The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity

Jan H. Schuur, Paul Selzer, and Johann Gasteiger*

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

Received November 31, 1995[⊗]

A molecular transform, derived from an equation used in electron diffraction studies, is developed that allows the representation of the three-dimensional structure of a molecule by a fixed number of values. Various atomic properties can be taken into account giving high flexibility to this representation of a molecule. This 3D-MoRSE (Molecule Representation of Structures based on Electron diffraction) code retains important structural features such as the mass (see ref 35) and the amount of branching as evidenced by an investigation of monosubstituted benzene derivatives. Furthermore, this molecular representation was able to distinguish between benzene, cyclohexane, and naphthalene derivatives in a dataset of great structural variety. This molecular representation was used in counterpropagation neural networks to distinguish between dopamine D1 and D2 agonists and to group 31 steroids binding to the corticosteroid binding globulin receptor into compounds of high, medium, and low activity. Great promise is given to this representation of molecular structures for the simulation of infrared spectra as revealed by an investigation of monosubstituted benzene derivatives.

INTRODUCTION

Many physical, chemical, or biological properties of compounds are dependent on the three-dimensional arrangement of the atoms in a molecule. The analysis of such structure−property relationships should therefore take account of the 3D structure. This has become standard usage in qualitative molecular modeling studies of substrate-receptor interactions. However, in the absence of a proper 3D structure coding the investigation of *quantitative* relationships between structure and biological activity usually has to be performed without a detailed representation of the 3D structure. This is one of the reasons why datasets of molecules with different skeletons are difficult to handle in a single QSAR investigation. At most, the 3D structure is only accounted for by gross features of the shape of a molecule given by the main extensions of a molecule (length and thickness) as expressed by Verloop′s STERIMOL parameters.[1]

Certain spectral data also depend on the 3D structure of a molecule. This is particularly true for the infrared spectrum that monitors the vibrations of different parts of a molecule in 3D space. The empirical study of the relationships between structure and infrared spectrum has reached a point that still leaves much to be desired. This largely hinges on the poor representation of the chemical structure that usually is performed by an analysis of molecular fragments derived from the constitution of the molecule (topological fragments). Such an approach has several drawbacks. First, the list of fragments to be considered has to be fairly large, so systems with up to 700 fragments have been developed.[3,4] Yet, this fragmentation of a molecule will always be incomplete as the number of fragments to be accounted for is basically unlimited. Extension of such an approach to three-dimensional fragments would only aggravate the problems as then the number of fragments will have to be even larger. Thus, a completely different approach for coding the 3D structure of molecules is necessary before a genuine step forward in our empirical understanding of the relationships between chemical structure and infrared spectra is possible, particularly as concerns the fingerprint region.

Consideration of the 3D structure of organic compounds in QSAR studies or structure-spectra correlations has been hindered by the lack of data on the three-dimensional structure of the compounds to be considered. The three-dimensional structure of a molecule can experimentally be derived from electron or X-ray diffraction studies or from NMR measurements of Nuclear Overhouser Effects (NOE). However, in comparison to the number of known organic compounds (10−12 million), the number of compounds for which the 3D structure has been determined (approx. 100 000) is minute. Thus, for most datasets of compounds to be investigated in QSAR or structure-spectra studies, the three-dimensional structure is known only for a few molecules preventing consideration of the 3D structure from the very beginning. However, recently systems have been developed that can automatically generate a three-dimensional model of a molecule from information on its constitution and stereochemical descriptors for a wide range of organic compounds.[5−8] These systems come up with 3D structures that are quite close to the experimental structures, in those cases where they are known as revealed by an analysis of 639 organic and organometallic compounds and their X-ray structures.[8]

The 3D structure generator CORINA developed in our group was recently shown to have a broad applicability and high conversion rate. A dataset of 126 705 structures of the National Cancer Institute, Washington, DC, USA, had been made available in the form of connection tables. This information on the constitution of molecules was used as input to the CORINA system leading to 3D structures for 126 148 molecules. Of the 557 structures that were not

[*] Author to whom correspondence should be addressed.

[⊗] Abstract published in *Advance ACS Abstracts,* February 15, 1996.

0095-2338/96/1636-0334$12.00/0

converted into 3D coordinates, 327 structures had atoms with a coordination number higher than six and were therefore outside of the defined scope of CORINA. With such a small failure rate of 230 structures (0.2%) the availability of 3D structures for the entire range of organic molecules has safely been secured.[9]

Such a general access to the 3D structure of organic compounds can now lay the foundation to an analysis of the relationships between structure and physical, chemical, or biological data by accounting for the three-dimensional arrangement of the atoms in a molecule. This immediately raises the question of how to code the 3D structure for studies of structure−property relationships. The most obvious way for coding the 3D structure of a molecule by specifying the Cartesian or internal coordinates of the atoms is unfavorable for most applications. Nearly all data analysis methods, be it statistical or pattern recognition methods or neural networks, ask for a representation of the objects to be studied by the same number of variables. The number of 3D coordinates, however, is intimately tied to the number of atoms in a molecule—three coordinates for each atom. Thus, datasets with molecules having different numbers of atoms cannot be coded merely by giving atomic coordinates because this would result in a varying number of coordinates for the molecules under investigation.

We present here a representation of the three-dimensional structure of organic molecules that is characterized by a fixed number of variables. This opens the way to a wide range of studies of the relationships between structure and physical, chemical, and biological properties by statistical procedures, pattern recognition methods, or neural networks.

A TRANSFORMATION OF THE 3D STRUCTURE

The secret to success was found by a careful scrutiny of the experimental methods for the determination of the 3D structure of molecules. Methods such as electron and X-ray diffraction do not directly yield atomic coordinates but provide diffraction patterns from which the atomic coordinates are derived by mathematical transformations.

It was already realized some time ago in a seminal paper by Soltzberg and Wilkins[10] that three-dimensional atomic coordinates can be transformed into a molecular code by modification of an equation used in electron diffraction studies for preparing theoretical scattering curves. Several steps of simplifications were introduced in these equations to eventually arrive at a binary encoding pattern.

We first follow the reasoning of Soltzberg and Wilkins without making such drastic simplifications necessary to end up with a binary pattern. Rather, we will stop the simplifications at a much higher level of information and then open the way to great flexibility in molecular encoding by introducing a variety of atomic properties into the encoding scheme.

As pointed out by Soltzberg and Wilkins, the structure factors of X-ray data contain information that goes beyond the molecular structure, most prominently information on the orientation of molecules in the unit cell. This additional information is unwanted for the proper coding of the molecular structure. In this context, data from electron diffraction studies offer a better starting point for the representation of molecular structure, uncorrupted by crystal effects.

The general molecular transform used in electron diffraction studies is given by

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

This equation represents the scattering in various directions, $\vec{S}$, by an ensemble of N spherical scatters (atoms) located at points $\vec{r}_i$; f_i are the form factors that take into account the directional dependence of scattering from a spherical body of finite size. The relationship shown in eq 1 is usually handled in electron diffraction studies in a form as proposed by Wierl[11] and represented in eq 2

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

In this equation, s measures the scattering angle given by

$$s = 4\pi \sin(\vartheta/2)/\lambda \qquad (3)$$

with ϑ being the scattering angle and λ the wavelength. $I(s)$ is the intensity of the scattered radiation, r represents the interatomic distances, $P_{ij}(r)$ is the probability distribution of the vibrational variation in the distance between atoms i and j, f_i and f_j are the form factors of atoms i and j. K collects various constants that depend on the instrument.

