Chemical Information in 3D Space

Johann Gasteiger,*,† Jens Sadowski, Jan Schuur, Paul Selzer, Larissa Steinhauer, and Valentin Steinhauer

Computer-Chemie-Centrum, Institut für Organische Chemie, Universität Erlangen-Nürnberg,
Nägelsbachstrasse 25, D-91052 Erlangen, Germany

Received March 22, 1996⊗

Progress in recent decades in the representation of chemical structures is briefly outlined. A reasonable model of the three-dimensional structure of any organic compound can nowadays be generated. This opens the door to the study of the relationships between the 3D structure of organic molecules and physical, chemical, or biological properties. A code has been developed that allows the representation of the 3D structure of molecules by a fixed (constant) number of variables and thus is amenable to an analysis by statistical and pattern recognition methods or neural networks. Examples for the investigation of biological data and the simulation of infrared spectra are given. Further work on the relationships between structure and infrared spectra has shown that the 3D structure of an organic molecule can be derived from its infrared spectrum.

INTRODUCTION

The way molecular structures are represented profoundly influences the level of insight that can be gained from chemical information. In the 1960s and early 1970s chemical structures were largely represented by fragment codes and line notations such as the Wiswesser Line Notation. This allowed a highly concise coding of chemical structures. However, fragment codes are notoriously insufficient for the representation of chemical reactions where individual bonds are broken and made that might be buried within a fragment.

The 1970s saw the rise of coding molecular structures by connection tables, a method that has now gained universal acceptance. An important problem that has to be dealt with when working with connection tables is to find a unique and unambiguous coding of a chemical structure. This is usually achieved by canonical numbering the atoms of a molecule. Clemens Jochum, one of the recipients of the 1995 Herman Skolnik Award, has early on solved this problem. The paper[1] describing this work was met with controversy,[2] but eventually all misunderstandings could be clarified (see Figure 1).[3−5] Clearly, however, molecules are three-dimensional objects, and any in-depth analysis of chemical information, particularly the analysis of relationships between structure and physical, chemical, or biological properties has to account for the three-dimensional arrangement of the atoms in a molecule. With the advent of automatic 3D structure generators[6] the door has been opened for the representation of the 3D structure of molecules, a method that will gain a lot of momentum in the near future.

The three-dimensional arrangement of the atoms in molecules had caught the interest of Reiner Luckenbach, the other recipient of the 1995 Herman Skolnik Award, many years ago as underlined by his authorship of a book on the dynamic stereochemistry of pentacoordinated phosphorous compounds (see Figure 2).[7]

FROM CONNECTION TABLES TO 3D STRUCTURES

It should be obvious that the coding of chemical structures has to correspond to our knowledge of chemical bonding. Clear as this requirement might seem, it has not always been fulfilled. The introduction of connections tables acknowledged the fact that molecules consist of atoms and bonds. In fact, a connection table is basically a valence bond representation of a molecule. This already shows that connection tables are quite inappropriate for the representation of structures that should better be described by a molecular orbital representation. The separation of σ- and π-electrons in the representation of molecules is a first step in this direction[8] and has also been adopted in the Beilstein Structure Registry System.[9] A more detailed representation has to take account of the orthogonality of π-orbitals.[10]

Deeper insight into the structure of chemical species has to consider the three-dimensional arrangement of the atoms in molecules. For more than 100 000 organic compounds the 3D structure has been determined by X-ray structure analysis and stored in the Cambridge Crystallographic Database.[11] Impressive as this number might seem, it is dwarfed by the number of known chemical compounds which already exceeds 14 million. This lack of data on the 3D structure of compounds has prevented attempts at coding molecules in their three-dimensional form.

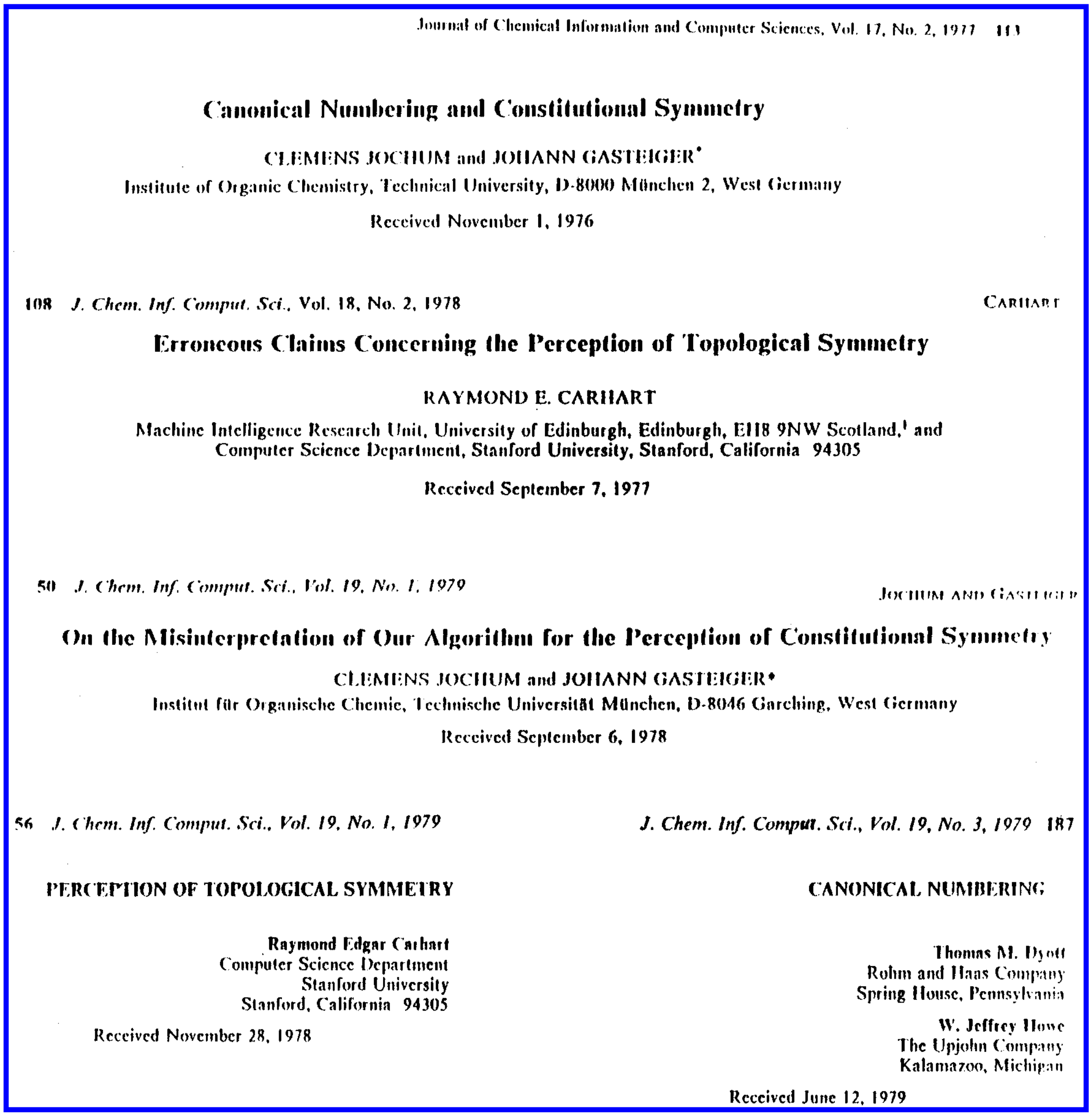
Recently, however, programs have become available that build a 3D model of a molecule from the information contained in a connection table (and stereochemical descriptors where appropriate).[6] Some of these systems are applicable to the entire range of organic chemistry and even cover certain classes of organometallic compounds.[12] Efforts have been made to make this 3D structure generation process automatic to allow the conversion of large datafiles without human intervention.

