Autocorrelation of Molecular Surface Properties for Modeling *Corticosteroid Binding Globulin* and Cytosolic *Ah* Receptor Activity by Neural Networks

Markus Wagener, Jens Sadowski, and Johann Gasteiger*

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

Received January 23, 1995[⊗]

Abstract: Molecular surface properties such as the electrostatic or the hydrophobicity potential were condensed into an autocorrelation descriptor. A vector of these autocorrelation descriptors based on the molecular electrostatic potential was successfully applied to modeling the affinities of a set of 31 steroid molecules binding to the *corticosteroid binding globulin* (CBG) receptor by using a combination of a Kohonen and a feedforward neural network. Similarly, an autocorrelation vector derived from the hydrophobicity potential was used to model the binding constant of a set of 78 polyhalogenated aromatic compounds to the cytosolic *Ah* receptor. The models found have a high predictive ability as established by cross-validation.

Introduction

It is generally accepted that receptor and substrate molecules recognize each other at their molecular surfaces. Therefore, the binding strength of a receptor−drug complex depends on the shape of the substrate surface and on the distribution of certain properties on this surface. Any method attempting to model biological activity should take into account this information and try to correlate it to biological activity. In general, the problem can be approached in two steps: First, several properties such as electrostatic potential, hydrogen bonding energy, or hydrophobicity potential are calculated at distinct points on the molecular surface or in the space surrounding the molecules. Second, these properties are correlated to the activity values using statistical methods or neural networks. Two major problems arise:

First, there is a large number of independent variables. For a steroid molecule, e.g., one obtains approximately 3500 points on the molecular surface when using a point density of 10 points per Å^2. Several approaches to overcome this problem have been published which are based on descriptors calculated from shape and potential differences[1,2] or similarity measures[3] or use special statistical techniques such as partial least squares as in the CoMFA method.[4]

Second, the independent variables are determined by the absolute positions of the points in space and therefore strongly depend on the orientation of the different molecules. Thus, one needs a method for finding the optimum alignment of the molecules of a dataset. Approaches for solving this problem were published in ref 5−7.

We propose here a 3D-QSAR descriptor that uses spatial autocorrelation coefficients for transforming the independent variables and for overcoming the alignment problem. In the second part of the paper, we present some applications of the novel descriptors using artificial neural networks for modeling biological activity.

The first application deals with a dataset of 31 steroids that bind to the *corticosteroid binding globulin* (CBG) receptor. This dataset has already been investigated by several research groups. It formed the basis for the introduction of the widely used Comparative Molecular Field Analysis (CoMFA) method.[4] In addition, molecular similarity calculations were performed for these molecules and analyzed by neural networks and statistical methods.[3]

The second dataset addresses the problem of modeling the toxicity of polyhalogenated aromatic compounds that include the highly toxic chlorinated and brominated dibenzo-*p*-dioxins, as well as chlorinated dibenzofurans and biphenyls. These compounds bind to the cytosolic *Ah* receptor.[8−10]

Methods

Spatial Autocorrelation. It is often necessary to consider the spatial distribution of some quality or phenomenon in an area consisting of several distinct regions. One question that arises then is whether the presence of that quality in one region makes its presence in a neighboring region more or less likely. If there is such an interdependence, the data exhibit spatial autocorrelation.[11] The concept of spatial autocorrelation is mainly applied to problems of geography, economics, ecology, or meteorology. A chemical example is the analysis of the amino acid sequence along a peptide backbone.[12] Statisticians have developed a number of measures for quantifying spatial autocorrelation.[11] One example is Moran's coefficient *I*:

$$I = \frac{n}{2L} \frac{\sum_{ij} \delta_{ij}(p_i - \bar{p})(p_j - \bar{p})}{\sum_i (p_i - \bar{p})^2} \qquad (1)$$

[⊗] Abstract published in *Advance ACS Abstracts*, July 1, 1995.

[1] Hopfinger, A. J. *J. Am. Chem. Soc.* **1980**, *102*, 7196−7206.

[2] Hopfinger, A. J. *J. Med. Chem.* **1983**, *26*, 990−996.

[3] Good, A. C.; So, S.; Richards, W. G. *J. Med. Chem.* **1993**, *36*, 433−438.

[4] Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

[5] Kearsley, S. K.; Smith, G. M. *Tetrahedron Comput. Methodol.* **1990**, *3*, 615.

[6] Good, A. C.; Hodgkin, E. E.; Richards, W. G. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 188−191.

[7] Manaut, M.; Sanz, F.; Jose, J.; Milesi, M. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 371−380.