Following Soltzberg and Wilkins[10] we have made the simplifications

$$K = 1$$

$$P_{ij}(r) = \delta(r - r_{ij})$$

effectively assuming the atoms to be point scatterers and the molecule to be rigid. In addition, for the form factors f_i different atomic properties A_i were used. This constitutes an extension of the approach taken in ref 10 where f_i was set equal to the atomic number Z_i. We, however, also use other atomic properties A_i, like atomic mass, partial atomic charge,[12,13] residual atomic electronegativities,[14] and atomic polarizabilities[15] calculated by previously published empirical methods.

This leads to eq 4

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

$$s = 0, ..., 31.0 \text{ Å}^{-1}$$

Values of this function were calculated at 32 evenly distributed values of s in the range of 0−31.0 Å^{-1} from the three-dimensional atomic coordinates of a molecule as obtained by the 3D structure generator CORINA.[5−8] Whereas Soltzberg and Wilkins only coded the zero crossings of the function shown in eq 4 (with A_i being the atomic number, Z_i) we took the 32 actual values of $I(s)$ between 0 and 31 Å^{-1}. These 32 values constitute the 3D-MoRSE (Molecule Representation of Structures based on Electron diffraction) code of the three-dimensional structure of a molecule.

Figure 1 shows the 32 values of $I(s)$ for benzene, cyclohexane, and naphthalene. It can be seen that these three structures have a code that indicates quite distinct common and differing features.

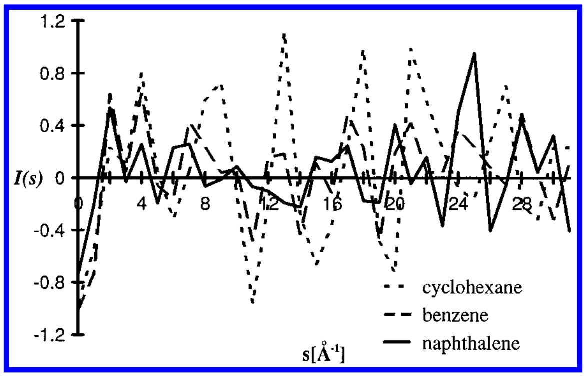


Figure 1. The 32 3D-MoRSE value $I(s)$ of the benzene, cyclohexane, and naphthalene molecule.

It is important to note that the molecular transform code derived here is independent of the orientation of a molecule in space. Small variations in the values of $I(s)$ observed on translation or rotation of a molecule result from the limits on resolution, by only choosing 32 discrete values of $I(s)$.

EXPERIMENTAL SECTION

Input of molecular structures was made by a graphical molecule editor that automatically generates a connection table (list of atoms and bonds in a molecule) together with stereochemical descriptors where appropriate. In other cases, sets of structures were downloaded from datafiles in a standard format such as the MOLFILE format. This information on the constitution of a molecule was converted into three-dimensional coordinates of the atoms by the 3D structure generator CORINA.[5−8]

Various atomic properties were calculated directly from the information in the connection table by previously published empirical methods: atomic charges, q_i, by the PEOE method and its extension to conjugated π-systems.[12,13] Concurrent with this, values on residual electronegativities, χ_i, as quantitative measures of the inductive effect are obtained.[14] Values for the effective polarizabilities, α_i, on the atoms of a molecule were calculated as published in ref 15. Representations of the 3D structure of molecules were calculated using different atomic properties, such as mass, m_i, atomic number, Z_i, or the variables mentioned above: charge, q_i, electronegativity, χ_i, or polarizability, α_i, for A_i and A_j in eq 4.

The values of A_i were scaled by multiplying them with 1/abs (max A_i) where max A_i was obtained from a standard data set. Sometimes, combinations of different atomic properties were used in place of A_i. Then, not A_i itself is scaled by this scaling procedure but all atomic properties being part of A_i are scaled individually.

Values of the function $I(s)$ (eq 4) were calculated at 32 evenly distributed points between 0 and 31.0 Å^{-1}. These 32 values were taken as a representation of the 3D structure of a molecule and input to data analysis methods such as principal component analysis (PCA) or a counterpropagation neural network (CPG). Clearly, both the number of discrete values $I(s)$ as well as the sampling range for these variables can be changed. A higher density of points corresponds to a better resolution in the representation of the 3D structure of molecules. For the studies reported here, the resolution chosen with 32 values between 0 and 31.0 Å^{-1} was found to be sufficient.

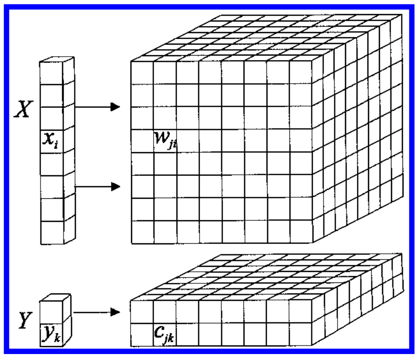


Figure 2. Model of a counterpropagation-network. The upper block takes the input variables X, and the lower block the output variables Y. Input and output vectors X and Y are separate columns at the left-hand side. The neurons are arranged in a two-dimensional manner viewed from the top. They are represented as columns of small boxes, each containing one weight w_{ji} or c_{jk}.

A principal component analysis (PCA) is basically an unsupervised method for the projection of data from a high-dimensional measurement space into a lower-dimensional space. The projection is such that as much as possible of the variation in the data goes into the first components (coordinates) of the lower-dimensional space. Central to a PCA is the calculation of eigenvalues for the data matrix. This was accomplished by standard techniques.[16] Many of the investigations were performed by a counterpropagation neural network (CPG)[17] and therefore this technique deserves a more detailed explanation.[18,19]

A counterpropagation neural network is used for finding a relationship between one or several quantities $Y(y_1, y_2, ...)$, the dependent variables, and input data $X(x_1, x_2,...)$, the independent variables. This relationship $Y = f(X)$ is not expressed in explicit mathematical form but is stored in weights of so-called (artificial) neurons.

A CPG network consists of two blocks, an input block for the x-variable(s) and an output block for the y-variable(s). The processing elements, the neurons, are arranged in a two-dimensional manner (Figure 2).

Each neuron, j, has as many weights, w_{ji}, as there are input variables, x_i, and weights, c_{jk}, as there are output variables, y_k, for each object (sample), X_s, Y_s.

The training of a CPG network is by a competitive learning process such that an object, X_s, is mapped into a specific neuron. That neuron, c, will be the winning (central) neuron that has weights most similar to the input data, x_{si}, of a sample, s, (eq 5); m = number of considered variables; n = number of neurons.

$$\mathrm{out}_{sc} \leftarrow \min\left[\sum_{i=1}^{m}(x_{si} - w_{ji})^2\right] \quad j = 1, 2, 3, ..., n \qquad (5)$$

In the learning mode, unsupervised or supervised learning determines if only the input variables, X, (unsupervised learning), or both input and output variables, X and Y, are compared in the search for the winning neuron. So if we use the terms supervised or unsupervised to describe the training methods of the following experiments it is always in relation to this definition, which may be different to other publications.[20]

However, both the weights, w_{ji} and c_{jk}, will be adjusted after each new input of $\{X,Y\}$ pair data. The weights of the neurons of the network will be adjusted such that they become more similar to the input data. In this process, the weights of the winning neuron are adjusted to the largest degree, the change in the weights decreases with increasing distance from the central neuron *c*.

The following two equations are invoked for weight correction

$$w_{ji}^{\text{new}} = w_{ji}^{\text{old}} + \eta(t)a(d_c - d_j)(x_i - w_{ji}^{\text{old}}) \qquad (6)$$

$$c_{jk}^{\text{new}} = c_{jk}^{\text{old}} + \eta(t)a(d_c - d_j)(x_k - c_{jk}^{\text{old}}) \qquad (7)$$

with η, learning rate; t; number of current training iteration; $d_c - d_j$; topological distance between neuron c and neuron j; $a(d_c - d_j)$, topology dependent scaling function of the learning rate η.

