Visualization and Substructure Retrieval Tools in the MOGADOC Database (Molecular Gasphase Documentation)

Jürgen Vogt,* Natalja Vogt, and Rüdiger Kramer

Sektion für Spektren- und Strukturdokumentation, University of Ulm, D-89069 Ulm, Germany

Received August 23, 2002

Over the years the MOGADOC database, which covers the literature for gas-phase electron diffraction, microwave spectroscopy, and molecular radio astronomy from the inception of each method, has been developed. Recently visualization and substructure retrieval tools have been implemented. New features are described by means of typical database applications.

INTRODUCTION

To facilitate access to structural and related properties of free molecules, the Section for Spectra and Structure Documentation at the University of Ulm has compiled and critically evaluated the literature in the following fields:

- gas-phase electron diffraction back to 1930,
- microwave spectroscopy back to 1945,
- molecular radio astronomy back to 1965.

On this basis the MOGADOC database (the acronym stands for Molecular Gasphase Documentation) has been established. Sponsored by the German Federal Government, the first version was developed for VAX mainframes under VMS disk operating system about 20 years ago in close cooperation with the Fachinformationszentrum in Karlsruhe.[1] Over the years the database was developed via a MS-DOS to a MS Windows version[2−4] using the database management system "STN Personal File System" (PFS) from the software house "Gesellschaft für elektronische Informationsdienstleistungen Kramer&Hofmann".[5] Recently MOGADOC has been converted by means of PFS 3000 to an HTML-based version, which uses conventional WWW browsers applying Java scripts and cascading style sheets. The present update MOGADOC 2000 comprises more than 27 700 references for about 8500 inorganic, organic, and organometallic compounds and contains numeric data sets of structural data, such as bond lengths, bond angles, dihedral angles, etc. for about 4700 compounds.

The database is applied by some dozen research groups in the fields of electron diffraction, microwave spectroscopy, radio astronomy, crystallography, and computational chemistry. It is a very helpful tool for the comparison of structural data from different methods. Moreover, it is used as a source of structural information for thermodynamical and force field calculations, in the analysis of conformational and other structural effects, etc.[6] The MOGADOC database serves also as a information resource for some Landolt Börnstein Tables.[7−11]

GENERAL FEATURES

To facilitate literature and compound retrievals, the database consists of a bibliographic file (MGDLIT) and a compound file (MGDCOM). In the MGDLIT file, the keywords, which form a hierarchically controlled thesaurus (/CT) and which are selected by reviewers, describe the main aspects of the quoted documents. Both files are interconnected by means of hyperlinks. Graphic user interfaces enable the users to perform retrievals by means of bibliographic and/or compound related search terms, such as compound names, molecular formulas in the Hill system, chemical composition, bond lengths, and angles. The search strings can be combined by means of the Boolean operators AND, OR, and NOT as well as the proximity operators (A) and (W). The last ones are known from the Messenger retrieval syntax.

Helpful features, such as sorting of hitlists, user-defined formats, and comments, are available. A tabular representation option is implemented in the retrieval program for performing statistical analyses.[12]

In the following the new features, namely the visualization tool, which has been recently made available in the update MOGADOC 2000, and the substructure retrieval tool, which will be implemented in the forthcoming update, are described by means of typical applications.

VISUALIZATION TOOLS

The visualization tool is incorporated in order to represent the results of statistical analyses. It can be applied both to all documents and to user-defined subsets of the MGDLIT or MGDCOM files by means of dialogue boxes. The user can choose among the most common diagram types, which are known from spreadsheet programs. In general, the independent variables of the plot are identical with the search fields, whereas the dependent variables are always the number of entries. The dialogue boxes also allow the interactive adjustment of plots (scaling, labeling, previewing). The resulting diagrams can be printed in the usual way.

Visualization of User-Defined Subsets. An example of the visualization of user-defined subsets is given in Figure 1. The screen shot illustrates the growth of publications about the magnetic nuclear hyperfine structure of microwave spectra of diamagnetic van der Waals molecules and other gas-phase complexes during the s 1990s. The bar chart results from the retrieval of the publication interval

1990−1999/PY

* Corresponding author phone: (+49) 731 50-31050; fax: (+49) 731 50-31059; e-mail: Juergen.Vogt@Chemie.Uni-Ulm.de.

