI' index
H	Harary number
twc	total walk count
$mwc^{(2)},\ldots,mwc^{(8)}$	molecular walk counts of length 2,...,8
twc_{unsat}	unsaturated total walk count
$mwc^{(2)}_{unsat},\ldots,mwc^{(8)}_{unsat}$	unsaturated molecular walk counts of length 2,...,8
$G_1\,(topol.)$	gravitational index (pairs, topol. dist.)
$G_1\,(topol.,incl.\,H)$	gravitational index (pairs, topol. dist., incl. H atoms)
$G_2\,(topol.)$	gravitational index (bonds, topol. dist.)
$G_2\,(topol.,incl.\,H)$	gravitational index (bonds, topol. dist., incl. H atoms)

Z	Hosoya Z–index
IC_0, IC_1, IC_2	Basak information content of order 0,1,2
TIC_0, TIC_1, TIC_2	Basak total information content of order 0,1,2
CIC_0, CIC_1, CIC_2	Basak complementary information content of order 0,1,2
$N * CIC_0, \ldots, N * CIC_2$	total complementary information content of order 0,1,2
SIC_0, SIC_1, SIC_2	Basak structural information content of order 0,1,2
$N * SIC_0, \ldots, N * SIC_2$	total structural information content of order 0,1,2
BIC_0, BIC_1, BIC_2	bonding information content of order 0,1,2
$N * BIC_0, \ldots, N * BIC_2$	total bonding information content of order 0,1,2
MSD	mean square distance index
w, w_{diag}	detour index,detour index (incl. half main diagonal)
P_{acyc}	total acyclic path count
${}^2P_{acyc}, \ldots, {}^8P_{acyc}$	molecular acyclic path count of length 2,...,8
${}^{\geq 9}P_{acyc}$	molecular acyclic path count of length 9 and higher
P	total path count
${}^2P, \ldots, {}^8P$	molecular path count of length 2,...,8
${}^{\geq 9}P$	molecular path count of length 9 and higher
$rings$	total ring count
${}^3rings, \ldots, {}^8rings$	molecular ring count of length 3,...,8
${}^{\geq 9}rings$	molecular ring count of length 9 and higher
$ch.\ G_1, \ldots, ch.\ G_8$	topological charge index of order 1,...,8
$ch.\ J_1, \ldots, ch.\ J_8$	mean topological charge index of order 1,...,8
$ch.\ J[k]$	global topological charge index of order k
D	topological diameter
ξ^c	eccentric connectivity index
λ_1^A	principal eigenvalue of A
$SCA1$	sum of coefficents of principal eigenvector of A
$SCA2$	mean coefficent of principal eigenvector of A
$SCA3$	log of sum of coefficients of principal eigenvector of A
λ_1^D	principal eigenvalue of D
χ_T	total χ index
T_m	number of methyl groups
T_3	number of pairs of methyl groups at distance 3
FRB	freely rotatable bonds
SZD	Szeged index
SZD_P	hyper-Szeged index
${}^3\chi_p, \ldots, {}^6\chi_p$	connectivity index ${}^3\chi, \ldots, {}^6\chi$ path

${}^3\chi_c, \ldots, {}^6\chi_c$	connectivity index ${}^3\chi, \ldots, {}^6\chi$ cluster
${}^4\chi_{pc}, \ldots, {}^6\chi_{pc}$	connectivity index ${}^4\chi, \ldots, {}^6\chi$ path-cluster
${}^3\chi_{ch}, \ldots, {}^6\chi_{ch}$	connectivity index ${}^3\chi, \ldots, {}^6\chi$ chain
${}^3\chi_p^v, \ldots, {}^6\chi_p^v$	valence connectivity index ${}^3\chi^v, \ldots, {}^6\chi^v$ path
${}^3\chi_c^v, \ldots, {}^6\chi_c^v$	valence connectivity index ${}^3\chi^v, \ldots, {}^6\chi^v$ cluster
${}^4\chi_{pc}^v, \ldots, {}^6\chi_{pc}^v$	valence connectivity index ${}^4\chi^v, \ldots, {}^6\chi^v$ path-cluster
${}^3\chi_{ch}^v, \ldots, {}^6\chi_{ch}^v$	valence connectivity index ${}^3\chi^v, \ldots, {}^6\chi^v$ chain
sym_top	size of topological symmetry group
R	topological radius
con. comp.	number of connectivity components

3.3 Electrotopological and AI Indices

$S(sCH3)$	sum of E-states of sCH3
$S(dCH2)$, $S(ssCH2)$	sum of E-states of dCH2, sum of E-states of ssCH2
$S(tCH)$, $S(dsCH)$	sum of E-states of tCH, sum of E-states of dsCH
$S(aaCH)$, $S(sssCH)$	sum of E-states of aaCH, sum of E-states of sssCH
$S(ddC)$, $S(tsC)$	sum of E-states of ddC, sum of E-states of tsC
$S(dssC)$, $S(aasC)$	sum of E-states of dssC, sum of E-states of aasC
$S(aaaC)$, $S(ssssC)$	sum of E-states of aaaC, sum of E-states of ssssC
$S(sNH3)$, $S(sNH2)$	sum of E-states of sNH3, sum of E-states of sNH2
$S(ssNH2)$	sum of E-states of ssNH2
$S(dNH)$, $S(ssNH)$	sum of E-states of dNH, sum of E-states of ssNH,
$S(aaNH)$	sum of E-states of aaNH
$S(tN)$, $S(sssNH)$	sum of E-states of tN, sum of E-states of sssNH
$S(dsN)$, $S(aaN)$	sum of E-states of dsN, sum of E-states of aaN
$S(sssN)$, $S(ddsN)$	sum of E-states of sssN, sum of E-states of ddsN
$S(aasN)$, $S(ssssN)$	sum of E-states of aasN, sum of E-states of ssssN
$S(sOH)$, $S(dO)$	sum of E-states of sOH, sum of E-states of dO
$S(ssO)$, $S(aaO)$	sum of E-states of ssO, sum of E-states of aaO
$S(sF)$	sum of E-states of sF
$S(sPH2)$, $S(ssPH)$	sum of E-states of sPH2, sum of E-states of ssPH
$S(sssP)$, $S(dsssP)$	sum of E-states of sssP, sum od E-states of dsssP
$S(ssssP)$	sum of E-states of ssssP
$S(sSH)$	sum of E-states of sSH
$S(dS)$, $S(ssS)$	sum of E-states of dS, sum of E-states of ssS

$S(aaS)$, $S(dssS)$	sum of E-states of aaS, sum of E-states of dssS
$S(ddssS)$, $S(ssssssS)$	sum of E-states of ddssS, sum of E-states of ssssssS
$S(sCl)$	sum of E-states of sCl
$S(sSeH)$	sum of E-states of sSeH
$S(dSe)$, $S(ssSe)$	sum of E-states of dSe, sum of E-states of ssSe
$S(aaSe)$, $S(dssSe)$	sum of E-states of aaSe, sum of E-states of dssSe
$S(ddssSe)$	sum of E-states of ddssSe
$S(sBr)$	sum of E-states of sBr
$S(sI)$	sum of E-states of sI
$S(sLi)$	sum of E-states of sLi
$S(ssBe)$, $S(ssssBe)$	sum of E-states of ssBe, sum of E-states of ssssBe
$S(ssBH)$, $S(ssssB)$	sum of E-states of ssBH, sum of E-states of ssssB
$S(sSiH3)$, $S(ssSiH2)$	sum of E-states of sSiH3, sum of E-states of ssSiH2
$S(sssSiH)$, $S(ssssSi)$	sum of E-states of sssSiH, sum of E-states of ssssSi
$S(sGeH3)$, $S(ssGeH2)$	sum of E-states of sGeH3, sum of E-states of ssGeH2
$S(sssGeH)$, $S(ssssGe)$	sum of E-states of sssGeH, sum of E-states of ssssGe
$S(sAsH2)$, $S(ssAsH)$	sum of E-states of sAsH2, sum of E-states of ssAsH
$S(sssAs)$, $S(sssdAs)$	sum of E-states of sssAs, sum of E-states of sssdAs
$S(ssssAs)$	sum of E-states of ssssAs
$S(sSnH3)$, $S(ssSnH2)$	sum of E-states of sSnH3, sum of E-states of ssSnH2
$S(sssSnH)$, $S(ssssSn)$	sum of E-states of sssSnH, sum of E-states of ssssSn
$S(sPbH3)$, $S(ssPbH2)$	sum of E-states of sPbH3, sum of E-states of ssPbH2
$S(sssPbH)$, $S(ssssPb)$	sum of E-states of sssPbH, sum of E-states of ssssPb
$AI(sCH3)$	AI of sCH3
$AI(dCH2)$ $AI(ssCH2)$	AI of dCH2, AI of ssCH2
$AI(tCH)$, $AI(dsCH)$	AI of tCH, AI of dsCH
$AI(aaCH)$, $AI(sssCH)$	AI of aaCH, AI of sssCH
$AI(ddC)$, $AI(tsC)$	AI of ddC, AI of tsC
$AI(dssC)$, $AI(aasC)$	AI of dssC, AI of aasC
$AI(aaaC)$, $AI(ssssC)$	AI of aaaC, AI of ssssC
$AI(sNH3)$	AI of sNH3
$AI(sNH2)$, $AI(ssNH2)$	AI of sNH2, AI of ssNH2
$AI(dNH)$, $AI(ssNH)$	AI of dNH, AI of ssNH
$AI(aaNH)$, $AI(sssNH)$	AI of aaNH, AI of sssNH
$AI(tN)$, $AI(dsN)$	AI of tN, AI of dsN
$AI(aaN)$, $AI(sssN)$	AI of aaN, AI of sssN
$AI(ddsN)$, $AI(aasN)$	AI of ddsN, AI of aasN

$AI(ssssN)$	AI of ssssN
$AI(sOH)$	AI of sOH
$AI(dO),\ AI(ssO)$	AI of dO, AI of ssO
$AI(aaO)$	AI of aaO
$AI(sF)$	AI of sF
$AI(sPH2),\ AI(ssPH)$	AI of sPH2, AI of ssPH
$AI(sssP),\ AI(dsssP)$	AI of sssP, AI of dsssP
$AI(sssssP)$	AI of sssssP
$AI(sSH)$	AI of sSH
$AI(dS),\ AI(ssS)$	AI of dS, AI of ssS
$AI(aaS),\ AI(dssS)$	AI of aaS, AI of dssS
$AI(ddssS),\ AI(ssssssS)$	AI of ddssS, AI of ssssssS
$AI(sCl)$	AI of sCl
$AI(sSeH)$	AI of sSeH
$AI(dSe),\ AI(ssSe)$	AI of dSe, AI of ssSe
$AI(aaSe),\ AI(dssSe)$	AI of aaSe, AI of dssSe
$AI(ddssSe)$	AI of ddssSe
$AI(sBr)$	AI of sBr
$AI(sI)$	AI of sI
$AI(sLi)$	AI of sLi
$AI(ssBe),\ AI(ssssBe)$	AI of ssBe, AI of ssssBe
$AI(ssBH)$	AI of ssBH
$AI(sssB),\ AI(ssssB)$	AI of sssB, AI of ssssB
$AI(sSiH3),\ AI(ssSiH2)$	AI of sSiH3, AI of ssSiH2
$AI(sssSiH),\ AI(ssssSi)$	AI of sssSiH, AI of ssssSi
$AI(sGeH3),\ AI(ssGeH2)$	AI of sGeH3, AI of ssGeH2
$AI(sssGeH),\ AI(ssssGe)$	AI of sssGeH, AI of ssssGe
$AI(sAsH2),\ AI(ssAsH)$	AI of sAsH2, AI of ssAsH
$AI(sssAs),\ AI(sssdAs)$	AI of sssAs, AI of sssdAs
$AI(sssssAs)$	AI of sssssAs
$AI(sSnH3),\ AI(ssSnH2)$	AI of sSnH3, AI of ssSnH2
$AI(sssSnH)$	AI of sssSnH
$AI(ssssSn)$	AI of ssssSn
$AI(sPbH3),\ AI(ssPbH2)$	AI of sPbH3, AI of ssPbH2
$AI(sssPbH),\ AI(ssssPb)$	AI of sssPbH, AI of ssssPb
$Xu,\ Xu^m$	Xu index, modified Xu index