As of now, we have discussed the situation that only the input variables, x_i, are considered in the learning process. This is an unsupervised learning process as the output values Y are not considered in the determination of the winning neuron. In this case, learning in a CPG network is equivalent to learning in a self-organizing feature map such as in a Kohonen network.[21,22]

Training of a CPG network results in mapping each sample from the dataset into a specific neuron. Clearly, several samples can be mapped into the same neuron if they are quite similar. Inspection of the two-dimensional arrangement of neurons (Figure 2) from the top shows how the entire dataset has been distributed across the network. A trained CPG network will show similarities in the objects of a dataset such that similar data are mapped into the same or closely adjacent neurons.

A CPG network, however, cannot only show *similarities* in the training data but can also be used for making *predictions* for these data. Test data are mapped into a trained CPG network in the same way as training data by using eq 5 for determining the winning neuron. However, no weight correction is anymore performed. By only using the input layer, the similarity of the test object to those in the training dataset can be determined. When the information on the output layer is considered, predictions on the results, *Y*, to be expected for the input object can be made. For example, when a counterpropagation network has been trained with the structure of molecules (as input) and their corresponding infrared spectra (as output), the input of a new structure will allow the prediction of its infrared spectrum, and vice versa.

RECOGNITION OF MOLECULAR SKELETONS

One of the major objectives for the development of a new structure coding was to find a consistent representation of molecular skeletons. Working with fragment codes for different skeletons and substitution patterns necessarily leads to a high— and always incomplete—number of fragments that have to be considered. What is needed is a mathematical transformation that can be applied to any molecular skeleton and any distribution of substituents on such skeletons in a consistent manner.

The merits of the 3D-MoRSE code will be shown with a dataset of substituted cyclohexanes, benzenes, and naphthalenes. This dataset was chosen in order to see whether the addition of six hydrogen atoms to the benzene ring (to give cyclohexanes) as well as the fusion of an extra aromatic ring to the benzene ring (to give naphthalenes) both provide enough variation to the benzene ring to be easily discernible in the 3D structure code.

The structures were derived from the SpecInfo database[23] containing infrared spectra of 13 373 molecules. The selection of molecules was restricted to compounds (i) only containing C, H, N, O, F, Cl, or Br atoms, (ii) bearing one to three substituents on the benzene ring, and (iii) having substituents with a maximum number of six non-hydrogen atoms (e.g. a phenyl ring). This provided a dataset of 871 benzene derivatives, 148 substituted cyclohexanes, and 81 naphthalene derivatives.

In order to have a statistically more balanced dataset between the different skeletons, the dataset of the benzene derivatives was further reduced simultaneously ensuring that the structural variation is as large as possible. This was achieved by a counterpropagation (CPG) network.[17−19] As previously shown[24] the selection of a training dataset based on projection by a Kohonen neural network leads to results that are superior to a random selection of the training dataset or to the selection by an experimental design technique. The CPG network was used in an unsupervised learning mode and thus is basically using a Kohonen learning procedure. The input to a 14 × 14 CPG network consisted of the 32 structural variables, the values of $I(s)$, as calculated by eq 4 from the 3D atomic coordinates of a molecule (cf. Figure 2). These 32 values constituted the input block of the counterpropagation network. The output block consisted of seven layers each reserved for one of the possible substitution patterns, mono, ortho, meta, para, 1,2,3-tri, 1,2,4-tri, and 1,3,5-tri substitution, respectively. These layers only served to illustrate the separation of the various substitution patterns.

The counterpropagation network was trained with all 871 benzene derivatives in an unsupervised manner, not taking account of the assignment of the molecules to one of the seven substitution classes. This training spread the dataset of 871 structures across the 196 neurons of the 14 × 14 CPG network. Necessarily, in many cases several structures were assigned to the same neuron. Only one of these structures was taken into the reduced dataset, the rest were discarded. Few neurons were empty; they did not receive any molecule at all. Thus, the reduced dataset contained 175 benzene derivatives; it would have had 196 structures if each neuron had been filled.

For each different atomic parameter A_i the selection of the dataset was individually performed. A typical dataset with the atomic parameter $A_i = \alpha_i$ consisted of 148 cyclohexane derivatives, 81 substituted naphthalenes, and 175 benzene derivatives.

The influence of the atomic parameter A_i on the separation of the derivatives of the three skeletons was studied. Each value of the code can be regarded as a coordinate of a space. Thus, the entire 32 3D-MoRSE values span a 32-dimensional space where each structure corresponds to a point in this space. The question is now whether this code is able to separate the points belonging to the different skeletal classes into individual clusters. The separation of *two* classes into separate clusters can be measured by the Fisher quotient, and its value was calculated for the three different combinations benzenes/cyclohexanes, benzenes/naphthalenes, and cyclohexanes/naphthalenes. However, in order to simulta-

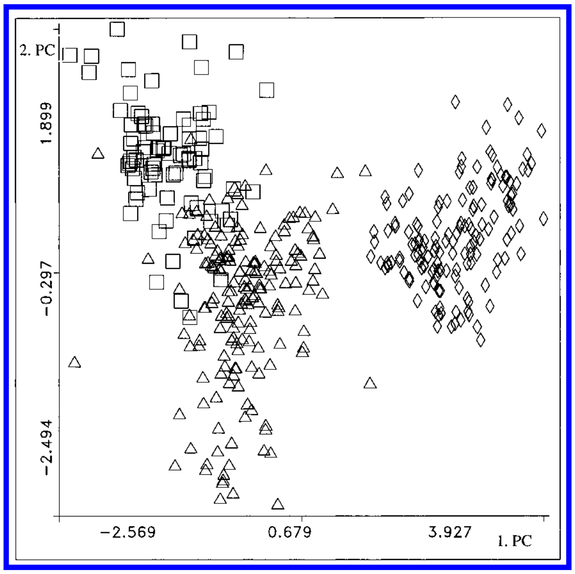


Figure 3. Plot of the first (PC-1) versus the second principal component (PC-2) from a principal component analysis (PCA) of 32 3D-MoRSE values $I(s)$ using atomic mass, m_i, as atomic parameter A_i in eq 4.

neously study the separation of *three* clusters a principal component analysis (PCA) of the 32-dimensional space was performed.[16] In a principal component analysis the largest amount of variance in the distribution of a dataset goes into the first few components. Therefore, in order to visualize the separation of the dataset, plots of the first few principal components were investigated. Figure 3 shows the results obtained by plotting the second against the first principal component of a PCA on the 32-dimensional encoding scheme embedded in eq 4 when using atomic mass, m_i, as atomic parameter A_i. The variance of the first principal component comprises 54.56% of the total variance in the dataset, the second principal component 11.38%.

In the plot, the benzene derivatives are marked with a triangle, the naphthalene derivatives with a square, and the cyclohexane derivatives with a rhombus. As can be seen, the first principal component separates clearly between cyclohexane derivatives and the benzene and naphthalene derivatives on the other hand. The second principal component distinguishes to a large extent between benzene and naphthalene derivatives.

However, a genuine separation of the three classes of compounds can only be achieved by using another atomic property as parameter A_i in eq 4. A good separation was observed by setting A_i equal to the atomic polarizability α_i.[15] A plot of the first two components of a principal component analysis of the 32-dimensional space spanned by the values of $I(s)$ obtained by eq 4 with $A_i = \alpha_i$ is shown in Figure 4. The variance of the first principal component now comprises 73.06% of the total variance in the dataset; the second principal component 16.02%.

The naphthalene compounds (squares) are well separated from all the other compounds. Furthermore, the benzene and cyclohexane compounds are clearly separated from each other, although not that distinctly. It is also gratifying that the benzene derivatives lie in between the cyclohexane and the naphthalene compounds, a fact reflecting that the benzene skeleton is similar both to the cyclohexane and to the naphthalene ring.