To illustrate the broad range of validity of such 3D structure generators, results from a fairly large datafile are given. The National Cancer Institute (NCI) has recently made available connection table information on nonproprietary compounds submitted to NCI. This datafile of 126 705 connection tables was submitted to the 3D structure generator CORINA (version 1.6) developed in our laboratory.[12−15] Most of the connection tables (126 148, 99.56%) could be converted into 3D structures.[15] Some of the structures (327, 0.28%) contained atoms with a coordination number higher

† gasteiger@eros.ccc.uni-erlangen.de.

⊗ Abstract published in *Advance ACS Abstracts,* June 15, 1996.

S0095-2338(96)00343-5 CCC: $12.00 © 1996 American Chemical Society



Journal of Chemical Information and Computer Sciences, Vol. 17, No. 2, 1977 113

Canonical Numbering and Constitutional Symmetry

CLEMENS JOCHUM and JOHANN GASTEIGER*
Institute of Organic Chemistry, Technical University, D-8000 München 2, West Germany

Received November 1, 1976

108 J. Chem. Inf. Comput. Sci., Vol. 18, No. 2, 1978 CARHART

Erroneous Claims Concerning the Perception of Topological Symmetry

RAYMOND E. CARHART
Machine Intelligence Research Unit, University of Edinburgh, Edinburgh, EH8 9NW Scotland,[1] and
Computer Science Department, Stanford University, Stanford, California 94305

Received September 7, 1977

50 J. Chem. Inf. Comput. Sci., Vol. 19, No. 1, 1979 JOCHUM AND GASTEIGER

On the Misinterpretation of Our Algorithm for the Perception of Constitutional Symmetry

CLEMENS JOCHUM and JOHANN GASTEIGER*
Institut für Organische Chemie, Technische Universität München, D-8046 Garching, West Germany

Received September 6, 1978

56 J. Chem. Inf. Comput. Sci., Vol. 19, No. 1, 1979

PERCEPTION OF TOPOLOGICAL SYMMETRY

Raymond Edgar Carhart
Computer Science Department
Stanford University
Stanford, California 94305

Received November 28, 1978

J. Chem. Inf. Comput. Sci., Vol. 19, No. 3, 1979 187

CANONICAL NUMBERING

Thomas M. Dyott
Rohm and Haas Company
Spring House, Pennsylvania

W. Jeffrey Howe
The Upjohn Company
Kalamazoo, Michigan

Received June 12, 1979

Figure 1. Scientific dispute on early work of Clemens Jochum on canonical numbering of connection tables.

than six and were thus outside the scope of CORINA; only 230 (0.18%) of the connection tables could not be converted into 3D structures because of deficiencies in the program. The entire process for the conversion of these more than 126 000 structures took 34 856 s on a Silicon Graphics Indigo 2, under IRIX 5.3 and did not need any human intervention and thus could be completed in one run. On average, the generation of a 3D structure took 0.28 s on this average size workstation.[16]

The entire datafile of 126 148 3D structures has been made publicly available and has been deposited at the file server of NCI. It can be accessed by anonymous file transfer (ftp):

ftp://ftp.helix.nih.gov/ncidata/3D/nciopen3d.mol.Z

(The file has been compressed with the UNIX command "compress".)

ftp://enar.organik.uni-erlangen.de/pub/nci/
nciopen3d.mol.Z

Access to CORINA is also provided over the world wide web (WWW). A maximum of 1000 connection tables can be converted by each individual or research group free on the following WWW address:

http://schiele.organik.uni-erlangen.de/services/3d.html

Alternatively, a structure coded as a connection table in MDL Molfile format can be sent to CORINA@eros.ccc.uni-erlangen.de with a subject line reading: PROPS =A_XYZ, ReturnFormat=xxx.

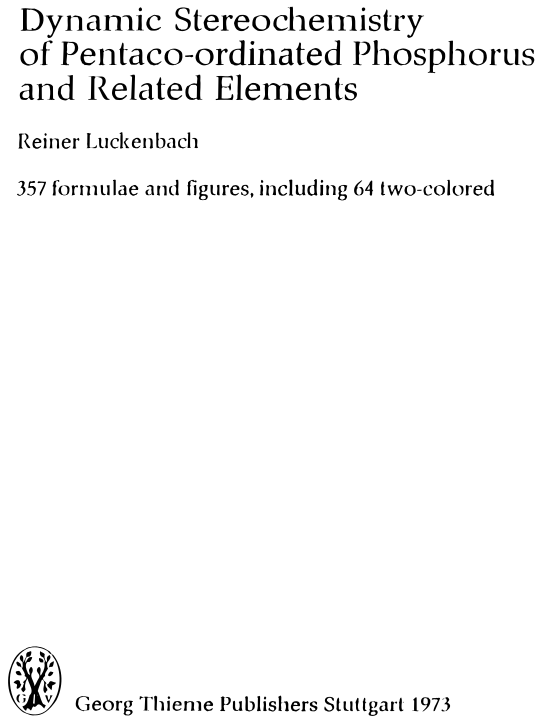
Three crucial questions have to be answered when dealing with automatic 3D structure generators:

(1) How broad is the scope of a 3D structure generator?

