

- (4) Domalski, E. S.; Hearing, E. D. *J. Phys. Chem. Ref. Data* **1988**, *17*, 4.
- (5) Ritter, E. R.; Bozzelli, J. W. Submitted to *Int. J. Chem. Kinet.*, Oct 1990.
- (6) Muller, C.; Scacchi, G.; Come, G. M. Presented at the AIChE 77th Annual Meeting, San Francisco, 1984.
- (7) Frurip, D. J.; Freedman, E.; Hertel, G. R. *Proc. Int. Symp. Runaway React.* **1989**.
- (8) Hoffmann, J. M.; Maser, D. C., Eds. *Chemical Hazard Process Review*; ACS Symposium Series 274; American Chemical Society: Washington, D.C., 1985.
- (9) Kee, R. J.; Miller, J. A.; Jefferson, T. H. Sanida Report No. SAND80-8003; 1980.
- (10) Lund, C. M. Lawrence Livermore National Laboratory Report No. UCRL-52504; 1978.
- (11) Reynolds, W. C. *STANJAN Version 3*; Stanford University: Stanford, 1986.
- (12) Cohen, N. Aerospace Report No. ATR-88 (7073)-2; Aerospace Corp.: El Segundo, CA, 1988.
- (13) Stein, S. E.; Golden, D. M.; Benson, S. W. *J. Phys. Chem.* **1977**, *81*, 4.
- (14) Shaw, R.; Golden, D. M.; Benson, S. W. *J. Phys. Chem.* **1977**, *81*, 18.
- (15) Stein, S. E.; Golden, D. M.; Benson, S. W. *J. Phys. Chem.* **1977**, *81*, 4.
- (16) Eigenmann, H. K.; Golden, D. M.; Benson, S. W. *J. Phys. Chem.* **1973**, *77*, 13.
- (17) Stein, S. E.; Fahr, A. *J. Phys. Chem.* **1985**, *89*, 3714.
- (18) Ritter, E. R.; Bozzelli, J. W. Manuscript in preparation, 1991.
- (19) This work.
- (20) Benson, S. W.; Buss, H. H. *J. Chem. Phys.* **1958**, *29*, 546.
- (21) Benson, S. W.; Cruickshank, F. R.; Golden, D. M.; Haugen, G. R.; O'Neal, H. E.; Rogers, A. S.; Shaw, R.; Walsh, R. *Chem. Rev.* **1969**, *69*, 279.
- (22) Burcat, A.; Zeleznik, F. J.; McBride, B. J. NASA Technical Memorandum No. 83800; 1985.
- (23) McBride, B. J.; Gordon, S. NASA Technical Memorandum, NASA TN-D-4097; 1967.
- (24) Hanson, R. J.; Haskell, K. H. Sandia National Laboratories Report, SAND77-0552; 1978.
- (25) Hanson, R. J.; Haskell, K. H. Sandia National Laboratories Report, SAND78-1290; 1979.
- (26) Dorafeeva, O. V.; Gurvich, L. V.; Jorish, V. S. *J. Phys. Chem. Ref. Data* **1986**, *15*, 2.
- (27) Ritter, E. R. Ph.D. Thesis, New Jersey Institute of Technology, 1989.
- (28) Ritter, E. R.; Bozzelli, J. W. *Chem. Phys. Proc. Combust.*; Paper 18, Eastern States Combustion Meeting, The Combustion Institute: Albany, 1989.
- (29) Rice, O. K. *Statistical Mechanics Thermodynamics and Kinetics*; W. H. Freeman and Company: San Francisco, 1967.
- (30) Soontag, R. E.; Van Wylen, G. J. *Fundamentals of Statistical Thermodynamics*; Robert E. Krieger Publishing Co.: Malabar, FL, 1985.
- (31) Herzberg, G. *Molecular Spectra and Molecular Structure II. Infrared and Raman Spectra of Polyatomic Molecules*; Van Nostrand Reinhold Co.: New York, 1945.
- (32) JANAF Thermochemical Tables. *J. Phys. Chem. Ref. Data* **1985**, *14*, 1.
- (33) Aly, F. A.; Lee, L. L. *Fluid Phase Equilib.* **1981**, *6*, 169.
- (34) Fakeeha, A.; Kache, A.; Rehman, Z. U.; Shoup, Y.; Lee, L. L. *Fluid Phase Equilib.* **1983**, *11*, 225.
- (35) Rehman, Z. U.; Lee, L. L. *Fluid Phase Equilib.* **1985**, *22*, 21.
- (36) Gardiner, W. C.; Burcat, In *Combustion Chemistry*, Gardiner, W. C., Jr., Ed.; Springer-Verlag: New York, 1984.
- (37) Reklaitis, G. V.; Ravindran, A.; Ragsdell, K. M. *Engineering Optimization*; John Wiley and Sons: New York, 1983.
- (38) Stull, D. R.; Westrum, E. F., Jr.; Sinke, G. C. *The Chemical Thermodynamics of Organic Compounds*; Robert E. Krieger Publishing Co.: Malabar, FL, 1987.
- (39) Pedley, J. B.; Naylor, R. D.; Kirby, S. P. *Thermochemical Data of Organic Compounds*; Chapman and Hall: New York, 1986.

## An Integrated Approach to Three-Dimensional Information Management with MACCS-3D<sup>1</sup>

OSMAN F. GÜNER,\* DAVID W. HUGHES, and LISE M. DUMONT

Molecular Design Limited, 2132 Farallon Drive, San Leandro, California 94577

Received January 18, 1991

In the past decade, the scientific community has realized the value of three-dimensional (3D) structural information and '3D searching' has started to become an important new methodology for computer-aided drug design. During this time, molecular modeling information generated from various sources has proliferated due to the growing availability of software and hardware and the increasing use of crystallographic and spectroscopic techniques. This information needs to be organized to allow for its effective storage and retrieval. This paper presents an approach to address this problem with a recently introduced program, MACCS-3D. In particular, this approach utilizes MACCS-3D's capability of handling data specific for atoms and atom pairs. With this software, various biological, computational, and spectroscopic data can be merged, allowing scientists from different disciplines to access and use this information more efficiently.