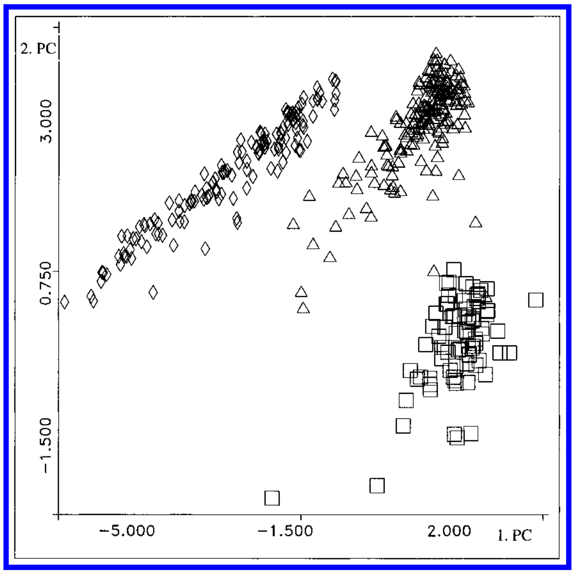
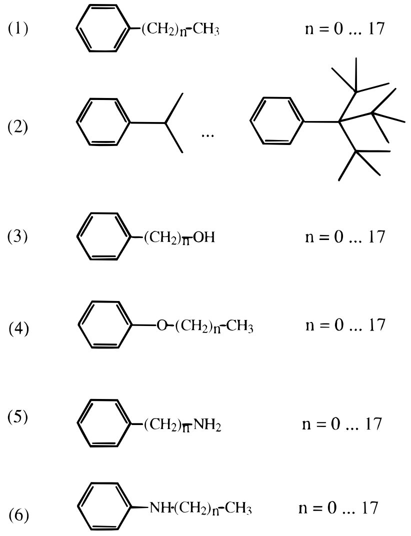


Figure 4. Plot of the first (PC-1) versus the second principal component (PC-2) from a PCA of 32 3D-MoRSE values $I(s)$ using atomic polarizability, α_i, as atomic parameter A_i in eq 4.

Chart 1



INFLUENCE OF THE SUBSTITUENT

In the transformation chosen here, each value, $I(s)$, contains information on the coordinates of all atoms (cf. eq 4). However, each value does so to a different extent. We therefore investigated the information-content of individual 3D-MoRSE values $I(s)$ and their ability to reproduce certain information such as mass or branching of parts of a molecule. In a previous section we have shown that the transformation code is able to retain information on the different *skeletons* of a molecule. Here, we concentrate the investigation on the question whether the transformation can also reflect information on the *substituent* of a molecule. To this end different sets of monosubstituted benzene derivatives were investigated. The 3D-MoRSE code values were calculated with $A_i = q_{\mathrm{tot},i}$. The following series of molecules were investigated (see Chart 1).

MASS

For each individual series, a separate correlation of each transformation value $I(s)$ with the mass of the substituent was made and the square of the regression coefficient r

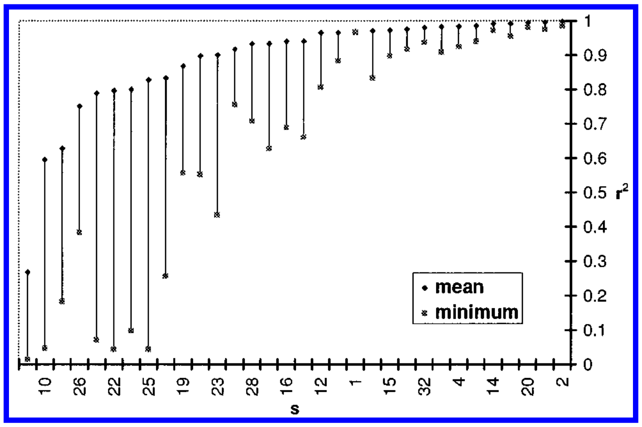
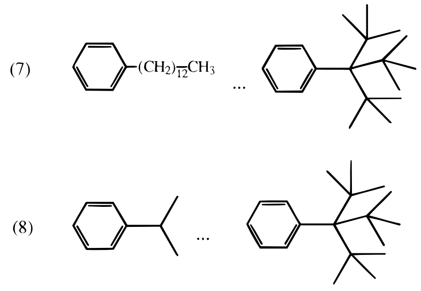


Figure 5. r^2 values for the correlation of the 32 3D-MoRSE values $I(s)$ with the mass of the substituent for the datasets (1)−(6) of Chart 1.

Chart 2



calculated. Then, the average of r^2 of all series of compounds was computed.

Figure 5 shows a plot of this average of r^2 and the minimum value of r^2 in one of these six series of compounds, for each of the 32 values. The figure shows that there is a high degree of correlation with the mass of the substituent for many 3D-MoRSE values $I(s)$ and that this correlation applies to all of the above six series of compounds. For example, in the series (6) of Chart 1 the second transformation value, $I(2)$, has a correlation of $r^2 = 0.999$ with the mass of the substituent, $m_{\text{side chain}}$. The corresponding equation is shown in eq 8.

$$m_{\text{side chain}} = 128.8\, I(2) + 109.5 \qquad (8)$$

An analogous equation (eq 9) with an $r^2 = 0.978$ is obtained for the series (2) of Chart 1.

$$m_{\text{side chain}} = 124.2\, I(2) + 96.7 \qquad (9)$$

BRANCHING

In the two series of Chart 2, the number of branching points in the hydrocarbon substituent is changing. It could be shown that $I(1)$ (and other values) have a high degree of correlation with the number of branching points, $n_{\text{branching points}}$. The corresponding equations are shown in eqs 10 and 11.

$$n_{\text{branching points}}\,(\text{series } 1) = 3.2\, I(1) + 4.1 \qquad (10)$$

$$n_{\text{branching points}}\,(\text{series } 2) = 10.6\, I(1) - 6.0 \qquad (11)$$

CLASSIFICATION OF DOPAMINE D1 AND D2 AGONISTS

Many biological activities of organic compounds strongly depend on the three-dimensional structure of the corresponding molecule. The transformation representation developed here offers a new approach to the representation of the 3D structure of molecules for finding structure−activity relationships. As a typical problem, the potential of this structure representation for the classification of molecules into substrates of different receptors was investigated. Dopamine can bind to two different receptors, the D1 and D2 receptor. Quite a few agonists are known that bind only to the D1 or the D2 receptor. A search in the MDDR-3D database[25] provided 22 dopamine D1 agonists and 67 D2 agonists. The potential of the description of the 3D structure of molecules developed here for the separation of D1 and D2 agonists was investigated by a counterpropagation network trained in unsupervised mode. Input to the network consisted of the 32 $I(s)$ values obtained from eq 4 with the atomic variable, A_i, set equal either to atomic number, Z_i, atomic mass, m_i, total charge, $q_{\text{tot},i}$, or polarizability, α_i, of the atom. In the output level an indicator value was given with 1 specifying D1 and 2 indicating D2 agonists. However, this information was not used in the (unsupervised) training of the network, it only served for identification of the results of the CPG training. Planar CPG networks with 10 × 10 neurons were used. Good separations of the two classes of agonists were obtained for all the above atomic properties. The number of misclassifications ranged between zero and one.