10.1021/ci020276l CCC: $25.00
Published on Web 02/07/2003

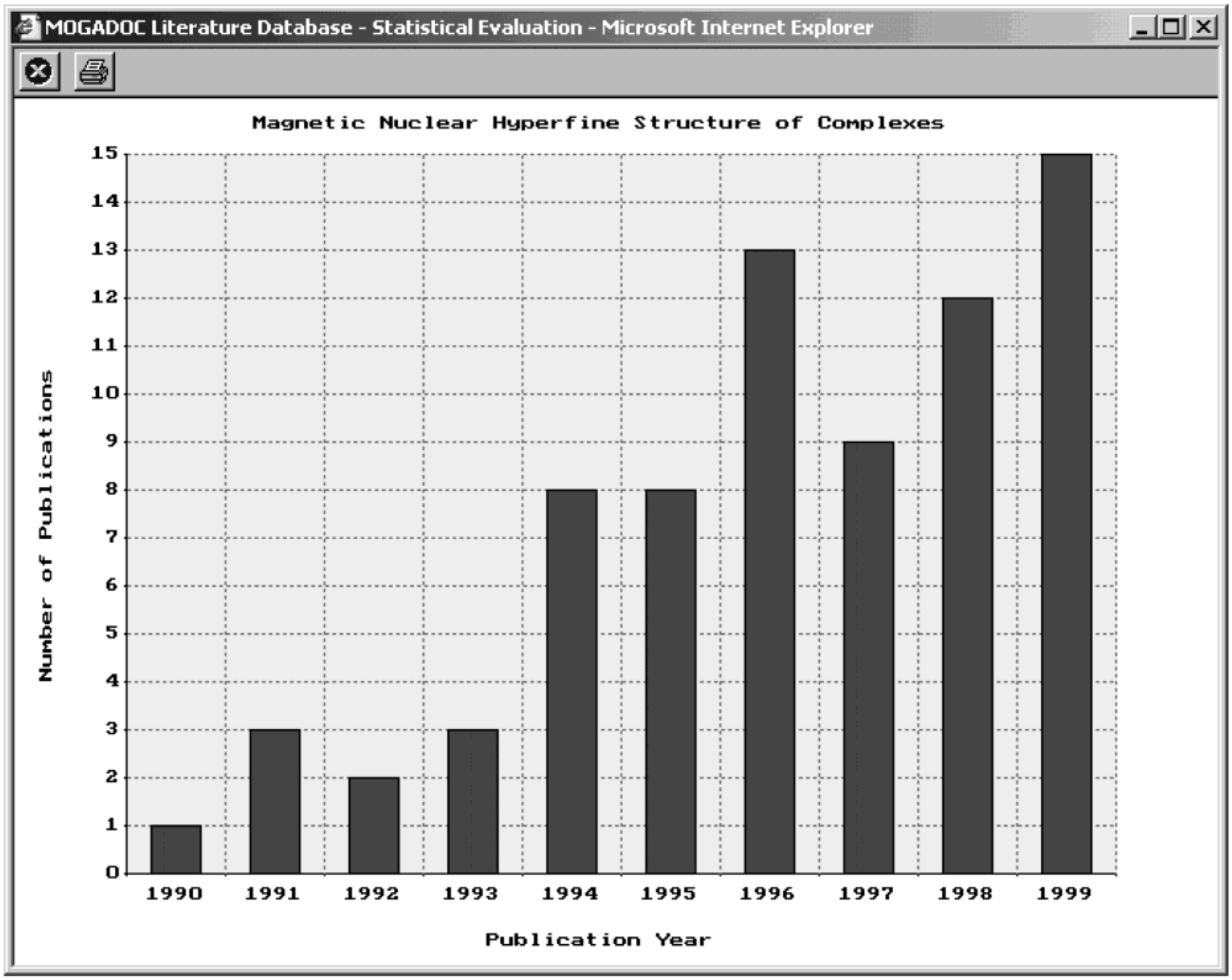

MOGADOC Literature Database - Statistical Evaluation - Microsoft Internet Explorer
Magnetic Nuclear Hyperfine Structure of Complexes
Number of Publications
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
Publication Year