### INTRODUCTION

In the past decade, computational chemistry, molecular modeling, and spectroscopic and crystallographic methods of 3D structural elucidation have become standard techniques in new chemical and drug design and have found several applications in a variety of chemical research areas. New developments in technology also affect the methods of information transfer; activities that were once considered impossible are commonplace today. The ability to interface instruments to data-collecting devices allows for the easy generation of vast amounts of data. In addition, communication and interfacing technology allows research sites to bring together different types of data from various experiments, permitting the scientist to merge information from multiple sources.<sup>2</sup>

Baker analyzes scientific information barriers across international borders: "users demand more, new, timely, high-quality, and complete information services; developing technology induces increased efficiency and effectiveness in in-

formation flow; and economics force sharing of resources and bartering and exchange of information wherever possible".<sup>3</sup> This analysis is also applicable to the information barriers across disciplines, even within the same organization.

For example, Brown explains that the development of a new human drug may involve the collaborative efforts of representatives from 30 to 50 distinct scientific disciplines. He then emphasizes the need for a multidisciplinary requirement for collecting, indexing, storing, retrieving, evaluating, and disseminating information over a decade of time between project definition and market introduction.<sup>4</sup> There are, however, obvious technological limitations in integrating such information. For example, the programs that computational chemists use to determine and later store the 3D structures and data are highly specialized and unwieldy to most other researchers in an organization. And programs historically used for managing chemical information corporate-wide do not have the capability to handle data specific to certain parts of

molecules; while they successfully manage data that describe the *entire* structure, they generally cannot handle data specific to individual *atoms* and *bonds*.

As understanding of the value of 3D structural information has increased, and as necessary hardware resources have become more affordable, 3D molecular databases and computational chemistry/molecular modeling programs have found applications in many research areas. As a result, 3D structures and data available from various sources, both proprietary and public, have proliferated. With this proliferation of data comes an increased need to access, share, and use 3D structural information within the multidisciplinary environment that is typical of today's research environment.

The management of the various types of 3D data available in modern research laboratories is in most cases not fully integrated within a research organization and thus not providing full benefits. Specific data are easily accessible only by a small portion of the research community, only by the specialists who generate them. In general, for example, it is only the computational chemists who can quickly locate specific conformational data, only the analytical chemists who can quickly locate spectroscopic data, and so on. Data are not readily accessible by other researchers from other departments and are not readily correlated with related data generated by other researchers. Because specific 3D data remain separate from other information generated by a research organization, valuable insight that could be gained from such data is often missed—insight that could, for example, substantially reduce the number of chemical or drug candidates to be synthesized and screened.

Consider that the average research and development expenditure for introduction of a new drug is given as \$70,<sup>5</sup> \$150,<sup>6</sup> and \$231<sup>7</sup> million by three sources; the more recent the source, the higher the cost estimate. More than 10 000 new chemical compounds are synthesized and tested in a pharmaceutical company's laboratories for merely one to reach the market.<sup>5,6</sup> Molecular modeling studies can provide information to significantly reduce the candidates needed (and reduce the time and cost of synthesizing and screening) by making the requirements for a successful candidate more apparent from the beginning. However, if the information gained by these studies is not readily available to a large part of the multidisciplinary research team, this valuable insight is easily missed.

For example, nuclear Overhauser enhancement (NOE) and *J*-coupling information generated from an NMR study by a spectroscopist, if available to a computational chemist, can give the computational chemist some accurate interatomic distances and torsional angles of a flexible molecule in solution phase. The computational chemist can then reevaluate the geometry of the molecule by fixing the positions of the protons and the torsional angles as verified by the spectroscopist. The new results then suggest to medicinal chemists a reasonable conformation that the molecule is more likely to assume in solution during interacting with the receptor. Thus, they develop a more accurate pharmacophore with which they can search corporate and commercial 3D databases for molecules that contain the pharmacophore, molecules that may be quality new leads.

Some examples already exist in industry that indicate a recognition of the importance of integrating scientific information. Most earlier work involved integrating chemical and biological information.<sup>8</sup> Hagadone and Lajiness demonstrate an example of using a relational database management system (RDBMS) to integrate structural and biological information.<sup>9</sup> They point out that the benefits of this system over the traditional approach are the reduced duplication of software functionality and training requirements and the ability to easily retrieve a broad range of data in unanticipated ways.<sup>9</sup> Barcza

et al. generated an integrated in-house preclinical research and development database<sup>10</sup> containing chemical structures with spectroscopic, chemical, physicochemical, and biological information using the MACCS-II system.<sup>11</sup>

All of the above approaches, however, are limited due to the lack of per-atom and atom-pair data handling capabilities of their DBMS. This capability has only recently become available with the introduction of MACCS-3D.<sup>12</sup> This program provides organized storage, searching, and retrieval of static 3D molecular models and model-related data, including atom and atom-pair data. It offers an interactive, customizable, and graphically oriented database management program with which chemical researchers can systemically access both commercial and proprietary 3D models and associated data.

Sharing of information is a necessary first step to a successful multidisciplinary collaboration. If scientific information generated by researchers in various disciplines throughout an organization can be stored in a *single pool*, with such data easily accessible by everyone who needs it, this all-important first step is taken. This paper demonstrates how such vital sharing of information can be accomplished with a technology available today. For demonstration purposes, a small database was built with MACCS-3D<sup>12</sup> using representative data from various spectroscopic and computational applications. The idea is to demonstrate the handling of data from entirely different sources that will be used in different ways by scientists from different disciplines, *all within a single database*. This way, each end-user uses an identical technology and standard user interface, thereby, overcoming one of the major obstacles in integration of chemical information. Bawden recently commented that integration of chemical information will only be truly effective when it operates at the level of the "single apparent system".<sup>13</sup> We are taking this approach further by operating at the level of a "single actual system".