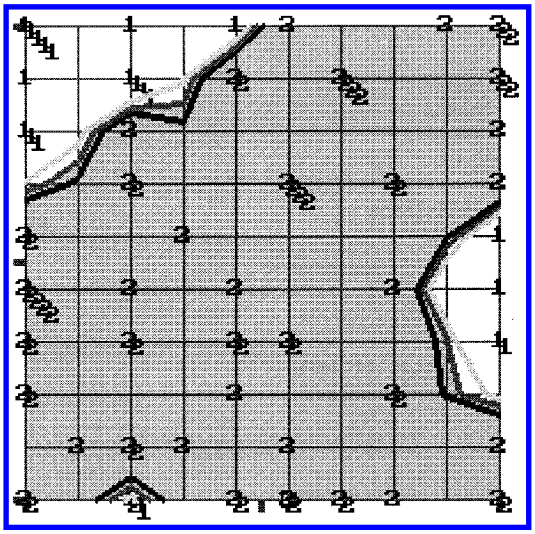
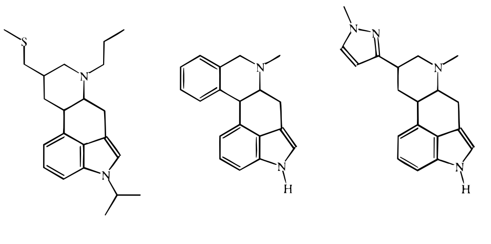


Figure 6. Contour plot of the output level for dopamin D2 agonists of a CPG-network trained with 32 3D-MoRSE values, obtained by using the total atomic charge, $q_{\text{tot},i}$, as atomic parameter A_i (cf eq 4). Points with number 1 or 2 indicate neutrons into which D1 or D2 agonists were mapped, respectively.

A good result was obtained with a 10 × 10 CPG network using the total atomic charge as the atom variable ($A_i = q_{\text{tot},i}$ in eq 4). Figure 6 shows the resulting classification in the output level. The isohypses indicate the strong separation of class 1 and 2. However, the compounds of class 1, the D1 agonists, do not fall into one group but into two distinct groups and a separate neuron with one misclassification.

The contents of the neuron with the conflict between D1 and D2 agonists are shown in Chart 3. Compound **1** is a D2 agonist but falls with compound **2** and **3**, D1 agonists, into the same neuron (Figure 6). All three compounds consist of a tetracyclic ring system consisting of three six- and one five-membered ring, with two nitrogen atoms at equivalent positions. There are two D1 agonists and six D2 agonists in this dataset with this common substructure.

Chart 3



1 2 3

Table 1. Steroids Binding to the Corticosteroids Binding Globulin (CBG) Receptor Together with Their Activity Values (p*K*) and the Activity Classification Taken From Ref 28

training set		
steroid no.	CBG affinity (p*K*)	affinity
s1	−6.279	medium
s2	−5.763	low
s3	−5.613	low
s4	−7.881	high
s5	−7.881	high
s6	−5.000	low
s7	−5.000	low
s8	−5.000	low
s9	−5.000	low
s10	−5.225	low
s11	−5.225	low
s12	−5.000	low
s13	−7.380	high
s14	−7.740	high
s15	−6.724	medium
s16	−7.512	high
s17	−7.553	high
s18	−6.779	medium
s19	−7.200	high
s20	−6.144	medium
s21	−6.247	medium
s22	−7.120	medium
s23	−6.817	medium
s24	−7.688	high
s25	−5.797	medium

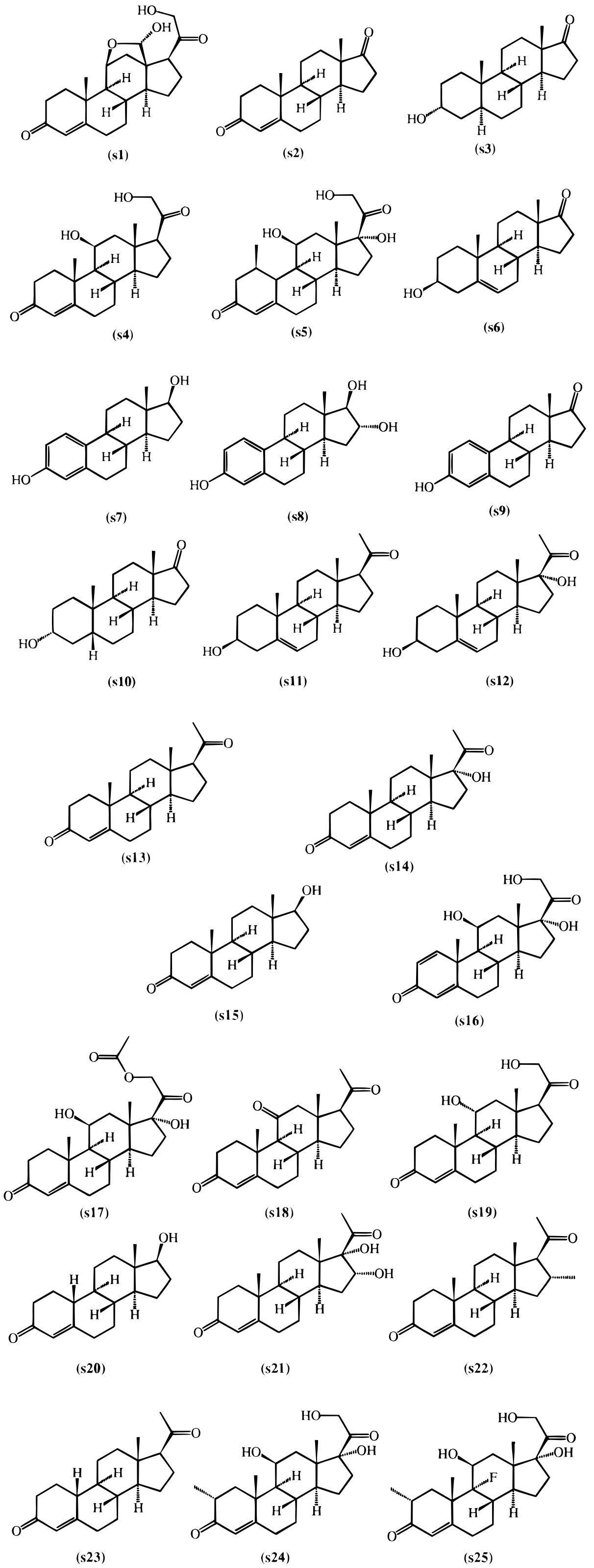
Apparently, this feature is not sufficient for distinguishing between D1 and D2 activity. The additional features responsible for either D1 or D2 activity are correctly perceived by the encoding chart using total atomic charge in eq 4 for all but one of the 8 compounds.

The fact that all atomic properties, m_i, Z_i, $q_{\mathrm{tot},i}$, and α_i can be used in eq 4 for distinguishing between D1 and D2 agonists in this dataset of 89 structures with at most one misclassification shows that not A_i but the atomic distances r_{ij} in eq 4 are the dominating factors for determining D1 and D2 activity. This underscores the need for consideration of 3D atomic coordinates in analyzing biological activity.

CLASSIFICATION OF STEROIDS BINDING TO THE CBG RECEPTOR

Corticosteroids are synthesized in the adrenal gland and are transported in the blood by binding to the corticosteroid binding globulin (CBG). A dataset of 31 corticosteroids for which affinity data were available in the literature[26−29] was investigated. This dataset had already been studied by several groups[26,28,29] using different techniques including the widely accepted Comparative Molecular Field Analysis (CoMFA). Furthermore, the dataset is distributed in computer-readable form.[30,31] Unfortunately, both publications[28,29] and the datafiles[30,31] on these molecules contain errors in the structural formulas or coding errors. This has been rectified by a careful reexamination[32] of the original literature; the dataset shown here is free of these coding errors and is consistent with the information reported in refs 26 and 27.