3.4 Geometrical Indices

$G_1,\ G_1\ (incl.\ H)$	gravitational index (pairs, 3D dist.)
$G_2,\ G_2\ (incl.\ H)$	gravitational index (bonds, 3D dist.)
$I_A,\ I_B,\ I_C$	principal moments of inertia A,B,C
$st.\ energy$	steric energy
$SHDW1, \ldots, 3$	XY shadow, XZ shadow, YZ shadow
$SHDW4, \ldots, 6$	standardized XY, XZ, YZ shadow
$SHDW1/SHDW2, \ldots$	XY/XZ, XY/YZ, XZ/YZ shadow
$ssSHDW1, \ldots, 3$	size sorted shadows 1,2,3
$ssSHDW4, \ldots, 6$	size sorted standardized shadows 1,2,3
$ssSHDW1/SHDW2, \ldots$	size sorted shadows 1/2,1/3,2/3
$V_{vdw},\ \rho_{vdw}$	Van der Waals volume, density by Van der Waals volume
V^s_{vdw}	standardized Van der Waals volume
V_{cub}	enclosing cuboid
S_{vdw}	Van der Waals surface
$SASA_{H_2O}$	solvent accessible surface area (H_2O)
$SASA_H$	solvent accessible surface area (H)
D_{3D}	geometrical diameter
V_{sphere}	enclosing sphere

3.5 Miscellaneous Indices

$slog\ P,\ sMR$	Crippen slog P, Crippen sMR
$at\ C01, \ldots, at\ C27$	Crippen atom types C01,..., C27
$at\ H01, \ldots, at\ H04$	Crippen atom types H01,...,H04
$at\ O01, \ldots, at\ O12$	Crippen atom types O01,...,O12
$at\ N01, \ldots, at\ N14$	Crippen atom types N01,...,N14
$at\ Hal,\ at\ Cl,\ at\ Br$	Crippen atom types Hal, Cl, Br
$at\ I,\ at\ F,\ at\ P$	Crippen atom types I, F, P
$at\ S01,\ at\ S02,\ at\ S03$	Crippen atom types S01, S02, S03
$at\ Me01,\ at\ Me02$	Crippen atom types Me01, Me02

3.6 Overall Indices

${}^{0-8}K$	sum of numbers of subgraphs of order 0 through 8
${}^{0}K, \ldots, {}^{8}K$	number of subgraphs of order 0, ..., 8
${}^{0}TC, \ldots, {}^{6}TC$	overall connectivity order $0, \ldots, 6$
TC	overall connectivity
${}^{1}TC^{*}, \ldots, {}^{6}TC^{*}$	overall connectivity subgraph order $1, \ldots, 6$
TC^{*}	overall connectivity subgraph
${}^{0}TC^{v}, \ldots, {}^{6}TC^{v}$	overall valence connectivity order $0, \ldots, 6$
TC^{v}	overall valence connectivity
${}^{0}TM_1, \ldots, {}^{6}TM_1$	overall first Zagreb order $0, \ldots, 6$
TM_1	overall first Zagreb
${}^{1}TM_1^{*}, \ldots, {}^{6}TM_1^{*}$	overall first Zagreb subgraph order $1, \ldots, 6$
TM_1^{*}	overall first Zagreb subgraph
${}^{1}TM_2, \ldots, {}^{6}TM_2$	overall second Zagreb order $1, \ldots, 6$
TM_2	overall second Zagreb
${}^{1}TM_2^{*}, \ldots, {}^{6}TM_2^{*}$	overall second Zagreb subgraph order $1, \ldots, 6$
TM_2^{*}	overall second Zagreb subgraph
${}^{1}TW, \ldots, {}^{6}TW$	overall Wiener order $1, \ldots, 6$
TW	overall Wiener
${}^{3}TC_p, \ldots, {}^{6}TC_p$	overall connectivity order $3, \ldots, 6$ path
TC_p	overall connectivity path
${}^{3}TC_p^{*}, \ldots, {}^{6}TC_p^{*}$	overall connectivity subgraph order $3, \ldots, 6$ path
TC_p^{*}	overall connectivity subgraph path
${}^{3}TC_p^{v}, \ldots, {}^{6}TC_p^{v}$	overall valence connectivity order $3, \ldots, 6$ path
TC_p^{v}	overall valence connectivity path
${}^{3}T(M_1)_p, \ldots, {}^{6}T(M_1)_p$	overall first Zagreb order $3, \ldots, 6$ path
$T(M_1)_p$	overall first Zagreb path
${}^{3}T(M_1)_p^{*}, \ldots, , {}^{6}T(M_1)_p^{*}$	overall first Zagreb subgraph order $3, \ldots, 6$ path
$T(M_1)_p^{*}$	overall first Zagreb subgraph path
${}^{3}T(M_2)_p, \ldots, {}^{6}T(M_2)_p$	overall second Zagreb order $3, \ldots, 6$ path
$T(M_2)_p$	overall second Zagreb path
${}^{3}T(M_2)_p^{*}, \ldots, {}^{6}T(M_2)_p^{*}$	overall second Zagreb subgraph order $3, \ldots, 6$ path
$T(M_2)_p^{*}$	overall second Zagreb subgraph path
${}^{3}TW_p, \ldots, {}^{6}TW_p$	overall Wiener order $3, \ldots, 6$ path
TW_p	overall Wiener path
${}^{3}TC_c, \ldots, {}^{6}TC_c$	overall connectivity order $3, \ldots, 6$ cluster
TC_c	overall connectivity cluster

${}^{3}TC_c^*, \ldots, {}^{6}TC_c^*$	overall connectivity subgraph order $3, \ldots, 6$ cluster
TC_c^*	overall connectivity subgraph cluster
${}^{3}TC_c^v, \ldots, {}^{6}TC_c^v$	overall valence connectivity order $3, \ldots, 6$ cluster
TC_c^v	overall valence connectivity cluster
${}^{3}T(M_1)_c, \ldots, {}^{6}T(M_1)_c$	overall first Zagreb order $3, \ldots, 6$ cluster
$T(M_1)_c$	overall first Zagreb cluster
${}^{3}T(M_1)_c^*, \ldots, {}^{6}T(M_1)_c^*$	overall first Zagreb subgraph order $3, \ldots, 6$ cluster
$T(M_1)_c^*$	overall first Zagreb subgraph cluster
${}^{3}T(M_2)_c, \ldots, {}^{6}T(M_2)_c$	overall second Zagreb order $3, \ldots, 6$ cluster
$T(M_2)_c$	overall second Zagreb cluster
${}^{3}T(M_2)_c^*, \ldots, {}^{6}T(M_2)_c^*$	overall second Zagreb subgraph order $3, \ldots, 6$ cluster
$T(M_2)_c^*$	overall second Zagreb subgraph cluster
${}^{3}TW_c, \ldots, {}^{6}TW_c$	overall Wiener order $3, \ldots, 6$ cluster
TW_c	overall Wiener cluster
${}^{4}TC_{pc}, \ldots, {}^{6}TC_{pc}$	overall connectivity order $4, \ldots, 6$ path-cluster
TC_{pc}	overall connectivity path-cluster
${}^{4}TC_{pc}^*, \ldots, {}^{6}TC_{pc}^*$	overall connectivity subgraph order $4, \ldots, 6$ path-cluster
TC_{pc}^*	overall connectivity subgraph path-cluster
${}^{4}TC_{pc}^v, \ldots, {}^{6}TC_{pc}^v$	overall valence connectivity order $4, \ldots, 6$ path-cluster
TC_{pc}^v	overall valence connectivity path-cluster
${}^{4}T(M_1)_{pc}, \ldots, {}^{6}T(M_1)_{pc}$	overall first Zagreb order $4, \ldots, 6$ path-cluster
$T(M_1)_{pc}$	overall first Zagreb path-cluster
${}^{4}T(M_1)_{pc}^*, \ldots, {}^{6}T(M_1)_{pc}^*$	overall first Zagreb subgraph order $4, \ldots, 6$ path-cluster
$T(M_1)_{pc}^*$	overall first Zagreb subgraph path-cluster
${}^{4}T(M_2)_{pc}, \ldots, {}^{6}T(M_2)_{pc}$	overall second Zagreb order $4, \ldots, 6$ path-cluster
$T(M_2)_{pc}$	overall second Zagreb path-cluster
${}^{4}T(M_2)_{pc}^*, \ldots, {}^{6}T(M_2)_{pc}^*$	overall second Zagreb subgraph order $4, \ldots, 6$ path-cluster
$T(M_2)_{pc}^*$	overall second Zagreb subgraph path-cluster
${}^{4}TW_{pc}, \ldots, {}^{6}TW_{pc}$	overall Wiener order $4, \ldots, 6$ path-cluster
TW_{pc}	overall Wiener path-cluster
${}^{3}TC_{ch}, \ldots, {}^{6}TC_{ch}$	overall connectivity order $3, \ldots, 6$ chain
TC_{ch}	overall connectivity chain
${}^{3}TC_{ch}^*, \ldots, {}^{6}TC_{ch}^*$	overall connectivity subgraph order $3, 6$ chain
TC_{ch}^*	overall connectivity subgraph chain
${}^{3}TC_{ch}^v, \ldots, {}^{6}TC_{ch}^v$	overall valence connectivity order $3, \ldots, 6$ chain
TC_{ch}^v	overall valence connectivity chain
${}^{3}T(M_1)_{ch}, \ldots, {}^{6}T(M_1)_{ch}$	overall first Zagreb order $3, \ldots, 6$ chain

$T(M_1)_{ch}$	overall first Zagreb chain
${}^3T(M_1)^*_{ch}, \ldots, {}^6T(M_1)^*_{ch}$	overall first Zagreb subgraph order $3, \ldots, 6$ chain
$T(M_1)^*_{ch}$	overall first Zagreb subgraph chain
${}^3T(M_2)_{ch}, \ldots, {}^6T(M_2)_{ch}$	overall second Zagreb order $3, \ldots, 6$ chain
$T(M_2)_{ch}$	overall second Zagreb chain
${}^3T(M_2)^*_{ch}, \ldots, {}^6T(M_2)^*_{ch}$	overall second Zagreb subgraph order $3, \ldots, 6$ chain
$T(M_2)^*_{ch}$	overall second Zagreb subgraph chain
${}^3TW_{ch}, \ldots, {}^6TW_{ch}$	overall Wiener order 3 chain
TW_{ch}	overall Wiener chain

3.7 Definitions of Descriptors

Leading references for the descriptors available in MOLGEN–QSPR :

Todeschini, R., Consonni, V.: Handbook of Molecular Descriptors. *Wiley-VCH, Weinheim and New York, 2000*; 2nd ed. 2009 under the new title Molecular Descriptors for Chemoinformatics.