## METHODS

The spectroscopic data is obtained from a standard textbook;<sup>14</sup> it involves partial data from ultraviolet (UV), infrared (IR), mass spectroscopic (MS), and nuclear magnetic resonance (NMR) analysis for the compounds in the database. We obtained the computational data for the same compounds with the AMPAC series of programs<sup>15</sup> utilizing the AM1 method;<sup>16</sup> vibrational frequency calculations followed the full geometry optimizations.

A MACCS-3D database contains per-model, per-atom, and atom-pair data as either numeric, text, or formatted data; the atom-pair data does not require two atoms to be connected. Table I lists the per-model data fields of the database, and Table II lists the per-atom and atom-pair data fields.<sup>17</sup>

Obviously, there are some costs associated with maintaining a multidisciplinary database. One cost is data redundancy; i.e., much of the data may reside in more than one place in the computer system. This may, however, provide advantages such as data integrity, validation, and protection. We suggest that each division or project maintain the data generated by their own group in a local departmental database generated with MACCS-3D. For example, spectroscopists could create and use a database of spectroscopic information, and computational chemists could use a database composed of computational data. In case these groups need to access the same general structural information, the appropriate data could be copied from the corporate database into the local databases. MACCS-3D also lets you protect fields (such as chemical structure), in local databases as necessary, which in turn enforces consistency from one database to another.

Of course, creating local databases could lead to a redundancy of tasks, as one person in each division will need to maintain each database. Knowledge about the maintenance

**Table I.** Per-Model Data Fields of the Example Database

type of data	instrument	description
spectroscopic	<sup>1</sup> H NMR	solvent source comments
	UV	$\lambda_{\max}$ log $\epsilon$
	MS	<i>m/e</i> parent <i>m/e</i> base M+1 % of M M+2 % of M
computational	energies	MM2 energy heat of formation ionization potential HOMO energy LUMO energy zero point energy
other		computational method occupied molecular orbitals unoccupied molecular orbitals vibrational frequencies

**Table II.** Per-Atom and Atom-Pair Data Fields of the Example Database

type of data	instrument	description
per atom	<sup>1</sup> H NMR	chemical shifts multiplicities
	<sup>13</sup> C NMR	chemical shifts
	Comp-Chem	HOMO atomic orbital coefficients LUMO atomic orbital coefficients net atomic charges atom electron densities
atom pair	<sup>1</sup> H NMR	coupling constants
	IR	bond-stretching frequencies
	Comp-Chem	bond orders

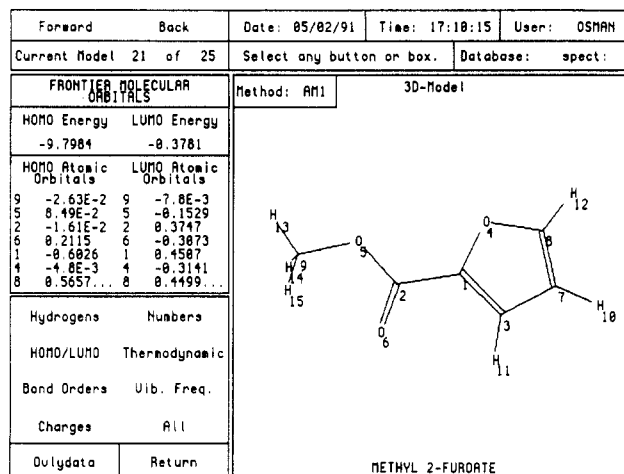
and administration of local databases lends researchers a greater understanding and appreciation of the corporate database. All too often, the corporate database is a "black box" to researchers, a mystery understood only by computer science personnel. By implementing local databases, researchers can unravel the mystery and perhaps discover how to better use corporate-wide information in their research.

Periodically the data in the departmental databases are transferred electronically to the central corporate database and merged by the corporate database supervisor. This also costs time and effort over individual registration of data by researchers. But in turn, it provides an additional opportunity to validate and check the data for consistency and completeness. The corporate database supervisor is in a position to monitor and control the flow of data into the corporate database, a process made more effectively by using background processing, database access, and command files. In a similar fashion, departmental database administrators are responsible for updating local databases with information from the corporate database. This entire process is a simple example of distributed system processing, which is rapidly becoming the standard for database access. In a networked environment, client/server programs will take over many of the tasks of data transfer control and database updating.

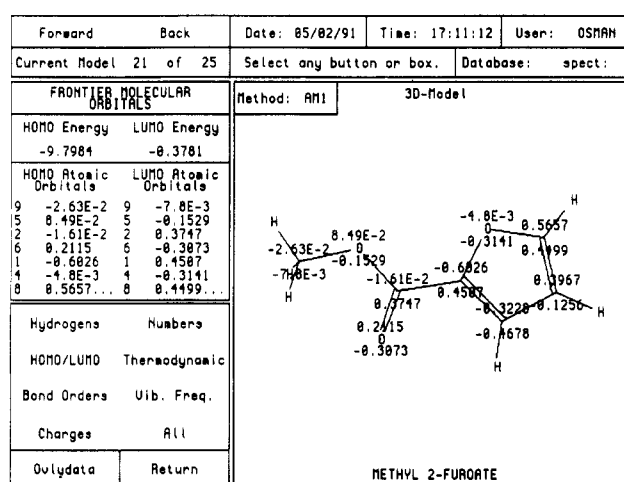
## DISCUSSION

How would researchers use such a database in a multidisciplinary environment? First, a variety of data related to a compound in a list resulting from any type of search—whether a substructure, similarity, or submodel search—can be readily available to the user. Hence, if a scientist conducts a search, he/she will now be able to easily access the spectroscopic and

a



b

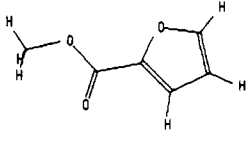
**Figure 1.** AM1 calculated (a) frontier molecular orbital data of methyl 2-furoate and (b) atomic orbital coefficients overlaid on the structure.

computational characteristics of the compounds in the hit list. For example, with the availability of the computational data, it is possible to compare the partial charges on a specific pharmacophoric group of each compound in a hit list in an attempt to differentiate among the active and inactive compounds via physical properties.