Chart 4

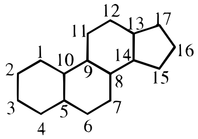


The entire range of activity values was divided into equally spaced intervals of high, medium, and low affinity. The dataset was now split into a training and a test set by randomly taking two molecules of each activity class into the test set. The remaining 25 molecules were taken for training. Table 1 gives the steroids of the training set and

Table 2

	test set	
steroid no.	CBG affinity (p*K*)	affinity
L1	−5.000	low
L2	−5.000	low
M1	−6.892	medium
M2	−5.919	medium
H1	−7.653	high
H2	−7.881	high

Chart 5



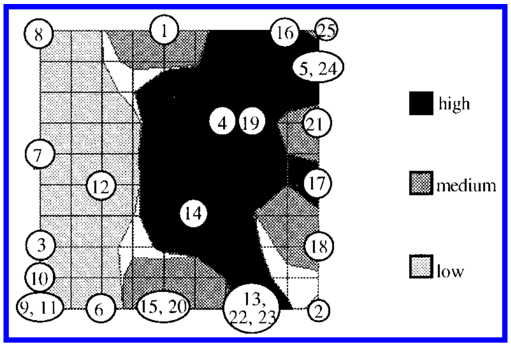


Figure 7. Mapping of the steroids **s1**−**s25** (see Chart 4 and Table 1) into a CPG-network of 10 × 10 neurons during training. The shading indicates areas into which compounds of high, medium, or low activity were mapped.

their CBG-affinities. The corresponding structures are shown in Chart 4. In Table 2 the activities and in Chart 6 the structures of the steroids of the test set are shown. Chart 5 shows the numbering of the steroid skeleton as used in the following discussion. A network of 10 × 10 neurons was trained with the representation of the 25 steroids by the 32 3D-MoRSE values obtained by eq 4 using the total charge on the atoms, $q_{\mathrm{tot},i}$, of the molecule as calculated by the PEOE method.[12,13] With this information as input the CPG network was trained in an unsupervised manner, the output layer contained the information to which class a compound belongs. However, this information was not used in the training; thus, in effect, learning in the CPG network only consisted of learning in the Kohonen layer.

The 25 steroids were grouped in the CPG network as shown in Figure 7. The compounds of low, medium and high CBG affinity separate quite well from each other, with the compounds of medium CBG affinity surrounding a network region of high affinity and the area of compounds with low affinity being further away.

No collisions occur; no molecules of different affinity classes were assigned to the same neuron. In the upper right-hand corner of the map in Figure 7, the compounds **s5**, **s16**, and **s24** of high affinity and compound **s25** of medium affinity were mapped close together. Examining the structural details of these compounds it can be seen that all have an α-hydroxy group in position 11, a *β*-hydroxy group at position 17, and an α-2-hydroxyacetyl group in position 17 together with a keto group in position 3 and a double bond between atoms 4 and 5 of the steroid ring system. Incidentally, compound **s25**, although of medium CBG affinity, also has all these structural features. However, the fluorine atom at position 9 brings in so much additional structural modification that this compound is only of medium affinity. Compounds **s4**, **s14**, and **s19** have the same structural features as the four compounds mentioned above, except that they lack a hydroxy group either at position 11, or at position 17, or at the acetyl group. Steroids **s13**, **s22**, and **s23** lack the three hydroxy groups at position 11, 17, and in the acetyl substituent and have still a high affinity. However, these structural modifications are large enough to separate these compounds from the others with high affinity in the CPG network. One feature that all highly active compounds have in common is a double bond between atoms 4 and 5.

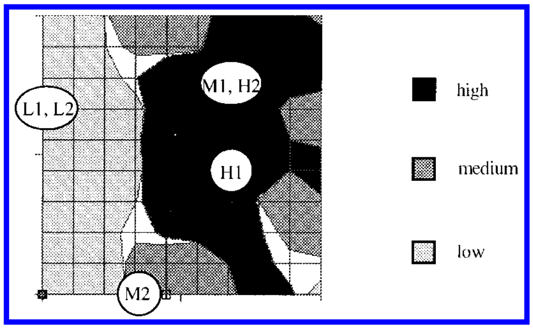
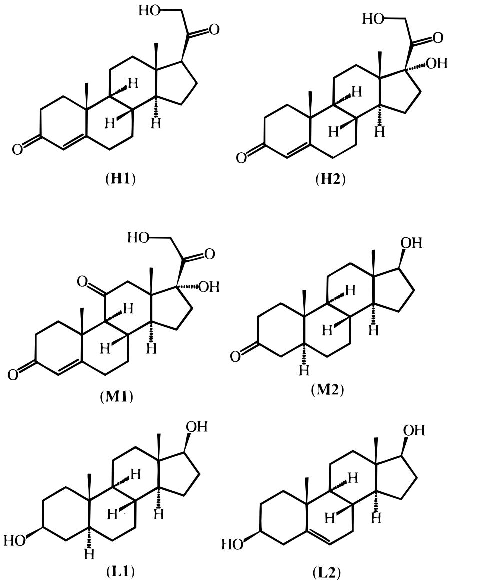


Figure 8. Mapping of a test set of steroids (see Chart 6 and Table 2) into the trained CPG-network of Figure 10.

Chart 6



In the following experiment it was tested to which neurons the six steroids of the test set were assigned. The aim was to find out whether the trained network can be used for the prediction of the activity of unknown compounds. The structures of the steroids of the test set are shown in Chart 6; the corresponding CBG-affinity values can be found in Table 2.

As mentioned before, the test molecules were not included in the training set, the network had not seen them before. Only the 3D-MoRSE code of each molecule was taken into account for choosing the winning neuron. The molecules were mapped into the network as shown in Figure 8.

Five molecules were assigned to the correct activity regions, whereas one steroid (**M1**) is misclassificated. This wrong assignment can easily be understood looking at the structural features of this molecule. It shows high structural similarities to the molecules of high affinity (**H1** and **H2**), especially the double bond between position 4 and position 5 that is supposed to be a major reason for high CBG-affinity. The keto group at position 11 is the reason for the medium activity of this molecule.

To summarize, the representation of the 3D structure of compounds making use of electronic effects as embodied in the partial atomic charges is able to separate steroids of high, medium, and low binding affinity to the corticosteroid binding globulin (CBG). An analysis of the clusters of molecules allows one to identify the essential features for high binding affinity. Consideration of compounds with medium affinity being grouped close to those of high affinity shows which structural changes can no longer be accommodated without loss of binding affinity.

SIMULATION OF INFRARED SPECTRA OF MONOSUBSTITUTED BENZENE DERIVATIVES

One of the incentives for the development of the new representation of the 3D structure of molecules was their use in structure-infrared spectra correlations. In a recent publication is shown that correlation tables between IR-frequencies and topological fragments cannot reliably be used for automatic spectra interpretation.[33] To overcome these limitations, consideration of the 3D structure seemed necessary. These relationships were investigated by a counter-propagation network containing the structure as represented by 32 3D-MoRSE values $I(s)$ in the input layer and the infrared spectrum given by 128 absorbance values in the output layer (see Figure 2). The 128 absorbance values were obtained by the following procedure: The infrared spectra were first brought to an equal resolution in the frequency range by interpolation. From 3500 to 2000 cm^{-1} the infrared spectra were digitized to 150 equally spaced points, one per 10 cm^{-1}. In the range from 2000−552 cm^{-1} the spectra were digitized to 362 points, one per 4 cm^{-1}. These 512 absorbance values were transformed into 512 Hadamard coefficients. The Hadamard coefficients 129 to 512 were set to zero. Then the 512 coefficients were transformed back by reverse Hadamard transformation into absorbances. This results in 512 values consisting of 128 groups of four equal values. For the representation of the IR spectrum only every fourth value was taken, leading to 128 absorbance values with a resolution of 40 cm^{-1} between 3500 and 2020 cm^{-1} and a resolution of 16 cm^{-1} between 2000 and 560 cm^{-1}.