(2) How good are the 3D models generated?

(3) How rapid are 3D structures obtained?

The first question is rather easy to answer. One has to choose a dataset of compounds and then determine the conversion rate. This has been done for the six 3D structure generators ALCOGEN,[17] CHEM-X,[18] COBRA,[19] CONCORD,[20] CORINA,[12−15] and MOLGEO[21] with a carefully selected dataset of 639 molecules, comprising a host of different classes of compounds for which also X-ray structure data were available.[12] Table 1 gives the conversion rate for these six structure generators. Recently, this investigation has been extended to include the version 1.6 of CORINA and the program CONVERTER.[22] The results with these two systems are also included in Table 1.



Dynamic Stereochemistry
of Pentaco-ordinated Phosphorus
and Related Elements

Reiner Luckenbach

357 formulae and figures, including 64 two-colored

Georg Thieme Publishers Stuttgart 1973

Figure 2. Monograph title page showing Reiner Luckenbach's interest in stereochemistry. Reprinted with permission. Copyright 1973 Georg Thieme Publishers Stuttgart.

The second question, addressing the problem of the quality of the 3D models, is much more difficult to approach. Comparisons of the generated 3D coordinates with those experimentally determined give a clear picture only for rather rigid structures such as those containing rigid ring systems. The root mean squares (RMS) deviations in the coordinates can be taken as a quantitative measure.[12] Figure 3 shows a visual comparison of structures generated by CORINA with those taken from the Cambridge Crystallographic Datafile.[11] As can be seen, CORINA can also be applied to the generation of 3D models of organometallic compounds.

Most molecules, however, are rather flexible, being able to exist in a variety of conformations having nearly equal energy. For those molecules, a direct comparison of the conformation generated by the 3D structure generator and that conformation existing in the crystal is not sufficient. Different quality criteria and a multistep evaluation procedure has been invoked to address the question on the quality of 3D models obtained from the seven 3D structure generators.[12] The results are given in detail in ref 12. As a further extension Figure 4 shows a plot of the number of structures converted against the root mean squares (RMS) deviation of the calculated 3D atomic coordinates determined by X-ray structure analysis.[23] The plot also includes results obtained with CONVERTER that has recently become available.[22]

The third question, the conversion speed, how much time is needed on average for calculating 3D atomic coordinates, becomes crucial when large datasets consisting of several hundreds of thousands or even millions of connection tables have to be converted into 3D structures. Table 1 also gives the results on the CPU time needed by the seven 3D structure generators mentioned above on the dataset of 639 molecules. It has to be mentioned again that some 3D structure generators could not convert all structures. The computation times refer only to those molecules that could be converted.

Table 1. Comparison of the Performance of Seven 3D Structure Generators on a Dataset of 639 Structures

	conversion rate [%]	CPU time [s/molec] (on a VAX 6000)
CONCORD	84	0.14
ALCOGEN	79	0.79
CHEM-X	74	0.33
MOLGEO	79	8.98
COBRA	75	3.49
CORINA 1.5	100	0.58
CORINA 1.6	100	0.32
CONVERTER	98	3.64

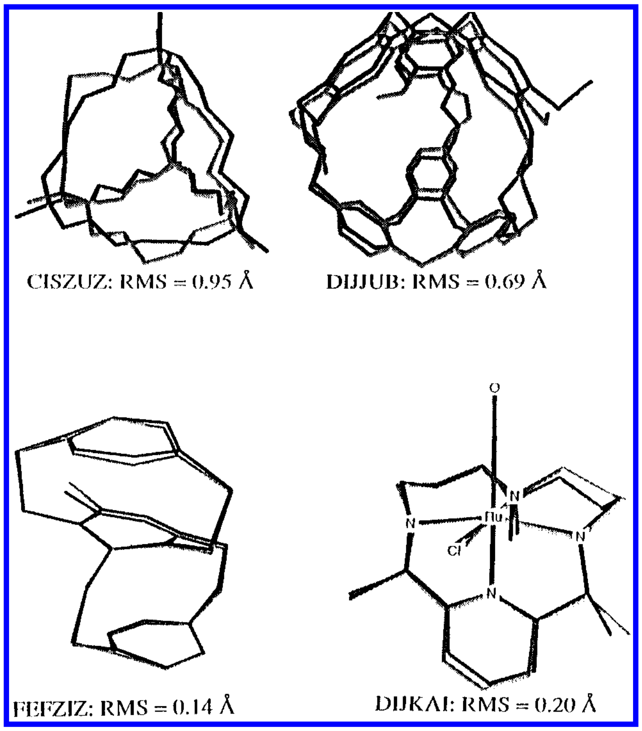


Figure 3. Examples of 3D structures generated by CORINA and their comparison to X-ray structures. The codes give the references to the Cambridge Crystallographic Datafile; the RMS value shows the root mean squares differences between the experimental and the calculated atom coordinates.

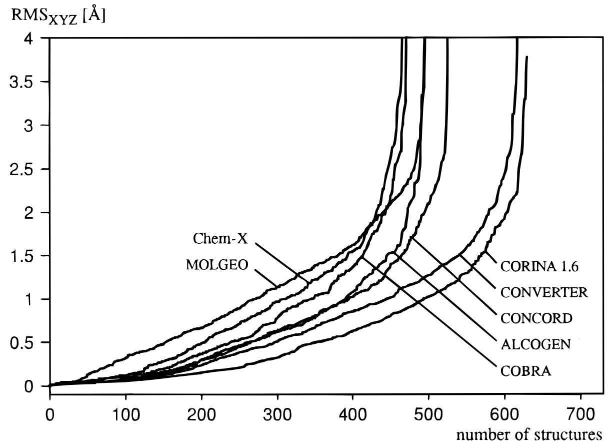


Figure 4. Quantity-quality characteristics of seven 3D structure generators. Conversion rate vs RMS value of non-hydrogen atoms.