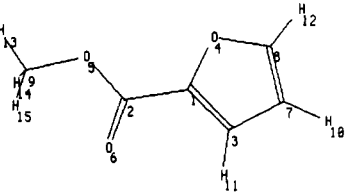
Using the MACCS-II System's Customization Module,<sup>18</sup> researchers can generate forms that display the information in any desired fashion. For example, Figure 1a displays the frontier molecular orbital (FMO) energies—i.e., information on the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) of methyl 2-furoate. Using a customized form, the researcher can overlay the per-atom and atom-pair data on the structure as shown in Figure 1b. Here, using different colors, the display differentiates HOMO and LUMO atomic orbital coefficients (not apparent with the black-and-white figure in this paper). Figure 2, parts a–c, displays three other pages of computational data showing thermodynamic data, bond orders, and vibrational frequencies, respectively. A form can contain more than a single page of information as shown in Figure 3a. This display can also overlay the HOMO/LUMO atomic orbital coefficients, net atomic charges, and  $\pi$ -electron densities on the compound as shown in Figure 3b; when a certain box gets too crowded, as in this case, it is possible to zoom into a particular box (Figure 3c).

In a similar way, customized forms can display spectroscopic data. Figure 4, parts a and b, respectively, gives examples for <sup>13</sup>C NMR chemical shifts and for <sup>1</sup>H NMR chemical shifts,

a

Forward	Back	Date: 85/02/91	Time: 17:12:01	User: OSMAN
Current Model 21 of 25		Select any button or box. Database: spect:		
THERMODYNAMIC DATA		Method: AM1 3D-Model		
Heat of Formation	Ionization Potential	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
-74.982	9.798			
Hydrogens	Numbers	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
HOMO/LUMO	Thermodynamic	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
Bond Orders	Uib. Freq.	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
Charges	All	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
Onlydata	Return	METHYL 2-FUROATE		

b

Forward	Back	Date: 85/02/91	Time: 17:12:58	User: OSMAN
Current Model 21 of 25		Select any button or box. Database: spect:		
BOND ORDERS		Method: AM1 3D-Model		
5 9 0.95				
2 5 1.834		<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
2 6 1.776		<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
1 2 0.953		<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
1 4 1.877		<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
4 8 1.14		<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
7 8 1.63		<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
3 7 1.224		<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
1 3 1.592		<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
Hydrogens	Numbers	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
HOMO/LUMO	Thermodynamic	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
Bond Orders	Uib. Freq.	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
Charges	All	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
Onlydata	Return	METHYL 2-FUROATE		

c

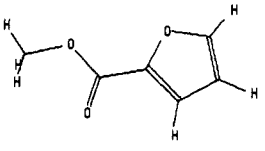
Forward	Back	Date: 85/02/91	Time: 17:13:28	User: OSMAN
Current Model 21 of 25		Select any button or box. Database: spect:		
VIBRATIONAL FREQUENCIES		Method: AM1 3D-Model		
Heat of Formation	Zero Point Energy	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
-74.982	72.556			
Hydrogens	Numbers	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
HOMO/LUMO	Thermodynamic	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
Bond Orders	Uib. Freq.	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
Charges	All	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
Onlydata	Return	METHYL 2-FUROATE		

Figure 2. AM1 calculated computational data is displayed: (a) thermodynamic data, (b) bond orders, and (c) vibrational frequencies.

proton multiplicities, and *J*-couplings. Analytical chemists can use such a database to store data on their analyzed structures. They can then utilize it to assist analysis of unknown compounds. For example, a spectroscopist may have difficulties in assigning a carbon to a certain peak in a <sup>13</sup>C NMR; by searching for those compounds with similar chemical shifts and identifying the chemical environment in those compounds in the database, the spectroscopist can hypothesize the environment of the unassigned carbon. Conversely, if the problem is in assigning a peak to a specific carbon, the spectroscopist

a

Forward	Back	Numbers	Select any button or box.
Current Model 21 of 25		Done	COMPUTATIONAL RESULTS
Only-FMO	Only-Charge	Only-BO	Heat of Formation -74.982 kcal/mol
3D Model		Method: AM1	Zero Point Energy 72.556 kcal/mol
			Ionization Poten. 9.798 kcal/mol
		HOMO Energy	LUMO Energy
		-9.7984	-0.3781
		HOMO Atomic Orbitals	LUMO Atomic Orbitals
		9 -2.63E-2	9 -7.8E-3
		5 8.49E-2	5 -0.1529
		2 -1.61E-2	2 0.3747
		6 0.2115	6 -0.3873
		1 -0.6826	1 0.4587
		4 -1.8E-3	4 -0.3141
		8 0.5657	8 0.4499
		7 0.3967	7 -0.1256
		3 -0.3228	3 -0.4678
			Bond Orders
			5 9 0.95
			2 5 1.834
			2 6 1.776
			1 2 0.953
			1 4 1.877...
			Net Atomic Charges
			13 0.189
			9 -7.5E-2
			5 -0.258
			2 0.383
			15 8.9E-2
			6 -0.347
			1 -9.6E-2
			4 -6.4E-2
			8 -5.7E-2
			7 -0.22
			3 -1.0...
			Electron Densities
			13 0.891
			9 4.875
			5 6.258
			2 3.617
			15 0.911
			6 6.347
			1 4.896
			4 6.064
			8 4.857
			7 4.22
			3 4.1...

b

Forward	Back	Numbers	Select any button or box.
Current Model 21 of 25		Done	COMPUTATIONAL RESULTS
Only-FMO	Only-Charge	Only-BO	Heat of Formation -74.982 kcal/mol
3D Model		Method: AM1	Zero Point Energy 72.556 kcal/mol
			Ionization Poten. 9.798 kcal/mol
		HOMO Energy	LUMO Energy
		-9.7984	-0.3781
		HOMO Atomic Orbitals	LUMO Atomic Orbitals
		9 -2.63E-2	9 -7.8E-3
		5 8.49E-2	5 -0.1529
		2 -1.61E-2	2 0.3747
		6 0.2115	6 -0.3873
		1 -0.6826	1 0.4587
		4 -1.8E-3	4 -0.3141
		8 0.5657	8 0.4499
		7 0.3967	7 -0.1256
		3 -0.3228	3 -0.4678
			Bond Orders
			5 9 0.95
			2 5 1.834
			2 6 1.776
			1 2 0.953
			1 4 1.877...
			Net Atomic Charges
			13 0.189
			9 -7.5E-2
			5 -0.258
			2 0.383
			15 8.9E-2
			6 -0.347
			1 -9.6E-2
			4 -6.4E-2
			8 -5.7E-2
			7 -0.22
			3 -1.0...
			Electron Densities
			13 0.891
			9 4.875
			5 6.258
			2 3.617
			15 0.911
			6 6.347
			1 4.896
			4 6.064
			8 4.857
			7 4.22
			3 4.1...