Trinajstić, N.: *Chemical Graph Theory,* 2nd edition, CRC Press, Boca Raton, FL, 1992.

3.7.1 Definitions of Arithmetic Descriptors

1. **Numbers of atoms:** A denotes the number of atoms excluding H atoms. $A\,(incl.\,H)$ means the number of atoms including H atoms. N_H is the number of H atoms. Correspondingly, we use the notations $N_C, N_O, N_N, N_S, N_F, N_{Cl}, N_{Br}, N_I$ and N_P.

2. **Relative numbers of atoms:** The descriptors

$$rel.\,N_H, rel.\,N_C, rel.\,N_O, rel.\,N_N, rel.\,N_S, rel.\,N_F, rel.\,N_{Cl}, rel.\,N_{Br}, rel.\,N_I, rel.\,N_P$$

mean the number of the respective atoms in the index, divided by the total number of atoms (including H atoms). For example,

$$rel.\,N_H = \frac{N_H}{A\,(incl.\,H)}.$$

3. **Numbers of bonds:** B denotes the number of bonds in the H-suppressed molecule, while $B\,(incl.\,H)$ is the number of bonds in a molecule containing H atoms.

4. **Numbers of localized bonding electron pairs:** $loc.\,B$ is the number of localized bonding electron pairs in an H-suppressed molecule. Aromatic π electrons are delocalized and therefore not counted here. $loc.\,B\,(incl.\,H)$ is analogous but it includes bonds to H atoms.

5. **Numbers of single bonds:** $n-$ is the number of single bonds in an H-suppressed molecule. $n-\,(incl.\,H)$ analogously includes bonds to H atoms.

6. **Relative numbers of single bonds:** $rel.\,n-$ and $rel.\,n-\,(incl.\,H)$ indicate the relative numbers of bonds of an H-suppressed molecule:

$$rel.\,n- \;=\; \frac{n-}{B} \quad \text{and} \quad rel.\,n-\,(incl.\,H) = \frac{n-\,(incl.\,H)}{B\,(incl.\,H)}.$$

7. **Numbers and relative numbers of multiple bonds:** $n=$ is the number of double bonds, $n\#$ the number of triple bonds, and n_{aroma} indicates the number of aromatic bonds. Correspondingly, we use the notations

$$rel.\,n{=}, rel.\,n{=}\,(incl.\,H), rel.n\#, rel.n\#(incl.\,H), rel.\,n_{aroma}, rel.\,n_{aroma}\,(incl.\,H)$$

for the relative numbers of multiple bonds (relative to B, or to $B\,(incl.\,H)$).

8. **The cyclomatic number:** C is defined as $C = B - A + 1$.

9. **The molecular weight**[1] MW and $MW\,(incl.\,H)$ are the sums of the atomic weights in an H-suppressed molecule and in the molecule including the H atoms, respectively. The atomic weight is that of the natural abundance isotope mixture.

10. **The mean atomic weight (or average atomic weight):** The mean atomic weights are defined as

$$mean\ AW = \frac{MW}{A} \quad \text{and} \quad mean\ AW\,(incl.\,H) = \frac{MW\,(incl.\,H)}{A\,(incl.\,H)}.$$

11. **The total charge:** cha is the charge of the molecule.

12. **The number of radical centers:** n_{rad}

13. **The number of hydrogen bond donors** HBD is assumed to be the number of H atoms attached to O and N atoms, in accord with the Chemical Abstracts/ACD definition.[74]

14. **The number of hydrogen bond acceptors** HBA is assumed to be the number of N and O atoms, in accord with the Chemical Abstracts/ACD definition.[74]

15. **The number of charged atoms** is indicated as n_{cha}.

16. **Monoisotopic mass (exact and integer):** These are the sums of the (exact or integer) masses of the most abundant isotope for all atoms (incl. H), denoted by $mass\_exact$ and $mass\_int$, respectively.

3.7.2 Definitions of Topological Indices

Definitions of graph theoretical matrices

The graph theoretical indices are based on the following important graph theoretical notions:

- The **adjacency matrix** $A = (A_{ij})$ of the molecular graph. A_{ij} is defined to be 1 if there is a covalent bond between atoms i and j, and 0 otherwise, or, in terms of the corresponding molecular graph,

$$A_{ij} = \begin{cases} 1 & \text{if } edge(i,j) \text{ exists,} \\ 0 & \text{otherwise.} \end{cases}$$

  The *degree* of vertex i or atom i, δ_i, is the i-th row sum:

$$\delta_i = \sum_j A_{ij}.$$

- **The unsaturated adjacency matrix** $\hat{A} = (\hat{A}_{ij})$ is defined by

$$\hat{A}_{ij} = \begin{cases} 1 & \text{if there is a single bond between atoms } i \text{ and } j, \\ 2 & \text{if there is a double bond between atoms } i \text{ and } j, \\ 3 & \text{if there is a triple bond between atoms } i \text{ and } j, \\ 1.5 & \text{if there is an aromatic bond between atoms } i \text{ and } j, \\ 0 & \text{otherwise.} \end{cases}$$

- The **distance matrix** $D = (D_{ij})$, where D_{ij} means the distance (=shortest path length) between atoms i and j in the H-suppressed molecular graph.

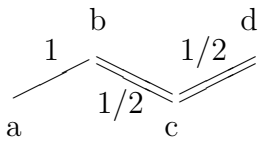
  The maximal entry in its i-th row is called *eccentricity* of atom i,

$$\eta_i = \max\{D_{ij} \mid 1 \leq j \leq A\}.$$

The *vertex distance degree* σ_i is defined as the i-th row sum of the distance matrix D of an H-suppressed molecular graph:

$$\sigma_i = \sum_j D_{ij}.$$

- **The unsaturated distance matrix** $\hat{D} = (\hat{D}_{ij})$, the rows and columns of which correspond to the non-H atoms. The entry $\hat{D}_{ij}$ is the length of the shortest path from atom i to atom j, where single bonds represent a distance of 1, double bonds represent a distance of 1/2, triple bonds represent a distance of 1/3, aromatic bonds represent a distance of 2/3. Here is an example:



In this example, the distance $\hat{D}_{ac}$ from a to c is $1 + 1/2 = 3/2$, and the distance $\hat{D}_{ad} = 1 + 1/2 + 1/2 = 2$.

The *unsaturated vertex distance degree* $\hat{\sigma}_i$ is defined as the i-th row sum of the unsaturated distance matrix $\hat{D}$ of an H-suppressed molecular graph:

$$\hat{\sigma}_i = \sum_j \hat{D}_{ij}.$$

- The **charge term matrix** $CT = (CT_{ij})$, a square matrix, the rows and columns of which correspond to the non-H atoms,

$$CT_{ij} = \begin{cases} \delta_i & \text{if } i = j, \\ M_{ij} - M_{ji} & \text{otherwise,} \end{cases}$$

where M is defined as $M = A \cdot D^{(-2)}$, and

$$D_{ij}^{(-2)} = \begin{cases} \frac{1}{(D_{ij})^2} & \text{if } i \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

- **The detour matrix** $\Delta = (\Delta_{ij})$, the rows and columns of which correspond to the non-H atoms. The entries are the lengths of longest paths between atoms,

$$\Delta_{ij} = \begin{cases} 0 & \text{if } i = j, \\ l_{ij} & \text{otherwise,} \end{cases}$$

  where l_{ij} is the length of the longest path between atoms i and j.

  A more logical definition includes closed detours from atom i to itself (rings of maximal length):

$$\Delta^*_{ij} = \begin{cases} l_{ii} & \text{if } i = j, \\ l_{ij} & \text{otherwise.} \end{cases}$$

  where l_{ii} is the size of the largest ring containing atom i, $l_{ii} = 0$ if atom i is not in a ring.

- **The Szeged matrix** $SZ = (SZ_{ij})$, the rows and columns of which correspond to the non-H atoms. The entry SZ_{ij} is the number of atoms in the H-suppressed molecule that are closer to i than to j,

$$SZ_{ij} = |\{a \mid a \text{ atom with } D_{ia} < D_{ja}\}|.$$

Definition of graph theoretical indices

1. **Wiener index:** W is the half-sum of the distance matrix entries of the H-suppressed molecule:[3]

$$W = \frac{1}{2} \cdot \sum_{i,j} D_{ij}.$$

2. **1st and 2nd Zagreb index:** M_1 is the sum (over all vertices) of squares of vertex degrees. M_2 is the sum (over all edges) of products of vertex degrees of atoms i and j forming an edge (i, j),[2,4]

$$M_1 = \sum_i (\delta_i)^2 \quad \text{and} \quad M_2 = \sum_{edge(i,j)} \delta_i \cdot \delta_j.$$

   The vertex degree δ_i of atom i is the number of its neighbors in an H-suppressed molecular graph.

3. **1st and 2nd modified Zagreb index:** These indices use the reciprocal vertex degrees of the atoms in an H-suppressed molecule,[5]

$$^{m}M_1 = \sum_i \frac{1}{\delta_i^2} \quad \text{and} \quad ^{m}M_2 = \sum_{edge(i,j)} \frac{1}{\delta_i \cdot \delta_j}\,.$$

Here m stands for "modified".

4. **Randić (or connectivity) indices:** They form the series of indices $^{m}\chi$ of order $m = 0, 1, 2, 3, \ldots$, defined by

$$^{m}\chi = \sum_{path\ p\ of\ length\ m} \prod_{i=1}^{A(p)} \frac{1}{\sqrt{\delta_i}}\,,$$

where the product is taken over the atoms in path p, and $A(p)$ means the number of atoms in that path.[6,7] For example, the Randić indices of order 0 and 1 are

$$^{0}\chi = \sum_i \frac{1}{\sqrt{\delta_i}} \quad \text{and} \quad ^{1}\chi = \sum_{edge(i,j)} \frac{1}{\sqrt{\delta_i \cdot \delta_j}}\,,$$

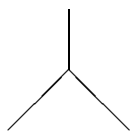
where the sum is taken over the vertices and the edges in an H-suppressed molecular graph, respectively.