CODING THE 3D STRUCTURE

Having secured a general and rapid access to the 3D structure of practically any organic molecule, the question is now how to code the 3D structure? An obvious approach would be to give the cartesian or internal coordinates of the individual atoms of the molecule. As each atom asks for three coordinates, the size of the code for a molecule then directly depends on the number of atoms. Methane with five atoms will be represented by 15 variables, cholesterol with 74 atoms by 222 variables. However, many methods used for finding relationships between structure and properties, particularly those used in quantitative structure activity relationship (QSAR) studies, such as statistical or pattern recognition methods or neural networks, require each molecule of a dataset to be represented by the same number of variables. The solution of this problem, the representation of the 3D structure of a molecule by a constant (fixed) number of variables irrespective of the number of atoms in the molecule, has been achieved by a novel code, the 3D MoRSE (*Mo*lecule *R*epresentation of *S*tructures based on *E*lectron diffraction) code.

The foundations of the 3D MoRSE code are now briefly outlined. More details are to be found in a recent publication.[24] The first step to a solution of the problem was made by considerations on the experimental methods used for 3D structure determination. One of the methods employed is electron diffraction. The general molecular transform used in electron diffraction studies is given by eq 1

$$G(\vec{S}) = \sum_{i=1}^{N} f_i(2\pi \vec{r}_i \vec{S}) \tag{1}$$

with $G(\vec{S})$, diffraction pattern; $\vec{S}$, observation point; $\vec{r}_i$, location of atoms; f_i, form factor of atom i; and N, number of atoms in molecule.

This relationship is usually handled in a form proposed by Wierl[25] and represented in eq 2

$$I(s) = K \sum_{i=2}^{\mathrm{N}} \sum_{j=1}^{i-1} f_i f_j \int_0^\infty P_{ij}(r) \frac{\sin(sr)}{sr} dr \tag{2}$$

with

$$\mathrm{s} = 4\pi \sin(\vartheta/2)/\lambda \tag{3}$$

and ϑ, scattering angle; λ, wavelength; $I(s)$, intensity of scattered radiation; $P_{ij}(r)$, probability distribution of the variation in the distance between atoms i and j due to vibrations; f_i, form factor of atom i; and K, collection of instrument constants.

This equation was further modified to give eq 4. In doing this we partly followed suggestions made by Soltzberg and Wilkins[26]

$$I(s) = \sum_{i=2}^{\mathrm{N}} \sum_{j=1}^{i-1} A_i A_j \frac{\sin(sr_{ij})}{sr_{ij}} \tag{4}$$

A_i, can be any atomic property such as atomic number, mass, partial atomic charge, or atomic polarizability and s, is a reciprocal distance.

The possibility for choosing an appropriate atomic property gives great flexibility to the 3D MoRSE code for adapting it to the problem under investigation. Quite a few different atomic properties have been investigated from the suite of physicochemical and topological values that are obtained by the program package PETRA (*P*arameter *E*stimation for the *T*reatment of *R*eactivity *A*pplications).[27] The value of s was considered only at discrete positions within a certain range. In many applications we have chosen 32 equidistant values between 0 and 31 Å^{-1}. The choice of the range of s and the number of values to be considered determines the resolution of the code for representing the 3D structure.

APPLICATIONS OF THE 3D MORSE CODE

The 3D MoRSE code with its fixed-length representation of 3D molecular structure allows the comparison of datasets comprising molecules of different size, with different number of atoms. Thus, the 3D MoRSE code will find many applications in establishing relationships between molecular structure and physical, chemical, or biological properties. One of the most promising fields of application lies in the area of quantitative structure−activity relationships (QSAR).

BIOLOGICAL ACTIVITY

Biological activity is intimately tied to the three-dimensional structure of molecules and to electronic properties of specific sites of the molecule. The potential of the 3D MoRSE code of simultaneously considering the 3D structure and atomic properties such as partial atomic charges (see eq 4) makes it particularly suited for studying biological data. It has been shown that the 3D MoRSE code is able to distinguish molecules that bind to the dopamine D1 receptor from those that bind to the dopamine D2 receptor.[24]

Furthermore, steroids that bind to the corticosteroid binding globulin (CBG) receptor can be clustered into compounds of high, intermediate, and low activity.[24] This is shown with a dataset of 31 steroids also, investigated by the widely used CoMFA method.[28] This dataset has been investigated by a variety of methods and is widely distributed. Unfortunately, some of the structures in the publications and in the electronic datafiles contain coding errors. This has been rectified in a recent publication.[29] The structures can also be obtained in MDL MOLfile format by anonymous ftp both as connection tables and with 3D coordinates generated by CORINA.[30] Each steroid of the dataset was converted into 3D coordinates by CORINA and represented by 32 3D MoRSE code values calculated by eq 4 using as atomic property A_i either the mass, m_i, of each atom, or the partial atomic charge, $q_{tot,i}$, as calculated by the PEOE method.[31,32]

The relationship between the 3D structure of the steroids and their CBG binding affinity was established by a Kohonen neural network.[33−35] The dataset was split into a set of 25 steroids for training the neural network and six molecules for testing its prediction ability, two compounds each of low, intermediate, and high affinity. The six molecules of the test set are given in Figure 5.