c

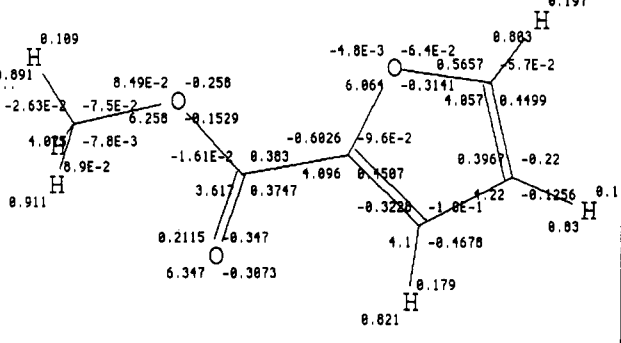
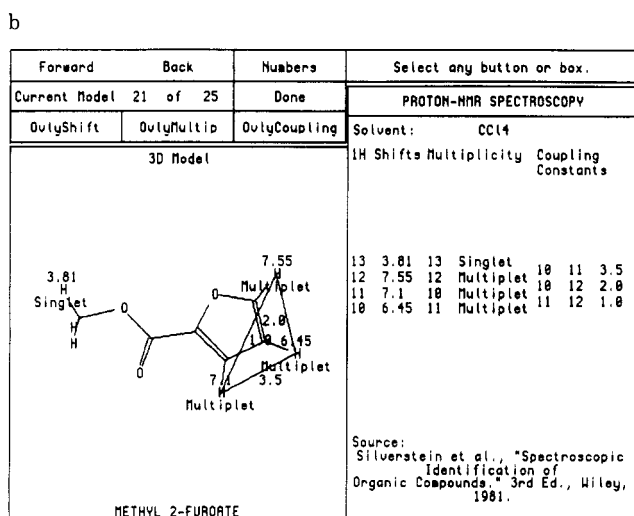
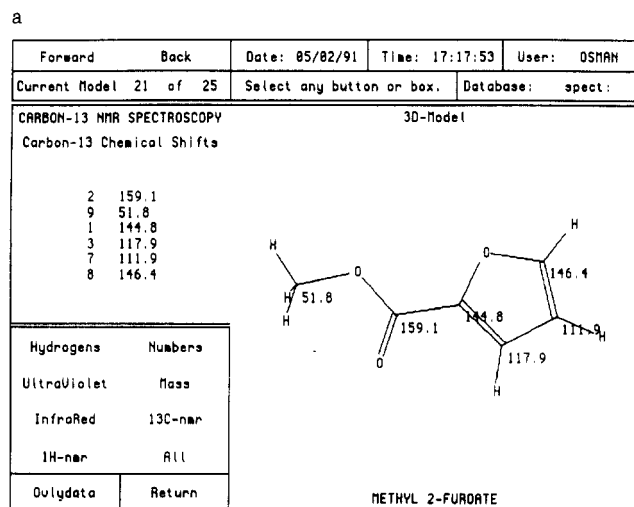
Forward	Back	Date: 85/02/91	Time: 17:13:28	User: OSMAN
Current Model 21 of 25		Select any button or box. Database: spect:		
VIBRATIONAL FREQUENCIES		Method: AM1 3D-Model		
Heat of Formation	Zero Point Energy	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
-74.982	72.556			
Hydrogens	Numbers	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
HOMO/LUMO	Thermodynamic	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
Bond Orders	Uib. Freq.	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
Charges	All	<div> <div>Occ. NO's</div> <div>Unocc. NO's</div> </div>		
Onlydata	Return	METHYL 2-FUROATE		

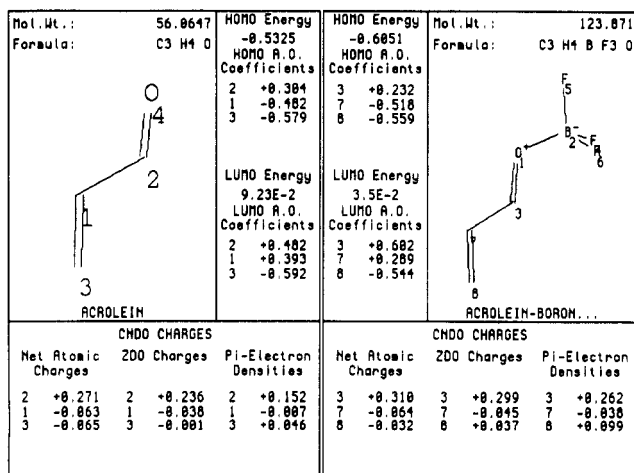
Figure 3. (a) Multiple pages of AM1 calculated information are displayed in one form. (b) Overlaid on the structure are the FMO coefficients, net atomic charges, and  $\pi$ -electron densities. (c) The structure with data overlaid is zoomed (the different data is more easily distinguished by color coding in the original application).

can search for carbons in a similar environment to get an idea of the typical chemical shifts for similar carbons.