Figure 1. Bar chart showing the growths of publications about magnetic nuclear hyperfine structure in spectra of weakly bound gas-phase complexes.

combined with the keyword search

((VDWAALS or COMPLEX or LBM+NT) and
(HFSR+NT or HFSN+NT))/CT

by means of the AND operator. Finally all conference abstracts

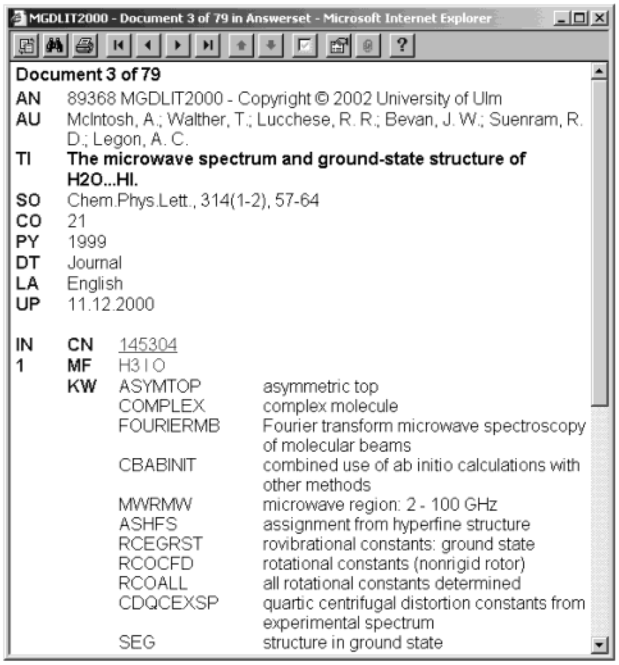
CONFERENCE/DT

are excluded by means of the NOT operator. The retrieval can be performed by means of the graphical user interface or in the expert mode by means of the list management. Only in the expert mode the corresponding search field codes /PY, /CT, and /DT have to be added to the search strings for the publication year, controlled term, and document type field, respectively.

The keywords have to be entered as abbreviations, which can be interactively selected from the thesaurus. Here the keywords COMPLEX and VDWAALS are abbreviations for “complexes” and “van der Waals compounds”, respectively. LBM is standing for a group of keywords devoted for “weakly bound molecules”, whereas HFSR and HFSN are abbreviations for the most general keywords for “magnetic nuclear spin rotation coupling” and “magnetic nuclear spin-nuclear spin coupling”, respectively. The relationship code NT (“narrow term”) automatically includes all more specific keywords of the given items, such as “chemical shielding from spin-rotation coupling”, “dipole−dipole interaction in spin−spin coupling”, etc. Figure 2 shows one of the resulting bibliographic entries which fulfills the logical requirement described above.

Visualization for Complete Files. In the MGDCOM file the visualization tools can only be applied to the specification of methods of structural investigations and to internal coordinates (bond lengths and angles). Because the tables of data sets consist of numeric values and element symbols in different search fields, the visualization tools cannot be directly and unambiguously applied to these data. To link the numeric values and the corresponding descriptions, predefined pseudoindices are temporarily created in the user directory.

The number of pseudoindices is limited now to the most frequent combinations of elements in the database, i.e., 15 for bond lengths (C−C, C−H, C−N, C−O, C−P, C−S, C−Si, C−F, N−O, N−Si, O−P, O−S, O−Si, P−S, P−Si) and 19 for bond angles (C−C−C, C−C−H, C−C−N, C−C−O, C−C−P, C−C−S, C−C−Si, C−N−C, C−O−C, C−P−C, C−S−C, C−Si−C, H−C−H, H−C−O, H−C−S, H−C−N, H−N−H, H−P−H, H−Si−H). It is emphasized that the bond orders are not specified in the numeric tables. Although the values of the internal coordinates are given by a varying number of significant decimal figures, the bond