Each of the structures of the training set was input into a Kohonen neural network consisting of 10 × 10 neurons. The architecture of the network corresponds to the upper part of Figure 7. Each neuron is represented by a column of the upper block of Figure 7 and contains as many weights as there are input variables; in our case with 32 values of $I(s)$ (eq 4), each neuron has 32 weights. The learning algorithm of a Kohonen network is such that each object, in our case, each steroid, is projected into a specific neuron of the two-dimensional arrangement of neurons.[34,35] The mapping of the training dataset of 25 steroids into the Kohonen network is shown in Figure 6. The view is now from the top onto the 10 × 10 network. Each neuron is represented by the intersection of lines and is marked by a number if a steroid

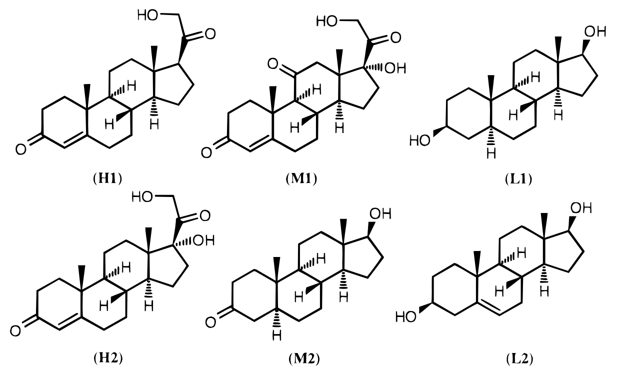


Figure 5. The six steroids of the test set, with low (L1, L2), medium (M1, M2), and high (H1, H2) affinity for binding to the corticosteroid binding globulin (CBG) receptor.

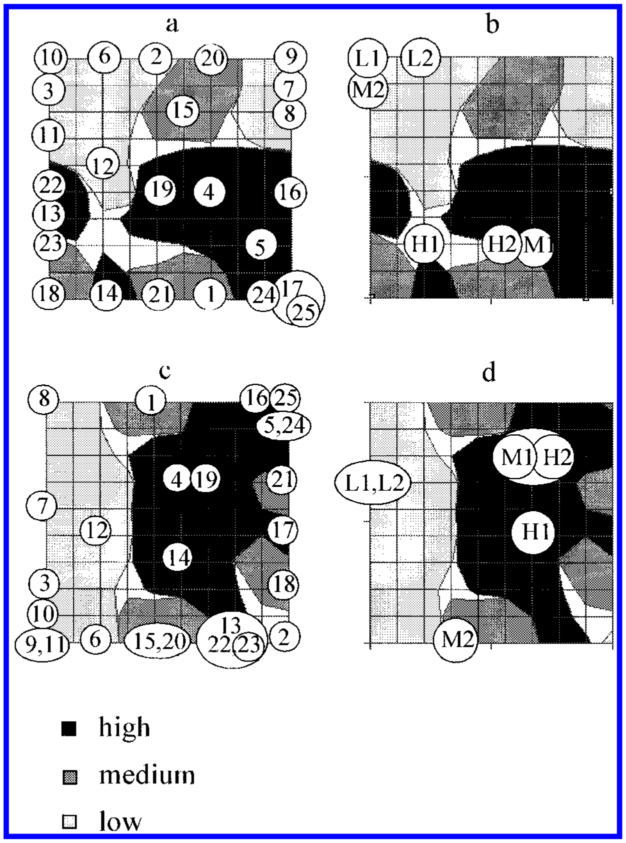


Figure 6. Mapping of a dataset of 31 steroids showing low, intermediate, and high affinity for binding to the corticosteroid binding globulin (CBG) receptor; (a) and (b) using atomic mass, m_i, as atomic property, A_i, in eq 4 for the calculation of the 3D MoRSE code (a) training set; (b) test set); (c) and (d) using partial atomic charge, $q_{tot,i}$, as atomic descriptor, A_i, (c) training set; (d) test set).

is projected into the network; the number representing the corresponding steroid of the dataset as numbered in ref 29.

Figure 6, part a (training set) and part b (test set), shows the results obtained with choosing mass, m_i, as atomic property, A_i, in eq 4. A certain amount of separation of compounds of low, intermediate, and high affinity to the CBG receptor can be distinguished for the training set (Figure 6a), although each activity class is represented by several clusters in the map of the Kohonen network. The poor separation of compounds according to CBG affinity is emphasized with the test set (Figure 6b): One of the steroids of intermediate activity (M2) is misclassified as having low activity. The other steroid with intermediate affinity (M1) and both compounds of high affinity (H1 and H2) are mapped into the border area between intermediate and high affinity and thus difficult to classify.

When partial atomic charges, $q_{tot,i}$, as calculated by the PEOE method[31,32] are used as atomic property, A_i, in eq 4, the picture becomes much clearer. Figure 6c shows that training the Kohonen network now leads to distinguished areas of compounds of low and high CBG affinity, with compounds of intermediate affinity being in areas of transition around the cluster of high affinity compounds. In addition, the prediction capability is much better as shown in Figure 6d. Compounds L1, L2, M2, H1, and H2 are all classified into their correct affinity range; only M1 is misclassified as it is mapped into an area of high affinity, although M1 is not far away from the cluster of compounds of intermediate affinity.

SIMULATION OF INFRARED SPECTRA

As an illustration of the power of the 3D MoRSE code for the prediction of physical data, we present here results on the simulation of infrared spectra. Infrared spectra can be calculated by quantum mechanical methods. However, in order to obtain spectra that agree with experimental data within an acceptable accuracy, quite extensive *ab initio* calculations with large basis sets or by a density functional theory are needed. On the other hand, many approaches to establish empirical relationships between structure and infrared spectra have been made. All those empirical approaches suffer from their representation of the structure of the molecules which is usually based on fragments derived from the constitution of a molecule. Whereas fragments are quite acceptable for the correlation of the vibrations of bonds, they are notoriously incapable of reproducing more complex vibrations like those observed in the fingerprint region between 1000 and 1500 cm^{-1}. Real progress in this field can only be expected when the three-dimensional structure of a molecule is considered because an infrared spectrometer monitors the vibrations of a molecule in 3D space.

In the following, the relationships between structure and infrared spectra were stored in a counterpropagation neural network.[34−36] Therefore, a brief outline of a counterpropagation (CPG) network is necessary. For a more extensive introduction see refs 33−36. A counterpropagation network can be used to establish relationships between two sets of data, input and output data, such as the structure (input) and the infrared spectrum (output) of a compounds.