A multidisciplinary database can be designed to serve individual needs as well as corporate needs. Hence, scientists from different disciplines can individually use such a database for their own research. They can extend the scope of such a database to accommodate their needs. For example, theoretical chemists can use MACCS-3D as an archival system



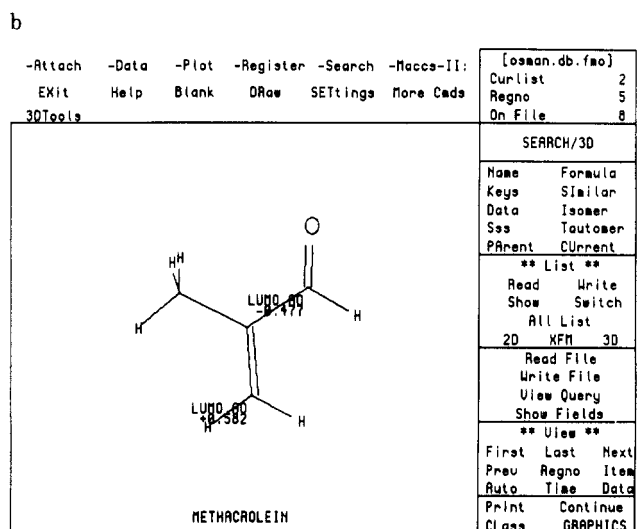
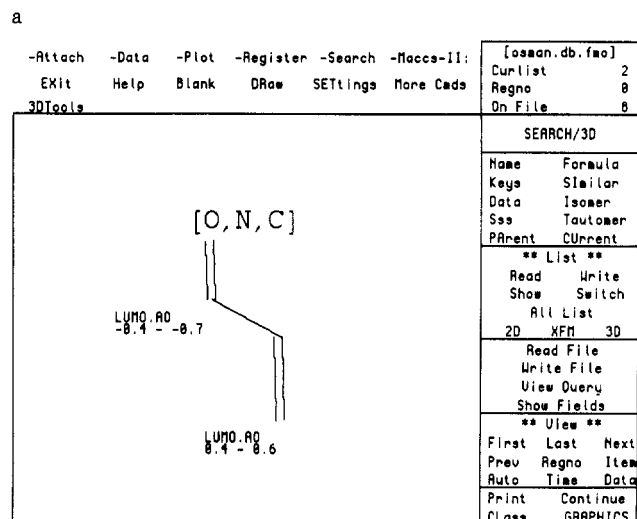
**Figure 4.** A sample of spectroscopic data is displayed: (a)  $^{13}\text{C}$  NMR chemical shifts and (b)  $^1\text{H}$  NMR chemical shifts, multiplicities, and coupling constants.



**Figure 5.** FMO energies and coefficients for acrolein and boron trifluoride complexed acrolein calculated by the CNDO/2 method.

to store the results of their quantum mechanical calculations. With easy access to earlier calculations, they do not have to unnecessarily repeat expensive calculations.

Figure 5 displays FMO energies and coefficients for acrolein and boron trifluoride complexed acrolein calculated by the CNDO/2 semiempirical method.<sup>19</sup> While a dramatic increase in the reaction rates upon catalysis accompanies loss of selectivity in many reactions, their loss does not occur for some



**Figure 6.** (a) MACCS-3D query to search for substituted olefins with a large LUMO coefficient at the terminal carbon as well as the substituent and (b) methacrolein, a hit satisfying the constraints set forth by the above query.

Diels-Alder reactions. Upon Lewis acid catalysis, both reaction rate and selectivity may increase.<sup>20</sup> This puzzling contradiction can be easily explained by examining the FMO energies and coefficients<sup>21</sup> of the two compounds (with and without Lewis acid catalysis). The form in Figure 5 allows the chemist to make this comparison visually.

Considerable lowering of the LUMO energy of acrolein from 0.092 eV to 0.035 eV (Figure 5) upon Lewis acid catalysis can explain the increasing rate. Because of the energy lowering, the LUMO of the dienophile will be energetically closer to the HOMO of the diene, allowing for easier transfer of electrons to form a bond and thus increasing the rate of the cycloaddition for 'normal-electron-demand'<sup>22</sup> Diels-Alder reactions. The higher regioselectivity for the ortho adduct can be explained by the increase in the ethylene LUMO coefficient differences from 0.199 to 0.255 (Figure 5). This increased polarity upon Lewis acid catalysis will favor the formation of the ortho adduct. Finally, the higher endo stereoselectivity can be explained by the sizable increase of the carbonyl carbon LUMO coefficient from 0.482 to 0.602 (Figure 5) upon Lewis acid catalysis. This increase causes a stronger secondary orbital interaction that will lower the activation energy for the endo transition state.<sup>23</sup>

Once a database is developed with data similar to those described above, scientists from entirely different disciplines can conduct specific substructure/data searches to benefit from

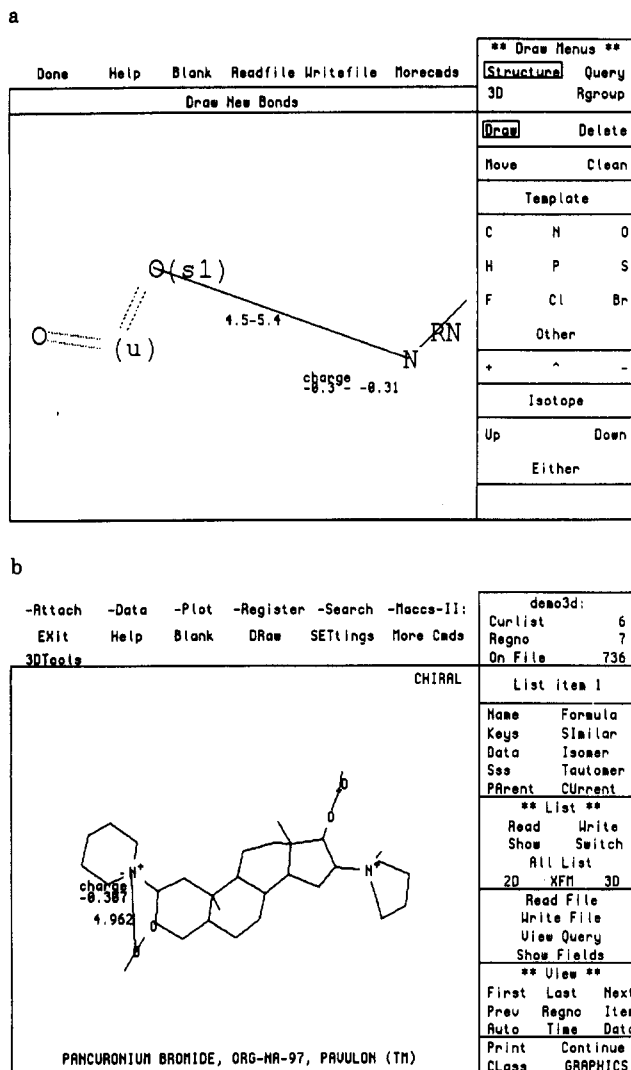


Figure 7. (a) MACCS-3D query to search for compounds with a carboxylate group 4.5–5.4 Å away from a ring nitrogen with a Gesteiger partial charge range of –0.3 to –0.31, and (b) pancuronium bromide, a hit with the above query.