5. **Solvation connectivity indices:** They form the series of indices $^{m}\chi^{s}$ with $m = 0, 1, 2, 3, \ldots$, defined by

$$^{m}\chi^{s} = \frac{1}{2^{m+1}} \cdot \sum_{path\ p\ of\ length\ m} \prod_{i=1}^{A(p)} \frac{L_i}{\sqrt{\delta_i}}\,,$$

where the product is taken over the atoms in the path, and L_i is the principal quantum number of atom i (= 2 for C, N, O, F, = 3 for Si, P, S, Cl, etc.).[1,10]

6. **Solvation connectivity index for clusters:** This index arises by taking the sum over all clusters of size 3, which means subgraphs of the following form:[1]



The index is defined by

$$^{3}\chi_c^s = \frac{1}{2^4} \cdot \sum_{cluster\ of\ size\ 3} \prod_{i=1}^{4} \frac{L_i}{\sqrt{\delta_i}} \cdot$$

7. **Kier and Hall (or valence) connectivity indices:** These form the series $^{m}\chi^v, m = 0, 1, 2, 3, \ldots$, and are defined as follows:[7,8,11]

$$^{m}\chi^v = \sum_{path\ p\ of\ length\ m} \prod_{i=1}^{A(p)} \frac{1}{\sqrt{\delta_i^v}} \cdot$$

δ_i^v, the *valence vertex degree* or *vertex valence* of atom i in an H-suppressed molecular graph, is defined as

$$\delta_i^v = \frac{Z_i^v - h_i}{Z_i - Z_i^v - 1},$$

where Z_i is the total number of electrons (= the atomic number) of atom i, Z_i^v the number of valence electrons, h_i the number of H atoms attached to atom i.

In MOLGEN–QSPR these indices are implemented for $m = 0, 1, 2, 3$.

8. **Kier shape indices 1, 2 and 3:** These are arithmetic expressions in terms of the number A of atoms and numbers ^{l}P of paths of length l in the molecular graph of the H-suppressed molecule:[12–14]

$$^{1}\kappa = \frac{A \cdot (A-1)^2}{(^{1}P)^2}, \quad ^{2}\kappa = \frac{(A-1) \cdot (A-2)^2}{(^{2}P)^2}, \quad ^{3}\kappa = \begin{cases} \frac{(A-3)\cdot(A-2)^2}{(^{3}P)^2} & \text{for even } A,\ A > 3, \\ \frac{(A-1)\cdot(A-3)^2}{(^{3}P)^2} & \text{for odd } A,\ A > 3\,. \end{cases}$$

Note that $^{1}P = B$, the number of bonds.

9. **Alpha-modified Kier shape indices 1, 2 and 3:**[12,13,15] These are

$$^{1}\kappa_\alpha = \frac{(A+\alpha) \cdot (A+\alpha-1)^2}{(^{1}P+\alpha)^2}, \quad ^{2}\kappa_\alpha = \frac{(A+\alpha-1) \cdot (A+\alpha-2)^2}{(^{2}P+\alpha)^2},$$

and

$$^{3}\kappa_\alpha = \begin{cases} \frac{(A+\alpha-3)\cdot(A+\alpha-2)^2}{(^{3}P+\alpha)^2} & \text{for even } A,\ A > 3, \\ \frac{(A+\alpha-1)\cdot(A+\alpha-3)^2}{(^{3}P+\alpha)^2} & \text{for odd } A,\ A > 3. \end{cases}$$

The modifying α is defined as follows:

$$\alpha = \sum_{i=1}^{A} \alpha_i = \sum_{i=1}^{A} \left(\frac{R_i}{R_{Csp^3}} - 1 \right),$$

where R_i is the covalent radius of the i-th atom in an H-suppressed molecule and R_{Csp^3} is the covalent radius of an sp^3 carbon atom. Here is a table with such values:

Atom / Hybrid i	R_i	α_i	Atom/Hybrid i	R_i	α_i
C_{sp3}	0.77	0.00	P_{sp3}	1.10	0.43
C_{sp2}	0.67	−0.13	P_{sp2}	1.00	0.30
C_{sp}	0.60	−0.22	S_{sp3}	1.04	0.35
N_{sp3}	0.74	−0.04	S_{sp2}	0.94	0.22
N_{sp2}	0.62	−0.20	F	0.72	−0.07
N_{sp}	0.55	−0.29	Cl	0.99	0.29
O_{sp3}	0.74	−0.04	Br	1.14	0.48
O_{sp2}	0.62	−0.20	I	1.33	0.73

10. **Kier molecular flexibility index, alpha modified and non-modified:** [1,16]

$$\Phi = \frac{{}^1\kappa_\alpha \cdot {}^2\kappa_\alpha}{A} \quad \text{and} \quad \Phi_{\overline{\alpha}} = \frac{{}^1\kappa \cdot {}^2\kappa}{A}\,.$$

11. **Platt number:** It is expressed in terms of the numbers $N(i)$ of neighbors of atoms,

$$F = \sum_{edge(i,j)} (N(i) + N(j) - 2),$$

The sum runs over all edges in the H-suppressed molecular graph.[17,18]

12. **Gordon-Scantlebury index:** N_{GS} is the number of path subgraphs of length 2 in an H-suppressed molecular graph.[1,2]

13. **Balaban index, saturated and unsaturated:** The saturated index is

$$J = \frac{B}{C+1} \sum_{edge(i,j)} \frac{1}{\sqrt{\sigma_i \cdot \sigma_j}}\,,$$

where B is the number of bonds, while σ_i means the i-th atom distance degree, i.e. $\sigma_i = \sum_j D_{ij}$. C is the cyclomatic number. The sum runs over all edges of an H-suppressed molecular graph.[19,20] The unsaturated index is

$$J_{unsat} = \frac{B}{C+1} \sum_{edge(i,j)} \frac{1}{\sqrt{\hat{\sigma}_i \cdot \hat{\sigma}_j}}\,,$$

where $\hat{\sigma}_i$ is the unsaturated distance degree, i.e. the i-th row sum in the unsaturated distance matrix.[21]

14. **Schultz molecular topological index MTI:** We introduce MTI' as the following scalar product of vectors:

$$MTI' = (\delta_1, \ldots, \delta_n)^t \cdot (\sigma_1, \ldots, \sigma_n)$$

and define the Schultz molecular index as

$$MTI = \sum_{i=1}^{n} \delta_i^2 + MTI'.$$

Quantities δ_i and σ_i are degree and distance degree, respectively, of atom i in the H-suppressed molecule.[22–25]

15. **Harary number:** This is defined as

$$H = \sum_{i=1}^{A} \sum_{j=i+1}^{A} \frac{1}{D_{ij}},$$

again for an H-suppressed molecular graph.[26–28]

16. **Walk counts:** We start with the molecular walk count of length k, defined by

$$mwc^{(k)} = \sum_{i,j} (A^k)_{ij},$$

where $A = (A_{ij})$ means the adjacency matrix of the H-suppressed molecular graph, $A^k = ((A^k)_{ij})$ its k-th power.

Remark: $mwc^{(0)}$ is equal to the number of atoms, $mwc^{(1)}$ is equal to $2B$, $mwc^{(2)} = M_1$, $mwc^{(3)} = 2M_2$.

Using this notion, we introduce the *total walk count*

$$twc = \sum_{k=1}^{n-1} mwc^{(k)}.$$

The sum runs over all lengths k (from 1 to $n-1$) of walks in an H-suppressed molecular graph, where n is the number of non-H atoms.[29–32]

Note: This is the original definition of twc.

17. **Unsaturated molecular walk counts:** These are defined in terms of powers of the unsaturated adjacency Matrix $\hat{A}$.

$$mwc^{(k)}_{unsat} = \sum_{i,j} (\hat{A}^k)_{ij} \, .$$

This expression is called the *unsaturated molecular walk count of length* k, while the *unsaturated total walk count* is the sum over these:

$$twc_{unsat} = \sum_{k=1}^{n-1} mwc^{(k)}_{unsat},$$

where n is the number of non-H atoms. The sum runs over all lengths k (from 1 to $n-1$) of walks in an H-suppressed molecular graph.

18. **Gravitational Indices (topo. dist.):** These are the indices

$$G_1\,(topol.) = \sum_{i=1}^{A} \sum_{j=i+1}^{A} \frac{w_i \cdot w_j}{D_{ij}^2} \quad \text{and} \quad G_1\,(topol., incl.\, H) = \sum_{i=1}^{A\,(incl.\,H)} \sum_{j=i+1}^{A\,(incl.\,H)} \frac{w_i \cdot w_j}{D_{ij}^2},$$

where w_i is the atomic weight of atom i (expressed in *amu*, i.e. 12.0110 for carbon), and the sum runs, in the first case, over all pairs of atoms in an H-suppressed molecular graph, while in the second case the hydrogen atoms are included.

If we restrict attention to bonds (pairs of distance 1), we obtain

$$G_2\,(topol.) = \sum_{edge(i,j)} w_i \cdot w_j \quad \text{and} \quad G_2\,(topol., incl.\, H) = \sum_{edge(i,j)} w_i \cdot w_j \, ,$$

where the latter includes bonds to H atoms.

19. **Hosoya index Z:**[34] Denoting by a_k the number of sets of k mutually non-adjacent edges in the H-suppressed molecular graph (so that, for example, $a_0 = 1$ and $a_1 = B$), while $\lfloor A/2 \rfloor$ denotes the biggest integer smaller than or equal to $A/2$, the Hosoya index is

$$Z = \sum_{k=0}^{\lfloor A/2 \rfloor} a_k.$$

20. **Basak Information Contents:** In order to obtain information content indices, Basak partitions the atoms of a molecule including H atoms into equivalence classes. Two atoms are considered equivalent if the numbers and atom types (chemical elements) of and the bond types to all their neighbors coincide, up to the neighborhood depth r. If for depth r G equivalence classes are found, then the number of atoms

in the g-th class is written as A_g^r, and the information content of order r, IC_r, is defined as

$$IC_r = \sum_{g=1}^{G} \frac{A_g^r}{A(incl.\,H)} \cdot \log_2 \frac{A_g^r}{A(incl.\,H)}.$$

The descriptors TIC_r, CIC_r, SIC_r and their multiples $N \cdot CIC_r, N \cdot SIC_r, N \cdot BIC_r$, for $r = 0, 1, 2, \ldots$, are defined as

$$\begin{aligned}
TIC_r &= A(incl.\,H) \cdot IC_r \\
CIC_r &= \log_2 A(incl.\,H) - IC_r \\
N \cdot CIC_r &= A(incl.\,H) \cdot CIC_r \\
SIC_r &= \frac{IC_r}{\log_2 A(incl.\,H)} \\
N \cdot SIC_r &= A(incl.\,H) \cdot SIC_r \\
BIC_r &= \frac{IC_r}{\log_2 B(incl.\,H)} \\
N \cdot BIC_r &= A(incl.\,H) \cdot BIC_r
\end{aligned}$$

Note: This definition of BIC_r is the original one.