The input data are stored in a two-dimensional network which is basically a Kohonen neural network.[33−35] In Figure 7, this part of the CPG network is represented by the upper block. Each neuron in the Kohonen (upper) block has as many weights as there are input data for each object. In our case, the structure of the molecules was represented by the 3D MoRSE code, using partial atomic charges, $q_{tot,i}$, as atomic property, A_i, and choosing 32 discrete values of s in the range of 0−31 $Å^{-1}$ to calculate *I(s)* by eq 4. The lower part of a counterpropagation network, used for storing the output data, in our case the infrared spectrum, is basically a look-up table. The infrared spectra were taken from the SpecInfo database[37] and represented by 128 absorbance values in the range of 3500 to 560 cm^{-1} (cf. refs 24 and 38 for further details on the IR spectra coding). A neuron in a CPG network consist of a column of weights in the Kohonen

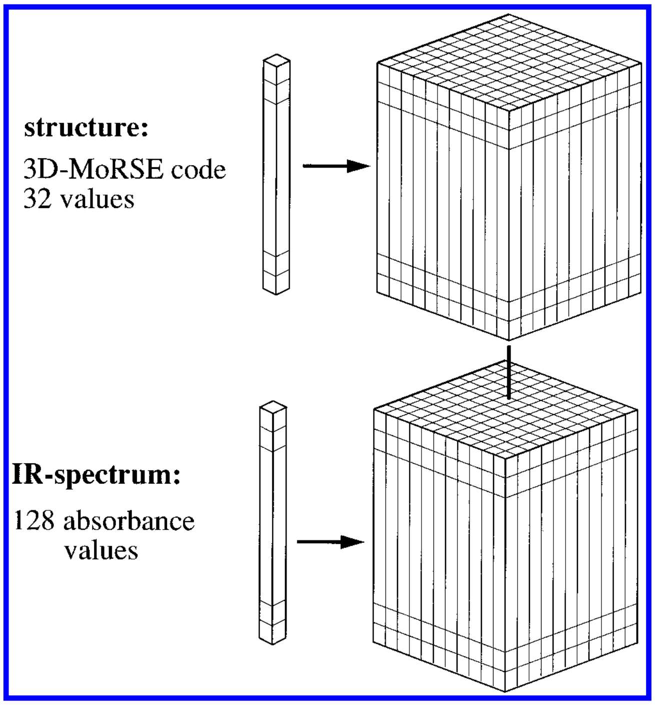


Figure 7. Architecture of a counterpropagation (CPG) network for the simulation of infrared spectra. The upper block is a two-dimensional arrangement of neurons, represented by columns of weights. This block is basically a Kohonen network and is used for storing the structure of molecules as represented by 32 values of the 3D MoRSE code. The lower block is used for storing the infrared spectrum; it is used as a look-up table. The neurons of the upper block extended into the lower block having as many additional weights as there are input data for each infrared spectrum.

block and a column of weights in the output block having the same location in the two-dimensional arrangement of neurons.

Training of a CPG network consists of presenting pairs of 3D structures and associated infrared spectra to adjust the weights in the two blocks of the CPG network by a competitive learning process.[34-36] After a number of such structure-spectra pairs have repeatedly been presented to a CPG-network the weights in the input and output block have stabilized.

The simulation of an infrared spectrum then asks for the input of the 3D MoRSE code as structure representation. This code will find a neuron having weights most similar to the input data. This neuron basically is the address where the infrared spectrum is to be found in the lower block (see Figure 7), the look-up table.

A study on the simulation of monosubstituted benzene derivatives has recently been published.[24] This investigation was further extended to include di- and trisubstituted benzene derivatives comprising all possible substitution patterns. All those benzene derivatives were taken from the SpecInfo database[37] that had substituents with up to six non-hydrogen atoms including C, N, O, F, Cl, and Br atoms. This provided 871 compounds. There were 487 molecules of this dataset taken to train a counterpropagation (CPG) network consisting of 25 × 25 neurons. Training needed 100 epochs. The rest of the dataset, consisting of 384 structures was used for testing the prediction performance of this network. All the examples contained in the following Figures 8−10 refer to molecules from the test set, molecules the CPG network had not seen before. The correlation coefficient between experimental and simulated infrared spectrum, calculated according to Bravais−Pearson, was taken as a measure of the quality of the prediction. A large percentage (89%) of the spectra had a correlation coefficient higher than 0.50, 65% higher than 0.70.

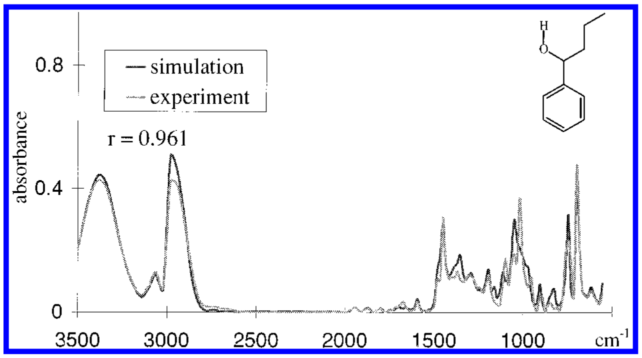


Figure 8. Experimental and simulated IR spectrum of 1-phenylbutanol-1 obtained from a 25 × 25 CPG network trained with 487 polysubstituted benzene derivatives and their associated IR spectra.

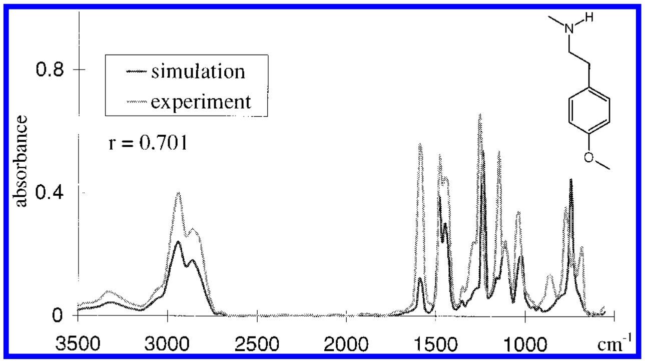


Figure 9. Experimental and simulated IR spectrum of 2-(4-methoxyphenyl)ethylmethylamine obtained from a 25 × 25 CPG network trained with 487 polysubstituted benzene derivatives and their associated IR spectra.