each other's information in their own research. Consider this scenario: A synthetic chemist needs a substituted dienophile to use in a synthetic route that will selectively provide a Diels–Alder cycloadduct with, for example, ortho regiochemistry and endo stereochemistry. By using a similar approach to that described above, the scientist discovers that she needs a monosubstituted dienophile with a large LUMO coefficient at the unsubstituted carbon for the desired regioselectivity. Furthermore, the substituent must be conjugated to the alkene, also with a large LUMO coefficient, which will provide a favorable secondary orbital interaction for the desired endo stereoselectivity.<sup>23</sup> Figure 6a shows a search query with such constraints. Figure 6b shows a dienophile from the resulting list of retrieved molecules with the desired properties. Using the wrong dienophile may result in the cycloadduct with the wrong regio- or stereochemistry. If the particular Diels–Alder reaction was one of the later steps in the multistep process, such a late mistake would cause a waste of effort and resources. However, the scientist can prevent such waste in this hypothetical situation, because she can predict the outcome of the reaction. By using the information generated and made available by the computational chemists, the synthetic chemist can select reactants that will comply with the requirements to give the desired product.

More importantly, such data (computational or spectroscopic) can also be incorporated into a 3D search query as part

of a geometric search. Figure 7a shows an example for such a search query: it specifies a carboxylate group and a ring nitrogen that are a required distance away from each other. Furthermore, a range of partial charge is specified on the nitrogen. Figure 7b displays pancuronium bromide, one of the compounds in the hit list. Here, the query is overlaid with the hit to show exactly how the compound satisfies the constraints. This query, run on the Demo3D database of 740 compounds provided with the tutorial of MACCS-3D, generated a hit list of compounds of which 67% were identified as neuromuscular blocking agents.

## CONCLUSIONS

In this paper we demonstrated, with examples, how MACCS-3D's per-atom and atom-pair data-storing capabilities can be used to integrate data from computational and spectroscopic groups into a central 3D database of models and data.

We showed how researchers of various disciplines can each use a single 3D database differently and how scientists from one discipline can easily access information provided by scientists from other disciplines, thereby allowing a truly multidisciplinary approach to research.

Using Brown's terminology for data (quantitative measurement), information (evaluated data), and knowledge (integrated information),<sup>24</sup> the goal of such an integrated information management system is to help all scientists to initially access needed *data*, evaluate it, and to finally transform *information* generated into *knowledge*. Achieving this goal is essential in today's multidisciplinary research environment.

## ACKNOWLEDGMENT

Thomas E. Moock, Bradley D. Christie, and Douglas R. Henry, who led the MACCS-3D development team, are gratefully acknowledged for their assistance and cooperation throughout this and many other application projects.

## REFERENCES AND NOTES

- (1) Presented in part at the 199th National Meeting of the American Chemical Society, Boston, MA, April 1990; paper CINF 7.
- (2) Bowman, C. M.; Nosal, J. A.; Rogers, A. E. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 147–151.
- (3) Baker, D. B. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 55–59.
- (4) Brown, H. D. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 78–80.
- (5) Seltzer, R. *J. Chem. Eng. News* **1983**, Aug 8, 16.
- (6) Layman, P. L. *Chem. Eng. News* **1989**, July 10, 9–11.
- (7) Jackson, J. *The Scientist* **1990**, June 25, 5.
- (8) (a) Howe, W. J.; Hagadone, T. R. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 8. (b) Barcza, S.; Kelly, L. A.; Wahrman, S. S.; Kirschenbaum, R. E. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 55. (c) Eakin, D. R.; Hyde, E.; Palmer, G. *Pestic. Sci.* **1973**, 319. (d) Brown, H. D.; Costlow, M.; Cutler, F. A., Jr.; Demott, A. N.; Gall, W. B.; Jacobus, D. P.; Miller, C. J. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 5. (e) Milne, G. W. A.; Miller, J. A. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 154. (f) Townsley, E. E.; Warr, W. A. In *Retrieval of Medicinal Chemical Information*; Howe, W. J., Milne, M. M., Pennell, A. F., Eds.; ACS Symposium Series 84; American Chemical Society: Washington, DC, 1978; Chapter 6. (g) Westland, R. D.; Holcomb, R. L.; Vinson, J. W.; Steele, J. D.; Cardwell, R. J.; Scott, R. L.; Harkaway, T. D.; Hyttinen, P. J.; Williams, T. *Ibid.* Chapter 9. (h) Dyott, T. M.; Edling, A. M.; Garton, C. R.; Johnson, W. O.; McNulty, P. J.; Zander, G. S. *Ibid.* Chapter 11. (i) Page, J. A.; Thiesen, R.; Kuhl, F. *Ibid.* Chapter 12. (j) Heller, S. R.; Milne, G. W. A. *Ibid.* Chapter 10.
- (9) Hagadone, T. R.; Lajiness, M. S. *Tetrahedron Comput. Methodol.* **1988**, *1*, 219–230.
- (10) Barcza, S.; Mah, H. W.; Myers, M. H.; Wahrman, S. S. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 198–204.
- (11) MACCS-II: Molecular Access System, available from Molecular Design Limited, 2132 Farallon Drive, San Leandro, CA 94577.
- (12) (a) Christie, B. D.; Henry, D. R.; Güner, O. F.; Moock, T. E. In *Online Information 90*; Raitt, D. I., Ed.; 14th International Online Information Proceedings; Learned Information, Oxford, 1990; pp 137–161. (b) Moock, T. E.; Christie, B. D.; Henry, D. R. In *Chemical Information Systems*; Bawden, D., Mitchell, E. M., Eds.; Ellis Horwood: Chichester, 1990; p 42.
- (13) Bawden, D. In *Chemical Information Systems*; Bawden, D., Mitchell, E. M., Eds.; Ellis Horwood: Chichester, 1990; p 163.