MGDLIT2000 - Document 3 of 79 in Answerset - Microsoft Internet Explorer

Document 3 of 79

AN	89368 MGDLIT2000 - Copyright © 2002 University of Ulm		
AU	McIntosh, A.; Walther, T.; Lucchese, R. R.; Bevan, J. W.; Suenram, R. D.; Legon, A. C.		
TI	**The microwave spectrum and ground-state structure of H2O...HI.**		
SO	Chem.Phys.Lett., 314(1-2), 57-64		
CO	21		
PY	1999		
DT	Journal		
LA	English		
UP	11.12.2000		
IN	CN	145304	
1	MF	H3 I O	
	KW	ASYMTOP	asymmetric top
		COMPLEX	complex molecule
		FOURIERMB	Fourier transform microwave spectroscopy of molecular beams
		CBABINIT	combined use of ab initio calculations with other methods
		MWRMW	microwave region: 2 - 100 GHz
		ASHFS	assignment from hyperfine structure
		RCEGRST	rovibrational constants: ground state
		RCOCFD	rotational constants (nonrigid rotor)
		RCOALL	all rotational constants determined
		CDQCEXSP	quartic centrifugal distortion constants from experimental spectrum
		SEG	structure in ground state

Figure 2. Example of a bibliographic entry in the MGDLIT file.

lengths and angles are stored in the corresponding pseudoindices with a constraint predefined accuracy of 0.01 Å and 0.1°, respectively. By this truncation equal intervals between the temporally obtained numeric values are created.

Figure 3, which shows the distribution of C−N bond lengths of all compounds in the MGDCOM file, is an example of the visualization of a predefined pseudoindex. The peaks at about 1.15 and 1.46 Å can be easily assigned to triple and single bonds, respectively, whereas the double bond is not forming a clear peak. However, the user should keep in mind, that the compounds entries, which simultaneously consist of different C−N distances (e.g. single and double bond in *N*-methylenemethanamine, see also Figure 4), give a multiple contribution to this distribution.

NUMERIC RETRIEVAL OF INTERNAL COORDINATES

In contrast to the visualization of internal coordinates, which is restricted by the predefined pseudoindices, all numeric bond lengths and angles can be retrieved. This feature can also be applied in a subsequent step for searching compounds, which contribute to C−N bond lengths, for example, between 1.25 and 1.35 Å (see Figure 3).

As can be seen in Figure 4, the numeric tables in general consists of many lines, each containing the element symbols (including nonsearchable locants in the structural formula) and the corresponding numeric values of that bond length or bond angle. Since in the numeric tables the bonds are stored by the element symbols in any order, the element symbols have to be combined by the proximity