To obtain a dataset, all monosubstituted benzene derivatives with less than nine non-H atoms in the substituent and their corresponding spectra were taken from the SpecInfo datafile.[23] This provided 185 structures. This datafile was split into a training and a test set.

It had been shown previously that the self-organizing capability of a Kohonen neural network provides an excellent basis for the selection of a training dataset that covers the information space as good as possible.[24] A Kohonen network consisting of 10 × 10 neurons was trained with the 185 monosubstituted benzene derivatives taking 32 3D-MoRSE values $I(s)$ obtained with $A_i = q_{\mathrm{tot},i}$, as input. The 185 structures were distributed by the Kohonen learning method over the 100 neurons in such a way that 75 neurons were filled with structures; 25 neurons stayed empty. From each of the occupied neurons one structure was selected for the training set, the others were taken into the test set. This ensured that the training data were taken from the entire space containing information. This gave a training set of 75 monosubstituted benzene derivatives and 110 structures in the test set.

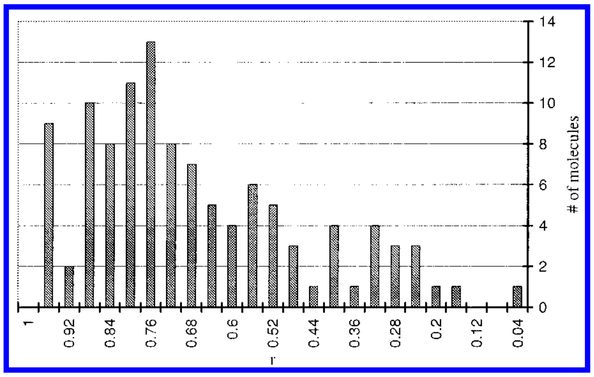


Figure 9. Distribution of the correlation coefficient *r* between simulated and experimental spectrum of monosubstituted benzene derivatives.

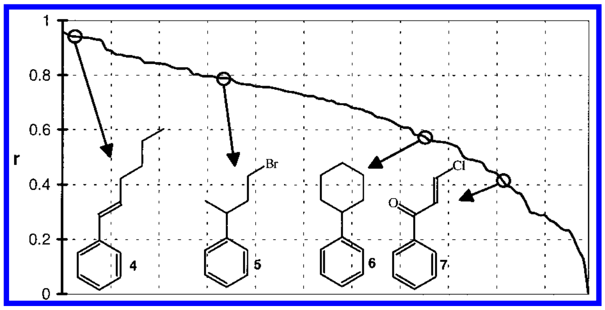


Figure 10. Four molecules, with different correlation coefficients *r* between simulated and experimental spectra, were selected as examples for simulations of varying quality (see Figures 11−14).

A counterpropagation network was trained with these 75 benzene derivatives taking 32 3D-MoRSE values $I(s)$ obtained from eq 4 with $A_i = q_{\mathrm{tot},i}$ and storing the 128 absorbances of the infrared spectrum in the output layer. Training was, however, performed in an unsupervised manner, only the structure representation and not the spectral information was considered in the competitive learning process. In determining the winning neuron by eq 5 only the 32 input values were considered. However, adjustment of weights by eqs 6 and 7 was performed on all 32 + 128 weights in the input and output layer.

To test the performance of this CPG network for the simulation of infrared spectra, the 32 transformation values of the 110 structures of the test set were sent into the CPG network. Each structure found a neuron in the input level corresponding to an address in the output layer from which a simulated spectrum could be retrieved. This simulated infrared spectrum was compared with the experimental infrared spectrum as contained in the SpecInfo datafile. The correlation coefficients *r* (Bravais-Pearson[34]) between the experimental and the simulated absorbances of all spectra were calculated. Figure 9 shows the distribution of *r* over the test set.

All 110 test objects were sorted according to decreasing value of *r*. Figure 10 shows these results. Four of the simulated spectra are discussed in more detail: 1-phenyl-1-hexene, **4** ($r = 0.94$) as an example for a good simulation, and 1-bromo-3-phenylbutane, **5** ($r = 0.79$), phenylcyclohexane, **6** ($r = 0.58$), and 3-chloro-1-phenyl-2-propen-1-one, **7** ($r = 0.42$) as examples for simulations with medium and low values of the correlation coefficient.

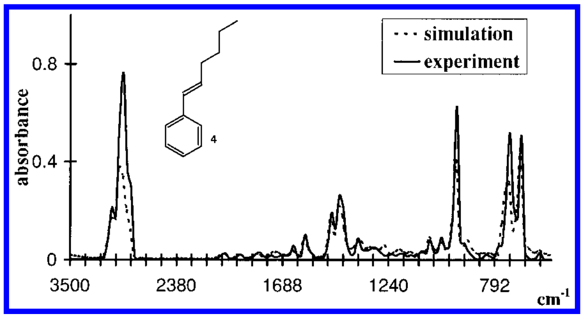


Figure 11. Simulated and experimental infrared spectrum of 1-phenyl-1-hexene, **4**.

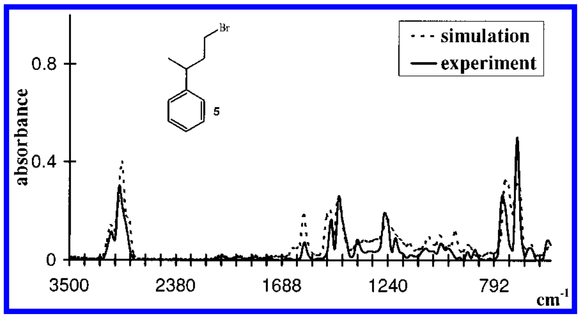


Figure 12. Simulated and experimental infrared spectrum of 1-bromo-3-phenylbutane, **5**.

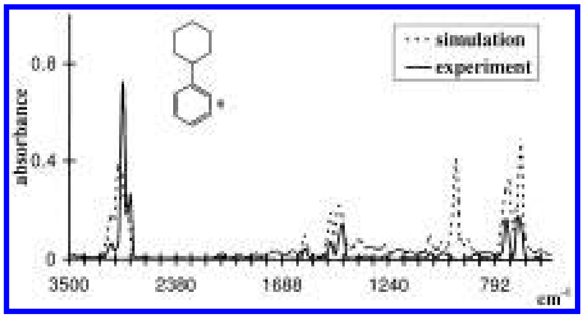


Figure 13. Simulated and experimental infrared spectrum of phenylcyclohexane, **6**.

Figure 11 shows a comparison of the simulated with the experimental infrared spectrum of 1-phenyl-1-hexene, **4**.

Many peaks are excellently reproduced in the simulation. Also the signals in the fingerprint region are very well simulated. This shows in particular the merits of this simulation method. For the simulation quality here it is unimportant how complex the signal causing vibrations are because this method recognizes the shape of the whole spectrum. Neural networks are learning from examples and do not require any initial physical functions as quantum chemical calculations do. Furthermore spectra simulation using this technique is much faster than spectrum calculation by any ab initio or semiempirical method.

In the next example (Figure 12), the signals of the experimental spectrum of **5** are well reproduced in the simulated spectrum, albeit with some discrepancies in the intensities.

Figure 13 shows the results obtained for phenylcyclohexane, **6**. Although the *r*-value is quite low (0.58), most of the peak positions and the spectrum shapes of the experimental spectrum and of the simulated spectrum are rather similar.

However, the simulated spectrum of **6** (Figure 13) shows an extra band at 952 cm^{-1}. This is caused by the training molecules **8**−**11** (Chart 7) that had been assigned to the neuron from which the simulated spectrum has been retrieved from. These molecules contain unsaturated hydrocarbon substituents which normally show absorption bands in this region.