The indices carry the following names:[35–37]

The index	its name
IC_r	Basak information content of order r
TIC_r	Basak total information content of order r
CIC_r	Basak complementary information content of order r
$N \cdot CIC_r$	total complementary information content of order r
SIC_r	Basak structural information content of order r
$N \cdot CIC_r$	total structural information content of order r
BIC_r	bonding information content of order r
$N \cdot BIC_r$	total bonding information content of order r

21. **Mean square distance index:** This index is defined as

$$MSD = \left(\frac{\sum_{i,j} (D_{ij})^2}{A \cdot (A-1)} \right)^{1/2},$$

where the sum is taken over all atoms in the H-suppressed molecular graph.[20]

22. **Detour indices:** If $\Delta = (\Delta_{ij})$ denotes the detour matrix of an H-suppressed molecular graph,

$$w = \frac{1}{2} \cdot \sum_{i,j} \Delta_{ij}$$

is the *detour index*. A variant is

$$w_{diag} = \frac{1}{2} \cdot \sum_{i,j} \Delta^*_{ij},$$

where $\Delta^* = (\Delta^*_{ij})$ means the detour matrix including main diagonal elements $\neq 0$.[38–42,73]

23. **Path counts:**[1,43,44] With ${}^lP_{acyc}$ being the number of paths of length l in the H-suppressed molecular graph without counting any closed paths (rings), and l_{max} being the maximum length of all unclosed paths, the *total molecular acyclic path count* is defined as

$$P_{acyc} = \sum_{l=1}^{l_{max}} {}^lP_{acyc}\,.$$

In MOLGEN–QSPR , acyclic path counts are implemented up to ${}^8P_{acyc}$. Longer paths (if any) are collectively counted in

$${}^{\geq 9}P_{acyc} = \sum_{l=9}^{l_{max}} {}^lP_{acyc}.$$

Considering also closed paths we get ${}^lP$, the number of paths of length l in the H-suppressed molecular graph, and the *total molecular path count*

$$P = \sum_{l=1}^{l_{max}} {}^lP\,.$$

Path counts are implemented in MOLGEN–QSPR up to ${}^8P$. Again, paths longer than 8 (if any) are collectively counted as

$${}^{\geq 9}P = \sum_{l=9}^{l_{max}} {}^lP.$$

24. **Ring counts:** Restricting attention to rings, we obtain the total ring count

$$rings = \sum_{l=3}^{l_{max}} {}^lrings\,,$$

where ${}^{l}rings$ is the number of rings of length (ring size) l in the H-suppressed molecular graph, l_{max} the maximum ring size.[1]

In MOLGEN–QSPR ring counts ${}^{3}rings, \ldots, {}^{8}rings$ are implemented, rings of size ≥ 9 (if any) are collectively counted as

$$ {}^{\geq 9}rings = \sum_{l \geq 9}^{l_{max}} {}^{l}rings. $$

25. **Topological charge indices of order k:** These indices use the charge term matrix $CT = (CT_{ij})$ as well as the distance matrix. They are defined in terms of the atoms in the H-suppressed molecule as follows,[45,46]

$$ ch.\, G_k = \frac{1}{2} \cdot \sum_{i,j} |CT_{ij}| \cdot \delta(k, D_{ij}), \;\; k = 1, 2, \ldots $$

where $\delta(k, D_{ij})$ is the Kronecker delta, i.e.

$$ \delta(k, D_{ij}) = \begin{cases} 1 & \text{if } k = D_{ij}, \\ 0 & \text{otherwise.} \end{cases} $$

These indices are called *topological charge indices of order k* ($k = 1, \ldots, 8$ in MOLGEN–QSPR), while *the mean topological charge indices of order k* are

$$ ch.\, J_k = \frac{ch.\, G_k}{A - 1}, \;\; k = 1, 2, \ldots $$

and the *global topological charge indices of order k* are

$$ ch.\, J[k] = \sum_{l=1}^{k} ch.\, J_k. $$

In MOLGEN–QSPR, mean topological charge indices are implemented up to $ch.\, J_8$, as well as the global topological charge index $ch.\, J[5]$.

26. **The diameter** is the maximal distance between two atoms in the H-suppressed molecule,

$$ D = \max\{D_{ij} \mid 1 \leq i < j \leq A\}. $$

27. **The eccentric connectivity index:** This is

$$\xi^c = \sum_{i=1}^{A} \eta_i \cdot \delta_i,$$

where η_i is the maximum entry in the i-th row of the distance matrix, δ_i the vertex degree of atom i.[50]

28. **The principal (leading, first) eigenvalue of A:** λ_1^A is the principal eigenvalue of the adjacency matrix. We note that A is a real symmetric matrix and therefore diagonalizable, with real diagonal elements.

29. **The sum of coefficients of the principal eigenvector of A:**[51] Denoting by c_i^{A1} the i-th coefficient of the eigenvector of the principal eigenvalue of A, we obtain the descriptors

$$SCA1 = \sum_i |\, c_i^{A1} |\,, \; SCA2 = \frac{SCA1}{n}\,, \; SCA3 = \frac{n}{10} \cdot \log\left(SCA1\right).$$

The sum runs over all n atoms of an H-suppressed molecule.

30. **The principal (leading, first) eigenvalue of D:** λ_1^D denotes the principal eigenvalue of the distance matrix.[52]

31. **The total Chi index** is defined as

$$\chi_T = \prod_{i=1}^{A} \frac{1}{\sqrt{\delta_i}}.$$

The product runs over all atoms of an H-suppressed molecular graph.[53]

32. **The number of methyl groups** is denoted by T_m.[53]

33. **The number of pairs of methyl groups at distance 3** is T_3.[53]

34. **The number of freely rotatable bonds** FRB means the number of bonds that are acyclic, single, not terminal (in the H-suppressed molecule), and not an amide $\mathsf{C}-\mathsf{N}$ bond.[54]

35. **Szeged indices:** These are expressed in terms of the Szeged matrix defined above:

$$SZD = \sum_{edge(i,j)} SZ_{ij} \cdot SZ_{j\,i} \quad \text{and} \quad SZD_P = \sum_{i,j=1}^{A} SZ_{ij} \cdot SZ_{j\,i}\,.$$

The edges and pairs are those in an H-suppressed molecular graph. SZD is called the *Szeged index*, while SZD_P is the hyper-Szeged index.[59–62]

36. **Connectivity indices for substructures:** These topological indices are expressed in terms of subgraphs of type q (which means paths, clusters, path-clusters or chains) in the H-suppressed molecular graph. m is the order, i.e. the number of edges of the subgraphs considered. $K(m,q)$ is the number of subgraphs of type q and order m. n is the number of atoms in the subgraph considered.[8,9]

$$^{m}\chi_q = \sum_{k=1}^{K(m,q)} \frac{1}{\sqrt{\prod_{i=1}^{n} \delta_i}}, \quad ^{m}\chi_q^v = \sum_{k=1}^{K(m,q)} \frac{1}{\sqrt{\prod_{i=1}^{n} \delta_i^v}} .$$

Available in MOLGEN–QSPR are the connectivity indices

$$^{m}\chi_p, 3 \le m \le 6,\ ^{m}\chi_c, 3 \le m \le 6,\ ^{m}\chi_{pc}, 4 \le m \le 6,\ ^{m}\chi_{ch}, 3 \le m \le 6\,,$$

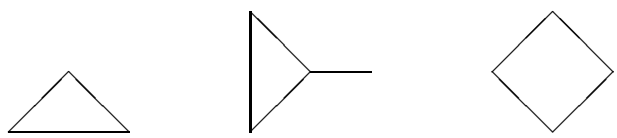
and the valence connectivity indices

$$^{m}\chi_p^v, 3 \le m \le 6,\ ^{m}\chi_c^v, 3 \le m \le 6,\ ^{m}\chi_{pc}^v, 4 \le m \le 6,\ ^{m}\chi_{ch}^v, 3 \le m \le 6\,,$$

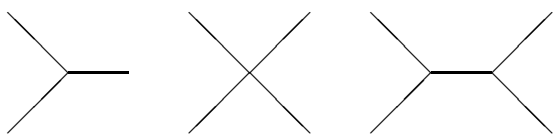
where a subgraph is

- of type *chain* (ch) if it contains a cycle ($m \ge 3$),
- otherwise if every vertex has either one or more than two non-H neighbors it is of type *cluster* (c) for $m \ge 3$,
- otherwise if every vertex has one or two non-H neighbors it is of type *path* (p) for $m \ge 3$,
- otherwise it is of type *path-cluster* (pc) for $m \ge 4$. So a path-cluster has no cycles but vertices with one, two and more than two non-H neighbors.

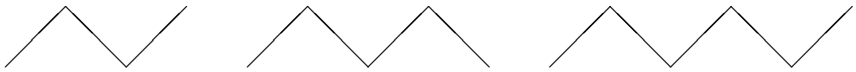
For example, chains of order $m = 3, 4, 4$ are



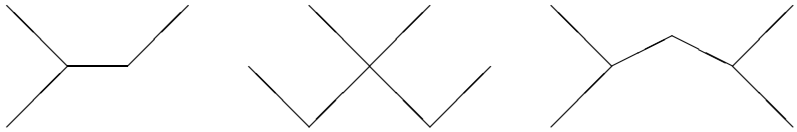
Clusters of order $m = 3, 4, 5$ are



Paths of order $m = 3, 4, 5$ are



Path-clusters of order $m = 4, 6, 6$ are



For classification of subgraphs, the numbers of non-H neighbors are taken as they are in the isolated subgraphs, whereas in the calculation of χ values the δ_i are taken as they are in the whole graph.

37. **Size of the topological symmetry group:** The topological symmetry group is the set of automorphisms of the H-suppressed molecular graph. An automorphism is the possibility to exchange vertices such that all neighborhood relations are conserved, that is, after this operation the graph looks the same as before.[75] The order or size of this group is indicated as *sym_top*. In a completely unsymmetric graph this number is 1, since there is always one automorphism, the trivial exchange of every vertex against itself. In the (H-suppressed) graph of 2-methylbutane (or of 2-methyl-2-butene) the two methyl groups bound to the same C atom are exchangeable, so that there is one nontrivial automorphism, and the size of the topological symmetry group is 2.

38. **The topological radius** is:[1]

$$R = \min_{1 \le i \le A} \left(\max_{1 \le j \le A} (D_{ij}) \right).$$

39. **The number of connectivity components** *con. comp* means the number of connected components of the molecular graph. In most cases, this index is equal to 1. If the compound is made of more than one component, the index increases.

3.7.3 Definitions of Electrotopological and AI indices

1. **Sum of E-state of atomic subgraphs:** Every non-H atom i is attributed a number S_i (*electrotopological state* or *E-state*) that is composed of two terms:

$$S_i = I_i + \sum_j \Delta I_{ij}.$$

The first term is the *intrinsic state* I_i, characteristic for an atom type plus its attached H atom and bonds, e.g. the methyl group, and defined as

$$I_i = \frac{(2/L)^2 \delta_i^v + 1}{\delta_i}.$$

The second term stands for the sum of influences of all other atoms j in the molecule on atom i, where

$$\Delta I_{ij} = \frac{I_i - I_j}{(D_{ij} + 1)^2}.$$

Thus, S_i characterizes a particular non-H atom, e.g. a particular methyl group in the ethyl acetate molecule. In MOLGEN–QSPR, the sum of E-state values of all such atoms is available, e.g. the sum of E-states of all methyl groups in a molecule, called $S(sCH_3)$, which in the case of ethyl acetate is the sum of E-states of the two methyl groups.