- (14) Silverstein, R. M.; Bassler, G. C.; Morrill, T. C. *Spectrometric Identification of Organic Compounds*, 4th ed.; John Wiley & Sons: New York, 1981.
- (15) AMPAC: QCPE No. 506, Indiana University, Chemistry Department.
- (16) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
- (17) Program runs on the entire range of VAX computers (VMS 5.1 or higher). The minimum hardware required to run MACCS-3D (rev. 1.0), which includes MACCS-II and the Customization Module, is a MicroVax-2000 with 8-Mbyte memory (16 Mbyte recommended). The program code takes less than 25 000 blocks of disk space; depending on the size of the database(s), up to 300 000 blocks of disk space is recommended initially (FCD-3D with 65 000 compounds takes 180 000 blocks). The price of the software depends on the size of the CPU and the number of users.
- (18) Customization Module: a module of MACCS-II that enables users to write applications and front-ends to the software, available from Molecular Design Limited.
- (19) Pople, J. A.; Segal, J. A. CNDO/2. *J. Chem. Phys.* **1966**, *44*, 3289.
- (20) Jain, P. C.; Mukerjee, Y. N.; Anand, N. *J. Am. Chem. Soc.* **1974**, *96*, 2996.
- (21) For an excellent review of the frontier molecular orbital concept see: Woodward, R. B.; Hoffmann, R. *Conservation of Orbital Symmetry*; Verlag Chemie GmbH: Weinheim, 1970; and Fleming, I. *Frontier Orbitals and Organic Chemical Reactions*; Wiley, New York: 1976. Results of ab initio calculations in concert with the observation in text is published: Güner, O. F.; Ottenbrite, R. M.; Shillady, D. D.; Alston, P. V. *J. Org. Chem.* **1987**, *52*, 391.
- (22) Sustmann, R. *Tetrahedron Lett.* **1971**, 2717 and 2721.
- (23) Güner, O. F.; Ottenbrite, R. M.; Shillady, D. D.; Alston, P. V. *J. Org. Chem.* **1988**, *53*, 5348.
- (24) Brown, H. D. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 155-158.

## A New Method of Computer Representation of Stereochemistry. Transforming a Stereochemical Structure into a Graph

TATSUYA AKUTSU

Mechanical Engineering Laboratory, 1-2 Namiki, Tsukuba, Ibaraki, Japan 305

Received February 2, 1991

A new method of computer representation of stereochemical structures, which include double bonds and asymmetric carbon atoms, is described. The method is very simple, and a stereochemical structure is transformed into a graph. From the results, graph algorithms, which have been intensively studied in computer science, can be directly applied to chemical structures. Especially, a polynomial time algorithm for stereochemically unique naming is implied, for which SEMA (stereochemically extended Morgan algorithm) does not work in polynomial time.

### INTRODUCTION

Representation and manipulation of chemical structures are very important for database systems and expert systems in chemistry. Especially, unique naming<sup>3,6,7,9,11</sup> and substructure matching<sup>12</sup> are most important. Usually, a chemical structure is represented as a graph, in which an atom corresponds to a vertex and a chemical bond corresponds to an edge. However, a graph is not sufficient for representing a chemical structure. Stereoisomers must be distinguished. Though they have the same graph structures, their geometric structures are different and they show different properties. How to represent stereoisomers in computer systems has been studied well. Especially, the works of Wipke and Dyott are well known. They developed the stereochemically unique naming algorithm<sup>13,14</sup> (stereochemically extended Morgan algorithm, abbreviated as SEMA) based on the ordered list representation of stereochemical structures by Petrarca et al.<sup>10</sup>

In this paper, another approach to distinguish stereoisomers is presented. A basic technique in computer science "transformation" is employed. A chemical structure is transformed into a usual graph (a structure which does not have stereochemical information). Besides, two structures are transformed into isomorphic graphs, if and only if they are stereochemically isomorphic.

### ORDERED LIST METHOD

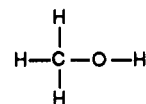
Before describing the transformation method, the ordered list method is reviewed. It was developed by Petrarca et al.<sup>10</sup> and adapted in the SEMA algorithm by Wipke and Dyott.<sup>13,14</sup> However, the original method is not described, but the method proposed in the CHAUS system<sup>2</sup> is described since it is simpler.

In this paper, only stereoisomers caused by the following local structures are considered (see Figure 1).

- (1) A pair of carbon atoms connected with a double bond
- (2) An asymmetric carbon atom adjacent to four atoms

Other cases (such as conformation) are considered to be handled in a similar way.

Basically, chemical structures are represented as graphs in the ordered list method. In graph representation, an atom corresponds to a vertex and a chemical bond corresponds to an edge. Note that, in this paper, hydrogen atoms are not graphically abbreviated. For example, methanol is represented as



but is not represented as CH<sub>3</sub>-OH.

The adjacency list, which is a famous data structure in computer science, is employed to represent a graph (see Figure 2). The adjacency list is essentially the same as the connection table, which is popular in chemical information processing. In the adjacency list, there is a list of attached atoms for each atom. The ordering of atoms in the list has no meaning, and an arbitrary ordering is allowed. However, the ordering is used to represent stereochemical information in the ordered list method.

At first, consider the case of a pair of carbon atoms connected with a double bond. For each carbon atom, adjacent atoms are listed in a clockwise order beginning with the other carbon atom (see Figure 3). Although four (= 2 × 2!) different orderings can be considered, two equivalent orderings are allowed (see Figure 3). The other two equivalent orderings are generated for the stereoisomer. Due to the ambiguity of the orderings, stereochemically isomorphic structures are not