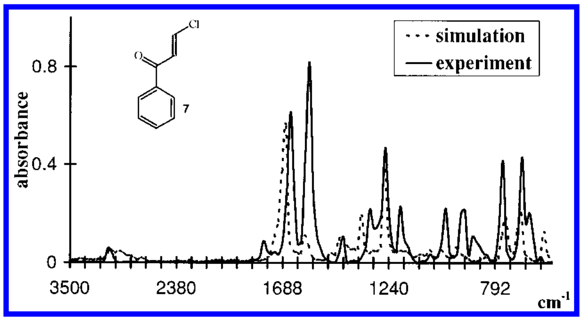
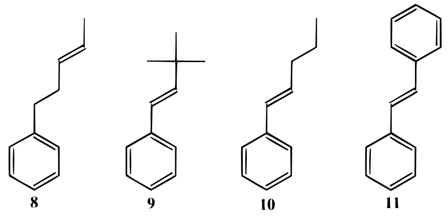


Figure 14. Simulated and experimental infrared spectrum of (3-chloro-1-phenyl-2-propen-1-one) **7**.

Chart 7



An example with a very low *r*-value (0.42) is shown in Figure 14. Comparing the simulated and the experimental spectrum it can be observed that some of the peak positions are slightly shifted. Furthermore some bands show discrepancies in their absolute intensities. The analytical chemist puts more emphasis on relative intensities and on the shapes of spectrum regions. Under this aspect also this example shows the potential of this spectrum simulation method.

CONCLUSIONS

The 3D-MoRSE code as a molecular transform developed here shows great potential for the representation of molecular structures. First, the number of values is independent of the size of the molecule and thus allows the study of datasets of great structural variety. Secondly, the number of these values can be changed and thus the resolution in the representation of a molecular structure can be scaled. Different atomic properties such as atomic number, mass, partial charge, polarizability, etc. can be considered providing great flexibility in the representation of molecules. And, last but not least, the three-dimensional structure of molecules is taken into account. This opens the way for the consideration of the 3D structure of molecules in investigations of the relationships between the structure and physical, chemical, and biological properties. With the recent availability of automatic 3D structure generators [5,8] the development of this representation of the 3D structure is a timely achievement. It can safely be expected that QSAR studies can in many cases be put on a higher level of insight and modelling power.

The classification of molecules into dopamine D1 and D2 agonists as well as of steroids into low, medium, or high affinity to the corticosteroid binding globulin are illustrative examples.

The new molecular representation opens exciting dimensions for the study of the relationships between the structure of organic molecules and their infrared spectra. We are actively pursuing this potential.

It should also be mentioned that the studies reported here point to the efficient capabilities of a counterpropagation neural network to implicitly model complex relationships.

ACKNOWLEDGMENT

Support of this work by the Bundesminister für Bildung und Forschung of the FRG is gratefully acknowledged. We thank the Fonds der Chemischen Industrie for financial assistance. We thank Prof. Jure Zupan and Dr. Marjana Novič, both of National Institute of Chemistry, Ljubljana, Slovenia, for providing us with the counterpropagation network code and many stimulating discussions. We thank Chemical Concepts, Weinheim, FRG for providing us with the SpecInfo database and MDL Information Systems, Inc. for giving access to the MDDR database.

REFERENCES AND NOTES

(1) Verloop, A.; Hoogenstaaten, W.; Tipker, J. *Drug Design, VII*; Ariëns, E. J., Ed.; Academic Press: New York, 1976; pp 165−207. Verloop, A. *The STERIMOL Approach to Drug Design*; Marcel Dekker: New York, 1987.

(2) Sasaki, S.; Abe, H.; Ouki, T.; Sakamoto, M.; Ochiai, S. Automated Structure Elucidation of Several Kinds of Aliphatic and Alicyclic Compounds. *Anal. Chem.* **1968**, *40*, 2220−2223. Funatsu, K.; Susuta, Y.; Sasaki, S. Introduction of Two-Dimensional NMR Spectral Information to an Automated Structure Elucidation System, CHEMICS. Utilization of 2D-INADEQUATE Information. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 6−11.

(3) Dubois, J. E.; Mathieu, G.; Peguet, P.; Panaye, A.; Doucet, J. P. Simulation of Infrared Spectra: An Infrared Spectral Simulation Program (SIRS) Which Uses DARC Topological Substructures. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 290−302.

(4) Huixiao, H.; Xinquan, X. ESSESA: An Expert System for Elucidation of Structures from Spectra. 1. Knowledge Base of Infrared Spectra and Analysis and Interpretation Programs. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 203−210.

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

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

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

(8) 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.

(9) This dataset of 3D-structures has been deposited on the file server of the National Cancer Institute, Bethesda, MD. It is publicly available and can be accessed via anonymous ftp helix.nih.gov.; cd ncidata/3D; get ciopen3d.mol.z.

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

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

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

(13) 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.

(14) Hutchings, M. G.; Gasteiger, J. Residual Electronegativity−An Empirical Quantification of Polar Influences and Its Application to the Proton Affinity of Amines. *Tetrahedron Lett.* **1983**, *24*, 2541−2544.

(15) Gasteiger, J.; Hutchings, M. G. Quantification of Effective Polarisability. Applications to Studies of X-ray Photoelectron Spectroscopy and Alkylamine Protonation. *J. Chem. Soc. Perkin Trans. 2*, **1984**, 559−564.

(16) Varmuza, K. *Pattern Recognition in Chemistry*; Springer Verlag: Berlin, 1980.

(17) Hecht-Nielsen, R. Counterpropagation networks. *Applied Optics* **1987**, *26*, 4979−4984.

(18) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists−An Introduction*; VCH: Weinheim, 1993; p 55.

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

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

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

(22) Kohonen, T. Self-Organized Formation of Topologically Correct Feature Maps. *Biol. Cyb.* **1982**, *43*, 59−69.

(23) Bremser, W. Structure Elucidation and Artifical Intelligence. *Angew. Chem* **1988**, *100*, 252−265; *Angew. Chem., Int. Ed. Engl.* **1988**, *27*, 247−260.

(24) Simon, V.; Gasteiger, J.; Zupan, J. A Combined Application of Two Different Neural Network Types for the Prediction of Chemical Reactivity. *J. Am. Chem. Soc.* **1993**, *115*, 9148−9159.

(25) MDL Drug Data Report 94.2, available from MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.

(26) Dunn, J. F.; Nisula, B. C.; Rodbard, D. Transport of Steroid Hormones: Binding of 21 Endogenous Steroids to Both Testosterone-Binding Globulin and Corticosteroid-Binding Globulin in Human Plasma. *J. Crin. Endocrin. Metab.* **1981**, *53*, 58−68.

(27) Mickelson, K. E.; Forsthoefel, J.; Westphal, U. Steroid Protein Interactions, Human Corticosteroid Binding Globulin: Some Physicochemical Properties and Binding Specifity. *Biochemistry* **1981**, *20*, 6211−6218.

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

(29) Good, A. C.; So, S.; Richards, W. G. Structure−Activity Relationships from Molecular Similarity Matrices. *J. Med. Chem.* **1993**, *26*, 990−996.

(30) Sybyl; Tripos Associates Inc.; St. Louis, MO, U.S.A.

(31) Automated Similarity Package; Oxford Molecular Ltd.: Oxford, UK.

(32) 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.

(33) Affolter, C.; Baumann, K.; Clerc, J. T.; Schriber, H.; Pretsch, E. Automatic Interpretation of Infrared Spectra. *Mikrochim. Acta*, in press.

(34) Bortz, J. *Statistik für Sozialwissenschaftler*; Springer: Berlin, 1989; S. 251.

(35) Much work has been devoted over the last two decades to the use of spectral data in automatic structure elucidation systems such as CHEMICS.[2]

CI950164C