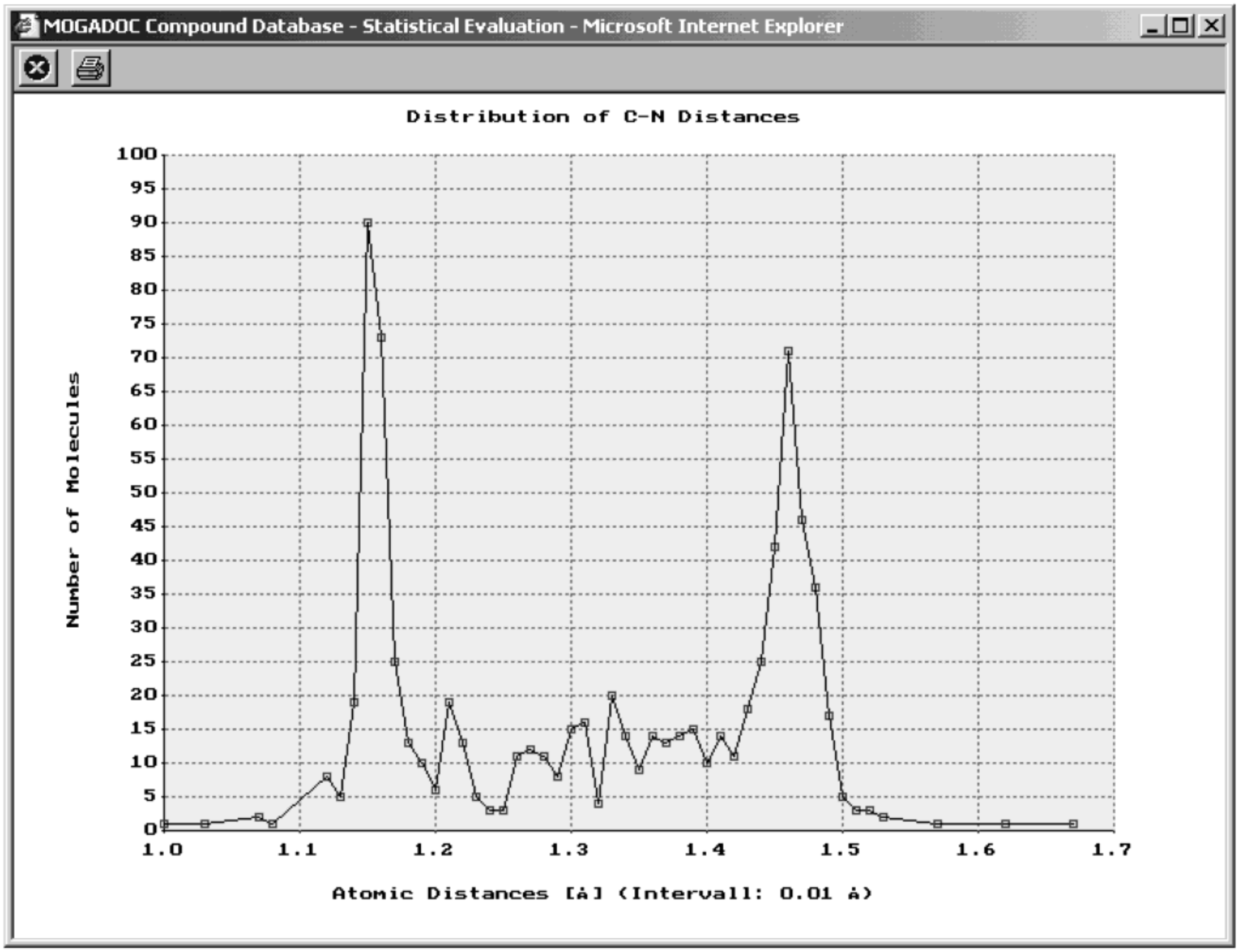
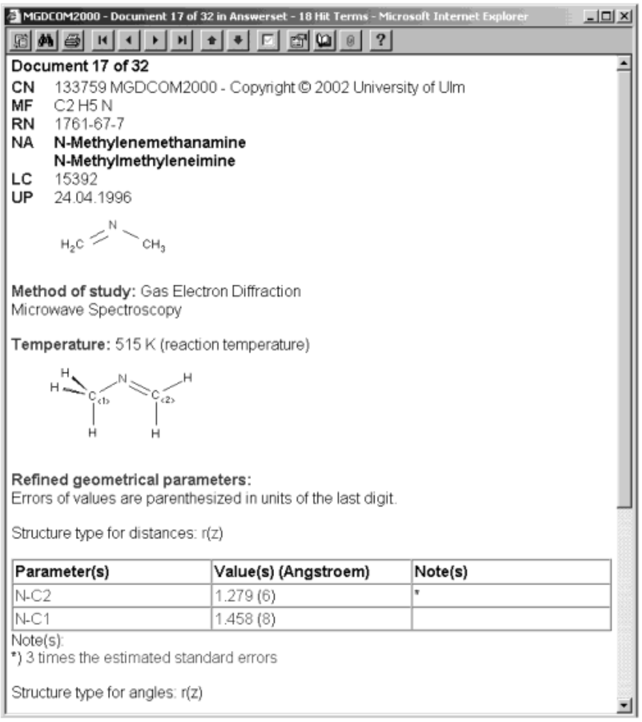


Figure 3. Scatter chart showing the distribution of numeric C−N bond lengths in the compound entries of the MGDCOM file.