Here is a table of the 80 available sums of E-states of atomic subgraphs:

$S(sCH3)$	$S(sssNH)$	$S(aaS)$	$S(ssSiH2)$
$S(dCH2)$	$S(dsN)$	$S(dssS)$	$S(sssSiH)$
$S(ssCH2)$	$S(aaN)$	$S(ddssS)$	$S(ssssSi)$
$S(tCH)$	$S(sssN)$	$S(sssssS)$	$S(sGeH3)$
$S(dsCH)$	$S(ddsN)$	$S(sCl)$	$S(ssGeH2)$
$S(aaCH)$	$S(aasN)$	$S(sSeH)$	$S(sssGeH)$
$S(sssCH)$	$S(ssssN)$	$S(dSe)$	$S(ssssGe)$
$S(ddC)$	$S(sOH)$	$S(ssSe)$	$S(sAsH2)$
$S(tsC)$	$S(dO)$	$S(aaSe)$	$S(ssAsH)$
$S(dssC)$	$S(ssO)$	$S(dssSe)$	$S(sssAs)$
$S(aasC)$	$S(aaO)$	$S(ddssSe)$	$S(sssdAs)$
$S(aaaC)$	$S(sF)$	$S(sBr)$	$S(ssssAs)$
$S(ssssC)$	$S(sPH2)$	$S(sI)$	$S(sSnH3)$
$S(sNH3)$	$S(ssPH)$	$S(sLi)$	$S(ssSnH2)$
$S(sNH2)$	$S(sssP)$	$S(ssBe)$	$S(sssSnH)$
$S(ssNH2)$	$S(dsssP)$	$S(ssssBe)$	$S(ssssSn)$
$S(dNH)$	$S(ssssP)$	$S(ssBH)$	$S(sPbH3)$
$S(ssNH)$	$S(sSH)$	$S(sssB)$	$S(ssPbH2)$
$S(aaNH)$	$S(dS)$	$S(ssssB)$	$S(sssPbH)$
$S(tN)$	$S(ssS)$	$S(sSiH3)$	$S(ssssPb)$

where s means a single bond, ss two single bonds, d a double bonds, t a triple bond, a an aromatic bond, etc. to the specified atom, disregarding bonds to H atoms specified.[58]

2. **AI of atomic subgraphs:** These are quantities similar to the electrotopological indices. For example,

$$AI(sCH_3) = m + \frac{\sum_{i=1}^{m} \delta_i^{mod} \cdot \sigma_i^2}{\sum_{i=1}^{A} \delta_i^{mod} \cdot \sigma_i^2},$$

where m is the number of $-CH_3$ subgraphs, and σ_i the distance degree of atom i. δ_i^{mod} is the *modified degree* of atom i,

$$\delta_i^{mod} = \delta_i + k_i, \quad \text{where} \quad k_i = \frac{1}{\left(\frac{2}{A}\right)^2 \cdot \frac{Z_i^v - h_i}{Z_i - Z_i^v - 1} + 1} = \frac{1}{\left(\frac{2}{A}\right)^2 \cdot \delta_i^v + 1},$$

h_i is the number of H atoms attached to atom i, Z_i^v the number of valence electrons of atom i and Z_i its atomic number. Remember that the term

$$\frac{Z_i^v - h_i}{Z_i - Z_i^v - 1} = \delta_i^v,$$

called *valence degree* of atom i, was introduced above, in connection with Kier and Hall (or valence) connectivity.

Here is the list of all AI descriptors available in MOLGEN–QSPR :

$AI(sCH3)$	$AI(sssNH)$	$AI(aaS)$	$AI(ssSiH2)$
$AI(dCH2)$	$AI(dsN)$	$AI(dssS)$	$AI(sssSiH)$
$AI(ssCH2)$	$AI(aaN)$	$AI(ddssS)$	$AI(ssssSi)$
$AI(tCH)$	$AI(sssN)$	$AI(sssssS)$	$AI(sGeH3)$
$AI(dsCH)$	$AI(ddsN)$	$AI(sCl)$	$AI(ssGeH2)$
$AI(aaCH)$	$AI(aasN)$	$AI(sSeH)$	$AI(sssGeH)$
$AI(sssCH)$	$AI(ssssN)$	$AI(dSe)$	$AI(ssssGe)$
$AI(ddC)$	$AI(sOH)$	$AI(ssSe)$	$AI(sAsH2)$
$AI(tsC)$	$AI(dO)$	$AI(aaSe)$	$AI(ssAsH)$
$AI(dssC)$	$AI(ssO)$	$AI(dssSe)$	$AI(sssAs)$
$AI(aasC)$	$AI(aaO)$	$AI(ddssSe)$	$AI(sssdAs)$
$AI(aaaC)$	$AI(sF)$	$AI(sBr)$	$AI(sssssAs)$
$AI(ssssC)$	$AI(sPH2)$	$AI(sI)$	$AI(sSnH3)$
$AI(sNH3)$	$AI(ssPH)$	$AI(sLi)$	$AI(ssSnH2)$
$AI(sNH2)$	$AI(sssP)$	$AI(ssBe)$	$AI(sssSnH)$
$AI(ssNH2)$	$AI(dsssP)$	$AI(ssssBe)$	$AI(ssssSn)$
$AI(dNH)$	$AI(sssssP)$	$AI(ssBH)$	$AI(sPbH3)$
$AI(ssNH)$	$AI(sSH)$	$AI(sssB)$	$AI(ssPbH2)$
$AI(aaNH)$	$AI(dS)$	$AI(ssssB)$	$AI(sssPbH)$
$AI(tN)$	$AI(ssS)$	$AI(sSiH3)$	$AI(ssssPb)$

where s means a single bond, ss two single bonds, d a double bonds, t a triple bond, a an aromatic bond, etc. to the specified atom, not counting bonds to H atoms specified.[63–66]

3. **Xu indices** are defined as follows:[1,67] The *Xu index* is

$$Xu = \sqrt{A} \cdot \log \frac{\sum_{i=1}^{A} \delta_i \cdot \sigma_i^2}{\sum_{i=1}^{A} \delta_i \cdot \sigma_i},$$

while the *modified Xu index* is

$$Xu^m = \sqrt{A} \cdot \log \frac{\sum_{i=1}^{A} \delta_i^{mod} \cdot \sigma_i^2}{\sum_{i=1}^{A} \delta_i^{mod} \cdot \sigma_i}.$$

3.7.4 Definitions of Geometrical Indices

1. **The steric energy:** $st.energy$ is calculated by molecular mechanics in MOLGEN, it is the target quantity minimized thereby. All other descriptors appearing in this subsection depend on geometry, that is on the particular conformer obtained in such optimization.

2. **Gravitational Indices (3D dist.):**[1,33] Using the geometrical distance (expressed in Ångström Å) of atoms i and j, we find the indices

$$G_1 = \sum_{i=1}^{A} \sum_{j=i+1}^{A} \frac{w_i \cdot w_j}{r_{ij}^2} \quad \text{and} \quad G_1\,(incl.\,H) = \sum_{i=1}^{A\,(incl.\,H)} \sum_{j=i+1}^{A\,(incl.\,H)} \frac{w_i \cdot w_j}{r_{ij}^2}.$$

Again, the summation runs, in the first case, over all pairs of atoms in an H-suppressed molecular graph, while in the second case H atoms are included.

If only bonded pairs are considered, the following indices are obtained, without and with consideration of bonds to H atoms,

$$G_2 = \sum_{edge\,(i,j)} \frac{w_i \cdot w_j}{r_{ij}^2} \quad \text{and} \quad G_2\,(incl.\,H) = \sum_{edge\,(i,j)} \frac{w_i \cdot w_j}{r_{ij}^2}.$$

3. **Principal moments of inertia:** I_A, I_B, I_C are the three principal moments of inertia of the molecule with $I_A \leq I_B \leq I_C$, i.e. moments of inertia for rotation about three mutually perpendicular axes oriented such that one of the moments is a maximum, another one a minimum.[1]

4. **Shadows:** $SHDW1, SHDW2$ and $SHDW3$ mean the areas of the projection of the molecular surface onto the planes XY, XZ and YZ, respectively. They are

called the XY shadow, the XZ shadow, the YZ shadow. X, Y and Z axes are the molecule's principal axes of inertia.[1,55–57]

From these indices we obtain the descriptors

$$SHDW4 = \frac{SHDW1}{L_x \cdot L_y}, \; SHDW5 = \frac{SHDW2}{L_x \cdot L_z}, \; SHDW6 = \frac{SHDW3}{L_y \cdot L_z},$$

where L_x, L_y and L_z are the maximal dimension of the molecular surface in X, Y and Z direction using vdw radii. They are called the *standardized* XY, XZ and YZ shadow.

We also introduce the quotients

$$\frac{SHDWi}{SHDWj}\,, \; i, j \in \{1, 2, 3\}\,, \; i < j\,.$$

These quotients are the XY/XZ shadow, etc..

Moreover, we introduce the *size sorted shadows*

$$ssSHDW1\,, \; ssSHDW2\,, \; ssSHDW3\,,$$

of which $ssSHDW1$ is the largest, $ssSHDW2$ is the second largest, and $ssSHDW3$ is the smallest. The prefix ss stands for size sorted.

In addition we have the *size sorted standardized shadows*

$$ssHDW4 = \frac{ssSHDW1}{L_x \cdot L_y}, \; ssHDW5 = \frac{ssSHDW2}{L_x \cdot L_z}, \; ssSHDW6 = \frac{ssSHDW3}{L_y \cdot L_z},$$

and the quotients

$$\frac{ssSHDWi}{ssSHDWj}, \; i, j \in \{1, 2, 3\}, \; i < j\,.$$

5. **Van der Waals volume** V_{vdw}, density ρ_{vdw}, V^s_{vdw} and V_{cub} are calculated for molecules including H atoms.

   V_{vdw} is the volume of the molecule, evaluated by using vdw radii for each atom. The other descriptors are obtained as follows:

$$\rho_{vdw} = \frac{MW\,(incl.\,H)}{V_{vdw}}\,, \; V_{cub} = L_x \cdot L_y \cdot L_z\,, \; V^s_{vdw} = \frac{V_{vdw}}{V_{cub}}\,,$$

   where L_x, L_y and L_z are the maximum dimensions of the molecular surface in X, Y and Z direction by using vdw radii, where X, Y and Z are the principal axes of

inertia of the molecule (incl. H atoms).

V_{vdw} is called the *Van der Waals volume*, ρ_{vdw} is the *density by Van der Waals volume*, V^s_{vdw} the *standardized Van der Waals volume*, V_{cub} the *enclosing cuboid*.[1]

6. **Van der Waals surface** S_{vdw} is the surface of the molecule by using vdw radii for each atom.