[8] Safe, S. *Annu. Rev. Pharmacol. Toxicol.* **1986**, *26*, 371−399.

[9] Safe, S. *Crit. Rev. Toxicol.* **1990**, *21*, 51−88.

[10] Bandiera, S.; Safe, S.; Okey, A. B. *Chem.-Biol. Interact.* **1982**, *39*, 259−277.

[11] Cliff, A. D.; Ord, J. K. *Spatial Autocorrelation*; Pion Limited: London, 1973.

[12] Van Heel, M. *J. Mol. Biol.* **1991**, *220*, 877−887.

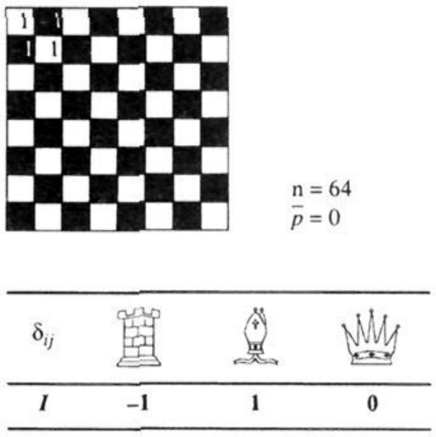
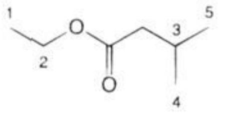


Figure 1. Values of Moran's I coefficient (eq 1) as a measure of autocorrelation of the colors on a chess board for the drawing rules of the rook, the bishop, and the queen.

Chart 1



where n is the total number of data points, $\{\delta_{ij}\}$ is a connection matrix with $\delta_{ij} = 1$ for neighboring points i, j and otherwise $\delta_{ij} = 0$, p_i is the property value of point i, $\bar{p}$ is the mean property value, and L is the total number of connections given by $\{\delta_{ij}\}$.

Equation 1 is a quantitative measure of the probability that at neighboring data points similar property values can be found. The meaning of I is illustrated by a simple example (Figure 1). The black and white colors of the fields of a chessboard are coded by -1 and 1, respectively. In the lower part of Figure 1, different values for I are shown as obtained by eq 1 using different neighborhood descriptions $\{\delta_{ij}\}$. These connection matrices refer to the directly adjacent fields that can be reached by a rook, a bishop, or the queen when moving only one field. The rook, which can move only horizontally and vertically, will always arrive at a field of opposite color to the one it started from—when only a move by one field is allowed. Thus, for the rook an I of -1 is obtained, a strictly negative correlation. A bishop always moves in a diagonal manner to a field of the same color. For the bishop, one gets an I of $+1$, a strictly positive correlation. The queen may move to black fields as well as to white ones and thus no correlation will be observed ($I = 0$).

Applications of Spatial Autocorrelation in Molecular Modeling. A number of applications of autocorrelation in molecular modeling and QSAR have been published. Moreau and Broto[13] first applied an autocorrelation function to the topology of molecular structures:

$$A(d) = \sum_{ij} p_i p_j \tag{2}$$

$A(d)$ is the autocorrelation coefficient referring to atom pairs i, j which are separated by d bonds. p_i is an atomic property, e.g. the partial charge on atom i. Thus, one obtains a series of coefficients for different topological distances d, a so-called autocorrelation vector. In this case, the entries of the neighborhood matrix $\{\delta_{ij}\}$ of eq 1 are equal to 1 if the topological distance between atoms i and j is equal to d. This is illustrated by an example. The molecule shown in Chart 1, 3-methylbutyric acid ethyl ester, has three pairs of atoms which are separated by five bonds: C_1-C_3, C_2-C_4, and C_2-C_5. Thus, the corresponding autocorrelation for the topological distance five computes to

$$A(5) = p_1p_3 + p_2p_4 + p_2p_5 \tag{3}$$

for a given atomic property p.

The autocorrelation vectors exhibit some useful qualities. First, a substantial reduction of data can be achieved by restricting d. Second, the autocorrelation coefficients are independent of the original atom numbering—they are canonical. Third, the length of the vector is independent of the size of the molecule. Finally, the vectors represent the degree of similarity between molecules. The topological autocorrelation vectors were used as molecular descriptors in QSAR studies.[14,15]