MGDCOM2000 - Document 17 of 32 in Answerset - 18 Hit Terms - Microsoft Internet Explorer
?
H_2C
N
CH_3
H
H
C<1>
N
C<2>
H
H
H

Document 17 of 32

CN 133759 MGDCOM2000 - Copyright © 2002 University of Ulm
MF C2 H5 N
RN 1761-67-7
NA **N-Methylenemethanamine**
N-Methylmethyleneimine
LC 15392
UP 24.04.1996

Method of study: Gas Electron Diffraction
Microwave Spectroscopy

Temperature: 515 K (reaction temperature)

Refined geometrical parameters:
Errors of values are parenthesized in units of the last digit.

Structure type for distances: r(z)

Parameter(s)	Value(s) (Angstroem)	Note(s)
N-C2	1.279 (6)	*
N-C1	1.458 (8)	

Note(s):
*) 3 times the estimated standard errors

Structure type for angles: r(z)

Figure 4. Example of a compound entry in the MGDCOM file.

operator (A) in the search field ATD (“Atoms defining Distances”)

$$(C\ (A)\ N)/ATD\ (L)\ 1.25-1.35/DIS$$

where the link operator (L) requires that the desired numeric values belong to the same table line.

It is emphasized that in the graphical user interface, which was designed for standard retrievals, the input boxes, which are assigned to particular search fields, can only be combined by the Boolean operators. Because the application of the AND operator would lead to false positive hits, this kind of numeric retrieval has to be performed in the expert mode of the list management window.

The resulting hitlist can be narrowed down, for example, for all organic compounds with two carbon atoms in the specific element count search field C: 2/C. Figure 4 shows one of the resulting compound entries which fulfills the logical requirement described above.

SUBSTRUCTURE RETRIEVAL

The implementation of a structure and substructure retrieval tool into the forthcoming MOGADOC update is possible due to the fortunate fact that the structural formulas have been already stored as connectivity tables in MOL format.[13] Presently the corresponding module is being developed in close cooperation with Dr. W. D. Ihlenfeldt, formerly of the University of Erlangen-Nürnberg. It is a part of the chemistry information toolkit CACTVS.[14,15] As in other structure databases the substructure retrieval is performed in two steps, the textual search of screens followed by a atom-by-atom and bond-by-bond matching.

The Java-based structure editor of the software house “enso Software”[16] is integrated both in the graphical user interface and in the list management window of the expert mode. The drawn fragments can be retrieved either in the “Exact Structure Search” or in the “Substructure Search” mode, which both can be combined by any fact retrieval. Figure 5 shows an adamantane-type fragment as an example; here one bridge atom is substituted by the atom list consisting of carbon, silicon, germanium, and tin. The retrieval in the “Substructure Search” mode presently yields the entries for 1-germaadamantane and 1-silaadamantane besides the entries for adamantane and its substituted derivatives (mainly at the bridge atom).

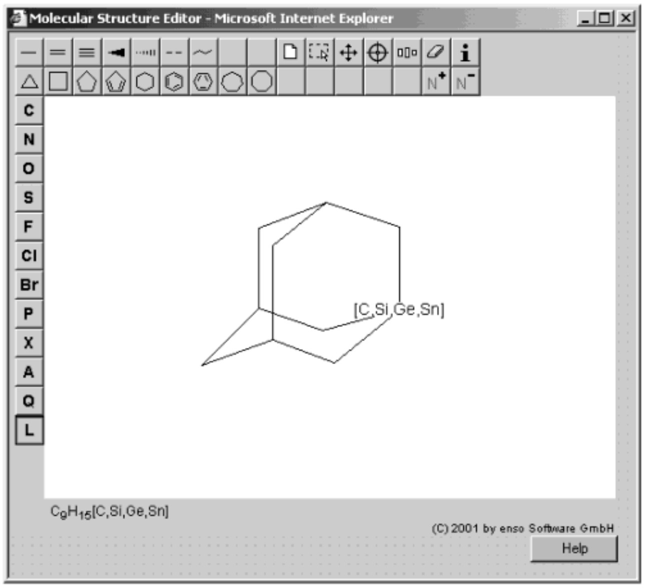

Molecular Structure Editor - Microsoft Internet Explorer
i
N+
N−
C
N
O
S
F
Cl
Br
P
X
A
Q
L
[C,Si,Ge,Sn]
C_9H_{15}[C,Si,Ge,Sn]
(C) 2001 by enso Software GmbH
Help

Figure 5. Java-based structure editor with an adamantane-type molecular fragment.