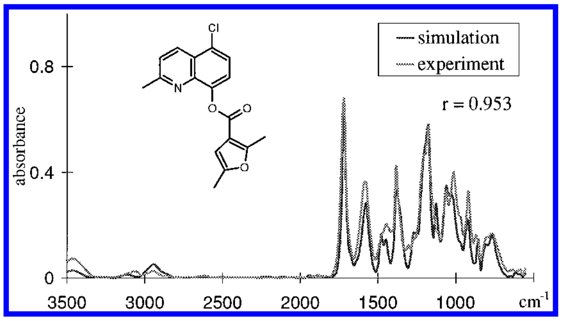


Figure 10. Experimental and simulated IR spectrum of a trisubstituted quinoline compound obtained from a 10 × 10 CPG network trained with 51 polysubstituted quinolines and isoquinolines.

Figure 8 shows the experimental and the simulated IR spectrum of 1-phenylbutanol-1. As can be seen, the correspondence between experimental and simulated IR spectrum is very high as also expressed by the value of 0.961 for the correlation coefficient. It is particularly gratifying that also the fingerprint region of the IR spectrum is quite well represented. This attests to the power of the 3D MoRSE code for representing the entire structure of a molecule, also those parts that consist of larger substructures including molecular skeletons. Figure 9 gives an example of a simulation with a lower correlation coefficient of 0.701.

Even here, the correspondence between the experimental and the simulated infrared spectrum is for a large part of the spectrum quite good. Most of the deviations are due to variations in intensities, the most drastic one at about 1600

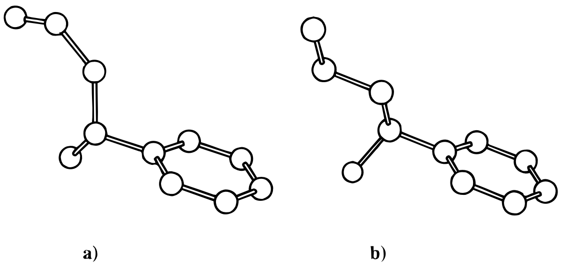


Figure 11. 3D structure of 1-phenyl-butanol-1 obtained from the infrared spectrum (b), compared with that generated by CORINA (a).

cm^{-1} for the band of the aromatic ring. It is known that this vibration is notorious for changing quite dramatically in intensity. Under these circumstances the simulated IR spectrum must be considered to be quite good. It has to be stressed again that 65% of the simulated spectra were of this or higher quality having a correlation with the experimental IR spectrum in the range of 0.70−1.00.

In another study, a CPG network consisting of 10 × 10 neurons was trained with 51 quinoline and isoquinoline derivatives bearing a variety and different numbers of substituents and their associated IR spectra. Another 72 molecules were tested against this network leading to a correlation coefficient higher than 0.50 for 75% of the molecules; for 44% of the structures the correlation between simulated and experimental IR spectrum was higher than 0.75.

Figure 10 gives an example, again from the test set, of a structure with a rather high correlation coefficient of 0.953. It can be seen that this approach gives good results for even quite complex structures.

In summary, the 3D MoRSE code gives quite good simulations of infrared spectra over the entire range of wavenumbers, even of the fingerprint region. This underscores the power of the 3D MoRSE code of reflecting the entire structure of a molecule, both of skeletons and substituents.

EPILOGUE

A counterpropagation (CPG) network stores in its upper part the structure and in its lower part the corresponding infrared spectrum (see Figure 7). So far we have used the CPG network for the simulation of IR spectra by inputting a structure in the form of its 3D MoRSE code and outputting an IR spectrum.

However, a CPG network can also be used the other way around, inputting an IR spectrum and outputting the 3D code of a molecule. Such a reverse use of a CPG network trained with 3D structures and associated IR spectra has been made a centerpiece of an approach for deriving the 3D structure of a compound from its IR spectrum.[39] Recently, Larissa and Valentin Steinhauer have succeeded in making this approach work. First examples for the successful prediction of the three-dimensional structure of a compound from its experimental infrared spectrum have been obtained.

Figure 11 shows the 3D structure of 1-phenyl-butanol-1 derived from the infrared spectrum of this compound and compared with the 3D structure obtained from CORINA. The two models of the 3D structure only differ in their conformations. As CORINA generates only one conformation from the manifold of low energy conformations, the two 3D structures are essentially identical. Clearly, this approach opens new dimensions for the experimental determination of the 3D structure of molecules, supplementing methods based on X-ray diffraction and multidimensional NMR spectroscopy.

CONCLUSION

The 3D MoRSE code has already been used for the correlation of various physical, chemical, and biological data with the structure of a molecule. In particular, as shown here, it has great promise for the modeling of biological data and the simulation of infrared spectra.

One of the most exciting perspectives of this work is that counterpropagation networks trained with the relationships between structure and infrared spectra can also be used for the prediction of the 3D structure of molecules from their infrared spectra.

The representation of molecules by the 3D MoRSE code is a major shift in paradigm in the computer representation of compounds. Connection tables are intimately tied to the concept of a chemical bond existing between two atoms. Thus, they are a representation of a single valence bond structure of a molecule. A connection table is quite adequate for representing a molecule as long as this molecule can reasonably well be represented by a single valence bond structure. Molecules where chemical bonding goes beyond two-center two-electron bonds such as organometallic π-complexes or boranes escape a reasonable representation by connection tables.

The 3D MoRSE code, on the other hand, does—at its very basis—not care about chemical bonds. It only reflects the three-dimensional arrangement of the atoms of a molecule. (It might only be that the calculation of atomic properties A_i, used in eq 4 is based on classical connection tables). Thus, the 3D MoRSE code is quite applicable to species having unusual types of bonding such as organometallic complexes or boranes.

ACKNOWLEDGMENT

The early days of jointly venturing with Clemens Jochum into the representation of chemical structures were exciting times. With Reiner Luckenbach I have shared much work on fostering the standing and image of the Division of Chemical Information and Computers (CIC) of the German Chemical Society (GDCh). We thank Chemical Concepts for providing access to the SpecInfo database. We appreciate the many discussions with Jure Zupan on neural networks and their use in finding structure-spectra correlations. Financial support of our work by the Bundesminister für Bildung, Wissenschaft, Forschung und Technologie (BMBF) and Deutsche Forschungsgemeinschaft is gratefully acknowledged.