7. **The solvent accessible surface area** $SASA_{H_2O}$ is the solvent accessible surface of the molecule by using vdw radii and an H_2O molecule (r = 1.5Å) as a probe, while $SASA_H$ is the solvent accessible surface of the molecule by using vdw radii and an H atom (r = 1.2Å) as a probe.

8. **The geometrical diameter** D_{3D} is the maximum distance of two points on the vdw surface of the molecule including H atoms:

$$D_{3D} = \max\{\,|\,b - a| \text{ for points } a, b \text{ in the vdw surface}\}.$$

9. **Enclosing sphere** V_{sphere} is the volume of the enclosing sphere (including vdw radii) of the molecule including H atoms:

$$V_{sphere} = \frac{4}{3} \cdot \pi \cdot \left(\frac{D_{3D}}{2}\right)^3 = \pi \cdot \frac{D^3_{3D}}{6}\,.$$

3.7.5 Definitions of Miscellaneous Indices

1. **Crippen atom type numbers:** $at\,C01{-}atC27, atH01{-}atH04, atO01{-}atO12, atN01{-}atN14, atHal, atCl, atBr, atI, atF, atP, atS01 - atS03, atMe01, atMe02$ are occurrence numbers of atom types. In Crippen's scheme, an atom is typified according to its nature and to that of its neighbors.[47] Thus,

   the C atom in a methyl group bonded to aliphatic C is of atom type C01,

   the C atom in a methyl group bonded to N or O is of atom type C03,

   the C atom in a methyl group bonded to aromatic C is of atom type C08, etc..

2. **slog P and sMR:** These are $\log P$ and molar refraction as calculated by Crippen's method.[47] Denote by N_k the number of atoms of Crippen type k, and by a_k the hydrophobicity increment of an atom of type k, then

$$slog\,P = \sum_k a_k \cdot N_k.$$

If b_k denotes the increment for the molar refractivity of an atom of type k, then we obtain sMR, the molar refractivity as calculated by Crippen's method,

$$sMR = \sum_k b_k \cdot N_k.$$

3.7.6 Definition of Overall indices

1. **Numbers of subgraphs:** Let ${}^mK$ denote the number of subgraphs of m edges in the H-suppressed molecular graph,

$${}^mK = |\{S \mid S \text{ a subgraph of } m \text{ edges}\}|\,,\ m = 0, 1, 2, \ldots.$$

Using these indices we obtain numbers of subgraphs with restricted number of edges. For example,

$${}^{0-8}K = \sum_{m=0}^{8} {}^mK$$

is the number of subgraphs of ≤ 8 edges.[48,49]

2. **Overall indices:**[68–71] These indices are denoted as ${}^mTO, {}^mTO^*, \ldots, TO_q^*$. T is the overall index sign. For the molecule each connected subgraph S up to size m is constructed. The letter O means one of these: M_1, the first Zagreb index, or M_2, the second Zagreb index, or W, the Wiener index, or C (for connectivity, stands for the sum over the vertex degrees of the atoms in the subgraph considered), or C^v (represents the sum over the valence vertex degrees of the atoms). In formal terms, we obtain the indices

$${}^mTO = \sum_{S\,of\,size\,m} O(S),\quad {}^mTO^* = \sum_{S\,of\,size\,m} O^*(S)\,,$$

$${}^mTO_q = \sum_{S\,of\,size\,m,\ type\,q} O(S),\quad {}^mTO_q^* = \sum_{S\,of\,size\,m,\ type\,q} O^*(S)\,.$$

If subgraphs of all sizes are considered, we obtain

$$TO = \sum_S O(S),\quad TO^* = \sum_S O^*(S)\,,$$

$$TO_q = \sum_{S\,of\,type\,q} O(S),\quad TO_q^* = \sum_{S\,of\,type\,q} O^*(S)\,.$$

MOLGEN–QSPR contains these indices for the following parameters:

descriptor	range of parameter m	unrestricted version
${}^mTC$	$0 \leq m \leq 6$	TC
${}^mTC^*$	$1 \leq m \leq 6$	TC^*
${}^mTC^v$	$0 \leq m \leq 6$	TC^v
${}^mTM_1$	$0 \leq m \leq 6$	TM_1
${}^mTM_1^*$	$1 \leq m \leq 6$	TM_1^*
${}^mTM_2$	$1 \leq m \leq 6$	TM_2
${}^mTM_2^*$	$1 \leq m \leq 6$	TM_2^*
${}^mTW$	$1 \leq m \leq 6$	TW
${}^mTC_p$	$3 \leq m \leq 6$	TC_p
${}^mTC_p^*$	$3 \leq m \leq 6$	TC_p^*
${}^mTC_p^v$	$3 \leq m \leq 6$	TC_p^v
${}^mT(M_1)_p$	$3 \leq m \leq 6$	$T(M_1)_p$
${}^mT(M_1)_p^*$	$3 \leq m \leq 6$	$T(M_1)_p^*$
${}^mT(M_2)_p$	$3 \leq m \leq 6$	$T(M_2)_p$
${}^mT(M_2)_p^*$	$3 \leq m \leq 6$	$T(M_2)_p^*$
${}^mTW_p$	$3 \leq m \leq 6$	TW_p
${}^mTC_c$	$3 \leq m \leq 6$	TC_c
${}^mTC_c^*$	$3 \leq m \leq 6$	TC_c^*
${}^mTC_c^v$	$3 \leq m \leq 6$	TC_c^v
${}^mT(M_1)_c$	$3 \leq m \leq 6$	$T(M_1)_c$
${}^mT(M_1)_c^*$	$3 \leq m \leq 6$	$T(M_1)_c^*$
${}^mT(M_2)_c$	$3 \leq m \leq 6$	$T(M_2)_c$
${}^mT(M_2)_c^*$	$3 \leq m \leq 6$	$T(M_2)_c^*$
${}^mTW_c$	$3 \leq m \leq 6$	TW_c
${}^mTC_{pc}$	$4 \leq m \leq 6$	TC_{pc}
${}^mTC_{pc}^*$	$4 \leq m \leq 6$	TC_{pc}^*
${}^mTC_{pc}^v$	$4 \leq m \leq 6$	TC_{pc}^v
${}^mT(M_1)_{pc}$	$4 \leq m \leq 6$	$T(M_1)_{pc}$
${}^mT(M_1)_{pc}^*$	$4 \leq m \leq 6$	$T(M_1)_{pc}^*$
${}^mT(M_2)_{pc}$	$4 \leq m \leq 6$	$T(M_2)_{pc}$
${}^mT(M_2)_{pc}^*$	$4 \leq m \leq 6$	$T(M_2)_{pc}^*$
${}^mTW_{pc}$	$4 \leq m \leq 6$	TW_{pc}
${}^mTC_{ch}$	$3 \leq m \leq 6$	TC_{ch}
${}^mTC_{ch}^*$	$3 \leq m \leq 6$	TC_{ch}^*
${}^mTC_{ch}^v$	$3 \leq m \leq 6$	TC_{ch}^v
${}^mT(M_1)_{ch}$	$3 \leq m \leq 6$	$T(M_1)_{ch}$
${}^mT(M_1)_{ch}^*$	$3 \leq m \leq 6$	$T(M_1)_{ch}^*$
${}^mT(M_2)_{ch}$	$3 \leq m \leq 6$	$T(M_2)_{ch}$
${}^mT(M_2)_{ch}^*$	$3 \leq m \leq 6$	$T(M_2)_{ch}^*$
${}^mTW_{ch}$	$3 \leq m \leq 6$	TW_{ch}

The sums run over the subgraphs (regarding m and q if specified) and sum up the values of the indices specified (e.g. W for Wiener index) of the subgraphs. In TC, TM_1, TM_2 calculations the δ values of the vertices of the subgraphs are used. If no asterisk appears in the symbol of an index, then these are taken as they are in the parent graph. If an asterisk appears in the symbol of an index, then δ values are taken as they are in the respective isolated subgraph.[68–71]

3.8 References

[1] Todeschini, R., Consonni, V., Handbook of Molecular Descriptors, *Wiley-VCH, Weinheim and New York, 2000*; 2nd ed. 2009 under the new title Molecular Descriptors for Chemoinformatics

[2] Trinajstić, N.: Chemical Graph Theory, *CRC Press, Boca Raton, FL, 2nd ed. 1992*

[3] Wiener, H.: Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc. 1947, 69, 17-20*

[4] Gutman, I.; Ruščić, B.; Trinajstić, N.; Wilcox, C. F.: Graph Theory and Molecular Orbitals. XII. Acyclic Polyenes. *J. Chem. Phys. 1975, 62, 3399-3405*

[5] Nikolić, S.; Kovačević, G.; Miličević, A.; Trinajstić, N.: The Zagreb Indices 30 Years After. *Croat. Chem. Acta, 2003, 76, 113-124*

[6] Randić, M.: On Characterization of Molecular Branching. *J. Am. Chem. Soc. 1975, 97, 6609-6615*

[7] Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H.: Molecular Connectivity V: Connectivity Series Applied to Density. *J. Pharm. Sci. 1976, 65, 1226-1230*

[8] Kier, L. B.; Hall L. H.: The Nature of Structure-Activity Relationships and their Relation to Molecular Connectivity. *Eur. J. Med. Chem. 1977, 12, 307-312*

[9] Kier, L. B.; Hall L. H.: Molecular Connectivity in Structure-Activity Analysis. *Research Studies Press - Wiley, Chichester (UK), 1986*

[10] Zefirov, N. S.; Palyulin, V. A.: QSAR for Boiling Points of "Small" Sulfides. Are the "High-Quality Structure-Property-Activity Regressions" the Real High Quality QSAR Models? *J. Chem. Inf. Comput. Sci. 2001, 41, 1022-1027*

[11] Kier, L. B.; Hall L. H.: Derivation and Significance of Valence Molecular Connectivity. *J. Pharm. Sci. 1981, 70, 583-589*

[12] Kier, L. B.: Shape Indexes of Orders One and Three from Molecular Graphs. *Quant. Struct.-Act. Relat. 1986, 5, 1-7*

[13] Kier, L. B.: Indexes of Molecular Shape from Chemical Graphs. *Acta Pharm. Jugosl. 1986, 36, 171-188*

[14] Kier, L. B.: A Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat. 1985, 4, 109-116*

[15] Kier, L. B.: Distinguishing Atom Differences in a Molecular Graph Shape Index. *Quant. Struct.-Act. Relat. 1986, 5, 7-12*

[16] Kier, L. B.: An Index of Molecular Flexibility from Kappa Shape Attributes. *Quant. Struct.-Act. Relat. 1989, 8, 221-224*

[17] Platt, J. R.: Influence of Neighbor Bonds on Additive Bond Properties in Paraffins. *J. Chem. Phys. 1947, 15, 419-420*

[18] Platt, J. R.: Prediction of Isomeric Differences in Paraffin Properties. *J. Phys. Chem. 1952, 56, 328-336*

[19] Balaban, A. T.: Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett. 1982, 89, 399-404*

[20] Balaban, A. T.: Topological Indices Based on Topological Distances in Molecular Graphs. *Pure Appl. Chem. 1983, 55, 199-206*