ACKNOWLEDGMENT

The project was supported by the Dr. B. Mez-Starck Foundation. The database has been developed over many years due to extensive and very valuable contributions of Dr. Barbara Mez-Starck until her death in May 2001.

REFERENCES AND NOTES

(1) Lohr, A.; Mez-Starck, B.; Schirdewahn, H. G.; Watson, D. G. MOGADOC (Molecular Gasphase Documentation) − An Interactive Computerized Search/Retrieval System. *J. Mol. Struct.* **1983**, *97*, 57−60.
(2) Vogt, J. Structural Chemical Database for Gasphase Compounds. *Struct. Chem.* **1992**, *3*, 147−153.
(3) Vogt, J. MOGADOC − A Bibliographic Numerical Resource for Gasphase Molecular Spectroscopy and Structure. *J. Mol. Spectrosc.* **1992**, *155*, 413−416.
(4) Vogt, J.; Mez-Starck, B.; Vogt, N.; Hutter, W. MOGADOC − A Database for Gasphase Molecular Spectroscopy and Structure. *J. Mol. Struct.* **1999**, *485−486*, 249−254.
(5) http://www.interhost.de.
(6) Rehm, D.; Sheldrick, G. M.; Boese, R. private communications.
(7) Demaison, J.; Hübner, W.; Hüttner, W.; Vogt, J.; Wlodarczak, G. *Molecular Constants Mostly from Microwave, Molecular Beam, and Sub-Doppler Laser Spectroscopy: Dipole Moments, Quadrupole Coupling Constants, Hindered Rotation and Magnetic Constants of Diamagnetic Molecules*; Hüttner, W., Ed.; Springer: Berlin, 2002; Landolt-Börnstein New Series II Vol. 24C, p 296.
(8) Graner, G.; Hirota, E.; Iijima, T.; Kuchitsu, K.; Ramsay, D. A.; Vogt, J.; Vogt, N. *Structure Data of Free Polyatomic Molecules. Inorganic Molecules*; Kuchitsu, K., Ed.; Springer-Verlag: Berlin, 1998; Landolt-Börnstein New Series II, Vol. 25A, p 359.
(9) Graner, G.; Hirota, E.; Iijima, T.; Kuchitsu, K.; Ramsay, D. A.; Vogt, J.; Vogt, N. *Structure Data of Free Polyatomic Molecules. Molecules Containing One or Two Carbon Atoms*; Kuchitsu, K., Ed.; Springer-Verlag: Berlin, 1999; Landolt-Börnstein New Series II, Vol. 25B, p 512.

(10) Graner, G.; Hirota, E.; Iijima, T.; Kuchitsu, K.; Ramsay, D. A.; Vogt, J.; Vogt, N. *Structure Data of Free Polyatomic Molecules. Molecules Containing Three or Four Carbon Atoms*; Kuchitsu, K., Ed.; Springer-Verlag: Berlin, 2000; Landolt-Börnstein New Series II, Vol. 25C, p 481.

(11) Graner, G.; Hirota, E.; Iijima, T.; Kuchitsu, K.; Ramsay, D. A.; Vogt, J.; Vogt, N. *Structure Data of Free Polyatomic Molecules. Molecules Containing Five or More Carbon Atoms*; Kuchitsu, K., Ed.; Springer-Verlag: Berlin, 2003; Landolt-Börnstein New Series II, Vol. 25D, in press.

(12) Vogt, J.; Vogt, N.: Statistical Tools of the MOGADOC Database (Molecular Gasphase Documentation). *Struct. Chem.* **2003**, *14*, 137−141.

(13) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical-Structure File Formats Used by Computer-Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.

(14) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, S.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach Toward Modularity and Flexibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109−116.

(15) http://www2.ccc.uni-erlangen.de/software/cactvs/index.html.

(16) http://www.enso-software.com.

CI020276L