REFERENCES AND NOTES

(1) Jochum, C.; Gasteiger, J. Canonical Numbering and Constitutional Symmetry. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 113−117.

(2) Carhart, R. E. Erroneous Claims Concerning the Perception of Topological Symmetry. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 108−109.

(3) Jochum, C.; Gasteiger, J. On the Misinterpretation of Our Algorithm for the Perception of Constitutional Symmetry. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 49−50.

(4) Carhart, R. E. Perception of Topological Symmetry. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 56.

(5) Dyott, T. M.; Howe, W. J. Canonical Numbering. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 187.

(6) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.

(7) Luckenbach, R. *Dynamic Stereochemistry of Pentaco-ordinated Phosphorus and Related Elements*; Georg Thieme Publishers: Stuttgart, 1973.

(8) Gasteiger, J. A. Representation of Pi-Systems for Efficient Computer Manipulation. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 111−115.

(9) Welford, S. M. In *Software − Entwicklung in der Chemie 1*; Gasteiger, J., Ed.; Springer: Berlin, 1987; pp 5−11.

(10) Bauerschmidt, S.; Gasteiger, J. Unpublished results.

(11) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; Smith, J. M.; Watson, D. G. The Development of Version 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187−204.

(12) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.

(13) Hiller, C.; Gasteiger, J. In *Software Entwicklung in der Chemie 1;* Gasteiger, J.; Ed.; Springer-Verlag; Berlin, 1987; pp 53−66.

(14) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Method.* **1992**, *3*, 537−547.

(15) Sadowski, J.; Rudolph, C.; Gasteiger, J. The Generation of 3D Models of Host−Guest Complexes. *Anal. Chim. Acta* **1992**, *265*, 233−241.

(16) Sadowski, J.; Schwab, C.; Gasteiger, J. Unpublished results.

(17) ALCOGEN is available from Chemical Concepts, Weinheim, Germany.

(18) Davies, K.; Dunn, D.; Upton, R. *An Algorithm to Generate 3D Structures from 2D Connection Tables;* Poster on the 5th Molecular Modeling Workshop; Darmstadt, 1991.

(19) (a) Leach, A. R.; Prout, K. Automated Conformational Analysis: Directed Conformational Search Using the A* Algorithm. *J. Comput. Chem.* **1990**, *11*, 1193−1205. (b) Leach, A. R.; Smellie, A. S. A Combined Model-Building and Distance Geometry Approach to Automated Conformational Analysis and Search. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 379−385. (c) COBRA is available from Oxford Molecular Ltd., Oxford, England.

(20) (a) Rusinko, A., III. Tools for Computer-Assisted Drug Design. Ph.D. Thesis, University of Texas at Austin, Austin, TX, 1988. (b) Pearlman, R. S. Rapid Generation of High Quality Approximate 3D Molecular Structures. *Chem. Des. Auto. News* **1987**, *2*, 1/5−6. (c) *Concord* User's Manual; TRIPOS Associates: St Louis, MO, 1988. (d) CONCORD is available from TRIPOS Associates.

(21) Gordeeva, E. V.; Katrizky, A. R.; Shcherbukhin, V. V.; Zefirov, N. S. Rapid Conversion of Molecular Graphs to Three-dimensional Representation Using the MOLGEO Program. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 102−111.

(22) CONVERTER, available from Biosym Technologies, San Diego, CA, U.S.A. We thank Dr. M. Waldman for providing us with the results of CONVERTER.

(23) This kind of representation of the quantity-quality characteristics has been suggested by Dr. V. van Geerestein (AKZO Organon, Oss, The Netherlands).

(24) Schuur, J. H.; Selzer, P.; Gasteiger, J. The Coding of the Three Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure − Spectra Correlations and Studies of Biological Activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334−344.

(25) Wierl, R. Elektronenbeugung und Molekülbau. *Ann. Phys.* (Leipzig) **1931**, *8*, 521−564.

(26) Soltzberg, L. J.; Wilkins, C. L. Molecular Transforms: A Potential Tool for Structure Activity Studies. *J. Am. Chem. Soc.* **1977**, *99*, 439−443.

(27) Gasteiger, J. Empirical Methods for the Calculation of Physicochemical Data of Organic Compounds. In *Physical Property Prediction in Organic Chemistry*; Jochum, C., Hicks, M. G., Sunkel, J., Eds.; Springer-Verlag: Heidelberg, 1988; pp 119−138.

(28) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(29) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.

(30) The steroid structures can be retrieved by anonymous ftp: ftp://cygnus.organik.uni-erlangen.de. login/ftp://cygnus.organik.uni-erlangen.de/pub/steroids/steroids.sd.

(31) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity − A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(32) Gasteiger, J.; Saller, H. Calculation of the Charge Distribution in Conjugated Systems by a Quantification of the Resonance Concept. *Angew. Chem.* **1985**, *97*, 699−701; *Angew. Chem., Int. Ed. Engl.* **1985**, *24*, 687−689.

(33) Kohonen, T. *Self-Organisation and Associative Memory*, 3rd ed.; Springer: Berlin, 1989.

(34) Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem.* **1993**, *105*, 510−536; *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503−527.

(35) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists − An Introduction;* VCH: Weinheim, 1993.

(36) Hecht-Nielsen, T. Counterpropagation Networks. *Applied Optics* **1987**, *26*, 4979−4984.

(37) SpecInfo, available from Chemical Concepts, Weinheim, Germany.

(38) Novic, M.; Zupan, J. Investigation of Infrared Spectra-Structure Correlation Using Kohonen and Counterpropagation Network, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 454−466.

(39) Steinhauer, L.; Steinhauer, V.; Gasteiger, J. Obtaining the 3D Structure from Infrared Spectra of Organic Compounds Using Neural Networks. In *Software Development in Chemistry 10;* Gasteiger, J., Ed.; Gesellschaft Deutscher Chemiker, Frankfurt am Main, Germany, 1996.

CI960343+