[21] Balaban, A. T.; Filip, P.: Computer Program For Topological Index J. *MATCH Commun. Math. Comp. Chem. 1984, 16, 163-190*

[22] Schultz, H. P.: Topological Organic Chemistry. 1. Graph Theory and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci. 1989, 29, 227-228*

[23] Schultz, H. P.; Schultz, T. P.: Topological Organic Chemistry. 6. Graph Theory and Molecular Topological Indices of Cycloalkanes. *J. Chem. Inf. Comput. Sci. 1993, 33, 240-244*

[24] Müller, W. R.; Szymanski, K.; Knop, J. v.; Trinajstić, N.: Molecular Topological Indices. *J. Chem. Inf. Comput. Sci. 1990, 30, 160-163*

[25] Mihalić, Z.; Nikolić, S; Trinajstić, N.: Comparative Study of Molecular Descriptors Derived from the Distance Matrix. *J. Chem. Inf. Comput. Sci. 1992, 32, 28-37*

[26] Ivanciuc, O.; Balaban, T.-S.; Balaban, A. T.: Design of Topological Indices. Part 4. Reciprocal Distance Matrix, Related Local Vertex Invariants and Topological Indices. *J. Math. Chem. 1993, 12, 309-318*

[27] Plavsić, D.; Nikolić, S; Trinajstić, N.; Mihalić, Z.: On the Harary Index for the Characterization of Chemical Graphs. *J. Math. Chem. 1993, 12, 235-250*

[28] Lucić, B.; Milicević, A.; Nikolić, S; Trinajstić, N.: Harary Index – Twelve Years Later. *Croat. Chem. Acta 2002, 75, 847-867*

[29] Rücker, G.; Rücker, C.: Counts of All Walks as Atomic and Molecular Descriptors. *J. Chem. Inf. Comput. Sci. 1993, 33, 683-695*

[30] Rücker, G.; Rücker, C.: Walk Counts, Labyrinthicity, and Complexity of Acyclic and Cyclic Graphs and Molecules. *J. Chem. Inf. Comput. Sci. 2000, 40, 99-106*

[31] Gutman, I.; Rücker, C.; Rücker, G.: On Walks in Molecular Graphs. *J. Chem. Inf. Comput. Sci. 2001, 41, 739-745*

[32] Nikolić, S; Trinajstić, N.; Tolić, I. M.; Rücker, G.; Rücker, C.: On Molecular Complexity Indices. Chapter 2, pages 29-89 in Complexity in Chemistry (Bonchev, D.; Rouvray, D. H., Eds.), *Taylor and Francis, London, 2003*

[33] Katritzky, A. R.; Mu L.; Lobanov, V. S.; Karelson, M.: Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem. 1996, 100, 10400-10407*

[34] Hosoya, H.: Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn. 1971, 44, 2332-2339*

[35] Basak, S. C.: Information Theoretic Indices of Neighborhood Complexity and Their Applications. Chapter 12 in Topological Indices and Related Descripors in QSAR and QSPR (Devillers, J.; Balaban, A. T., Eds.), *Gordon and Breach, Amsterdam, 1999*

[36] Basak, S. C.: Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach. *Med. Sci. Res. 1987, 15, 605-609*

[37] Basak, S. C.; Gute, B. D.: Characterization of Molecular Structures Using Topological Indices. *SAR QSAR Environ. Res. 1997, 7, 1-21*

[38] Ivanciuc, O.; Balaban, A. T.: Design of Topological Indices. Part 8. Path Matrices and Derived Molecular Graph Invariants. *MATCH Commun. Math. Comp. Chem. 1994, 30, 141-152*

[39] Amić, D.; Trinajstić, N.: On the Detour Matrix. *Croat. Chem. Acta. 1995, 68, 53-62*

[40] Lukovits, I.: The Detour Index. *Croat. Chem. Acta 1996, 69, 873-882*

[41] Lukovits, I.; Razinger, M.: On Calculation of the Detour Index. *J. Chem. Inf. Comput. Sci. 1997, 37, 283-286*

[42] Rücker, G.; Rücker, C.: Symmetry-Aided Computation of the Detour Matrix and the Detour Index. *J. Chem. Inf. Comput. Sci. 1998, 38, 710-714*

[43] Randić, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L.: Search for All Self–Avoiding Paths for Molecular Graphs. *Comput. & Chem. 1979, 3, 5-13*

[44] Randić, M.: Characterization of Atoms, Molecules, and Classes of Molecules Based on Paths Enumeration. *MATCH Commun. Math. Comp. Chem. 1979, 7, 5-64*

[45] Gálvez, J.; Garcìa, R.; Salabert, M. T.; Soler, R.: Charge Indexes. New Topologcal Descriptors. *J. Chem. Inf. Comput. Sci. 1994, 34, 520-525*

[46] Gálvez, J.; Garcìa-Domenech, R.; De Julián-Ortiz, V.; Soler, R.: Topological Approach to Drug Design. *J. Chem. Inf. Comput. Sci. 1995, 35, 272-284*

[47] Wildman, S. A.; Crippen, G. M.: Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci. 1999, 39, 868-873*

[48] Bonchev, D.: Novel Indices for the Topological Complexity of Molecules. *SAR QSAR Environ. Res. 1997, 7, 23-43*

[49] Rücker, G.; Rücker, C.: Automatic Enumeration of All Connected Subgraphs. *MATCH Commun. Math. Comp. Chem. 2000, 41, 145-149*

[50] Sharma, V.; Goswami, R.; Madan, A. K.: Eccentric Conectivity Index: A Novel Highly Discriminating Topological Descriptor for Structure-Property and Structure-Activity Studies. *J. Chem. Inf. Comput. Sci. 1997, 37, 273-282*

[51] Rücker, G.; Rücker, C.; Gutman, I.: On Kites, Comets, and Stars. Sums of Eigenvector Coefficients in (Molecular) Graphs. *Z. Naturforsch. A 2002, 57a, 143-153*

[52] Schultz, H. P.; Schultz, E. B.; Schultz, T. P.: Topological Organic Chemistry. 2. Graph Theory, Matrix Determinants and Eigenvalues, and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci. 1990, 30, 27-29*

[53] Needham, D. E.; Wei, I. C.; Seybold, P. G.: Molecular Modeling of the Physical Properties of the Alkanes. *J. Am. Chem. Soc. 1988, 110, 4186-4194*

[54] Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D.: Molecular Properties that Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem. 2002, 45, 2615-2623*

[55] Jurs, P. C.; Hasan, M. N.; Hansen, P. J.; Rohrbaugh, R. H.: Prediction of Physicochemical Properties of Organic Compounds from Molecular Structure. Pages 209-233 in Physical Property Prediction (Jochum, C., Ed.), *Springer, Berlin 1988*

[56] Rohrbaugh, R. H.; Jurs, P. C.: Description of Molecular Shape Applied in Studies of Structure/Activity and Structure/Property Relationships. *Anal. Chim. Acta 1987, 199, 99-109*

[57] Rohrbaugh, R. H.; Jurs, P. C.: Molecular Shape and the Prediction of HPLC Retention Indexes of Polycyclic Aromatic Hydrocarbons. *Anal. Chem. 1987, 59, 1048-1054*

[58] Kier, L. B.; Hall, L. H.: Molecular Structure Description. The Electrotopological State. *Academic Press, San Diego, CA, and London, 1999*

[59] Khadikar, P. V.; Deshpande, N.V.; Kale, P. P.; Dobrynin, A.; Gutman, I.; Dömötör, G.: The Szeged Index and an Analogy with the Wiener Index. *J. Chem. Inf. Comput. Sci. 1995, 35, 547-550*

[60] Gutman, I.; Klavzar, S.: An Algorithm for the Calculation of the Szeged Index of Benzenoid Hydrocarbons. *J. Chem. Inf. Comput. Sci. 1995, 35, 1011-1014*

[61] Zerovnik, J.: Computing the Szeged Index. *Croat. Chem. Acta. 1996, 69, 837-843*

[62] Zerovnik, J.: Szeged Index of Symmetric Graphs. *J. Chem. Inf. Comput. Sci. 1999, 39, 77-80*

[63] Ren, B.: Novel Atomic-Level-Based AI Topological Descriptors: Application to QSPR/QSAR Modeling. *J. Chem. Inf. Comput. Sci. 2002, 42, 858-868*

[64] Ren, B.: Atomic-Level-Based AI Topological Descriptors for Structure-Property Correlations. *J. Chem. Inf. Comput. Sci. 2003, 43, 161-169*

[65] Ren, B.: Novel Atom-Type AI Indices for QSPR Studies of Alcohols. *Comput. & Chem. 2002, 26, 223-235*

[66] Ren, B.: Application of Novel Atom-Type AI Topological Indices to QSPR Studies of Alkanes. *Comput. & Chem. 2002, 26, 357-369*

[67] Ren, B.: A New Topological Index for QSPR of Alkanes. *J. Chem. Inf. Comput. Sci. 1999, 39, 139-143*

[68] Bonchev, D.; Trinajstić, N.: Overall Molecular Descriptors. 3. Overall Zagreb Indices. *SAR QSAR Environ. Res. 2001, 12, 213-236*

[69] Bonchev, D.: The Overall Wiener Index – A New Tool for Characterization of Molecular Topology. *J. Chem. Inf. Comput. Sci. 2001, 41, 582-592*

[70] Bonchev, D.: Overall Connectivity – A Next Generation Molecular Connectivity. *J. Mol. Graphics Model. 2001, 20, 65-75*

[71] Bonchev, D.: Overall Connectivities/Topological Complexities: A New Powerful Tool for QSPR/QSAR. *J. Chem. Inf. Comput. Sci. 2000, 40, 934-941*

[72] Rücker, C.; Meringer, M.: How Many Organic Compunds are gt-nonplanar? *MATCH Commun. Math. Comput. Chem. 2002, 45, 159-172*

[73] Buckley, F.; Harary F.: Distance in Graphs. *Addison-Wesley, Redwood City, CA, 1990,* page 213

[74] Anonymus: Searching Properties in the CAS Registry File. *STNotes 2002, 28, 1-7*

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

[76] Augustin, V.: Computerunterstützte Berechnung von Symmetrien unscharfer Strukturen. *Diploma thesis, University of Bayreuth, 2004*

Chapter 4

Literature on MOLGEN–QSPR

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

A. Kerber, R. Laue, M. Meringer, C. Rücker: *Molecules in Silico: A Graph Description of Chemical Reactions.* J. Chem. Inf. Model. 47 (2007), 805-817.

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

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

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 (2005), 163-176.

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

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 (2004), 187-204.

M. Meringer: *Mathematische Modelle für die kombinatorische Chemie und die molekulare Strukturaufklärung.* PhD thesis, University of Bayreuth, 2004. *Logos-Verlag, xxxiv+354 pp., 2004, ISBN 3-8325-0673 -X.*

J. Braun: *Topologische Indizes und ihre computerunterstützte Anwendung in der Chemie.* Diploma thesis, University of Bayreuth, 1999.

Most of these papers may be downloaded in the form of preprints free of charge from the MOLGEN homepage at http://www.molgen.de