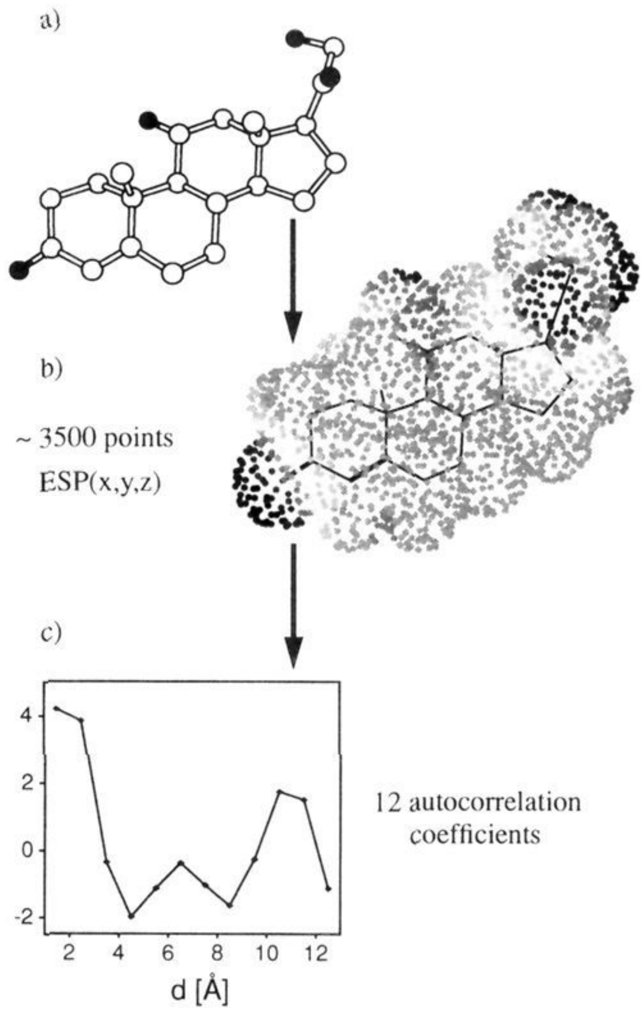


Figure 2. Calculation of the spatial autocorrelation vector for the electrostatic potential on the surface of a steroid molecule: (a) the 3D model, (b) the point representation of the electrostatic potential on the van der Waals surface, and (c) the autocorrelation vector.

Broto and Moreau[16] extended this concept to 3D molecular models by replacing the topological distance d in eq 2 by the interatomic distance in 3D space and thus correlating atomic properties on the three-dimensional molecular skeleton. Two other 3D autocorrelation studies appeared more recently, characterizing the results of molecular dynamics simulations[17] and correlating values of potential on a CoMFA-like grid around molecules.[18]

Spatial Autocorrelation of Molecular Surface Properties. In the present study, a 3D descriptor is introduced that is based on the autocorrelation of properties at distinct points on the molecular surface. The points are randomly distributed according to a preset point density in order to model a continuous surface and to avoid artefacts. The distances between surface points are sorted into preset intervals (d_{lower}, d_{upper}). The autocorrelation coefficient $A(d_{\text{lower}}, d_{\text{upper}})$ is obtained by summation of the products of property values p at points i,j having a distance d belonging to the distance interval (d_{lower}, d_{upper}) and by weighting the sum by the total number L of distances in the interval:

$$A(d_{\text{lower}}, d_{\text{upper}}) = \frac{1}{L}\sum_{ij} p_i p_j \quad (d_{\text{lower}} < d_{ij} < d_{\text{upper}}) \tag{4}$$

For a series of distance intervals with different lower and upper bounds d_{lower} and d_{upper}, a vector of autocorrelation coefficients is obtained. This vector is a compressed expression of the distribution of property p on the molecular surface. Figure 2 illustrates the complete sequence for the calculation of the autocorrelation vector. Starting from a 3D

[13] Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 359–360.

[14] Moreau, G.; Broto, P. *Nouv. J. Chim.* **1980**, *4*, 757–764.

[15] Zakarya, D.; Tiyal, F.; Chastrette, M. *J. Phys. Org. Chem.* **1993**, *6*, 574–582.

[16] Broto, P.; Moreau, G.; Vandycke, C. *Eur. J. Med. Chem. Chim. Ther.* **1984**, *19*, 66–70.

[17] Grassy, G.; Lahana, R. In: *Trends in QSAR and Molecular Modelling '92. Proceedings of the 9th Symposium on Structure-Activity-Relationships*: *QSAR and Molecular Modelling*; Wermuth, C. G., Ed.; ESCOM: Leiden, 1993; pp 216–219.

[18] Clementi, S.; Cruciani, G.; Riganelli, D.; Valigi, R.; Costantino, G.; Baroni, M.; Wold, S. *Pharm. Pharmacol. Lett.* **1993**, *3*, 5–8.

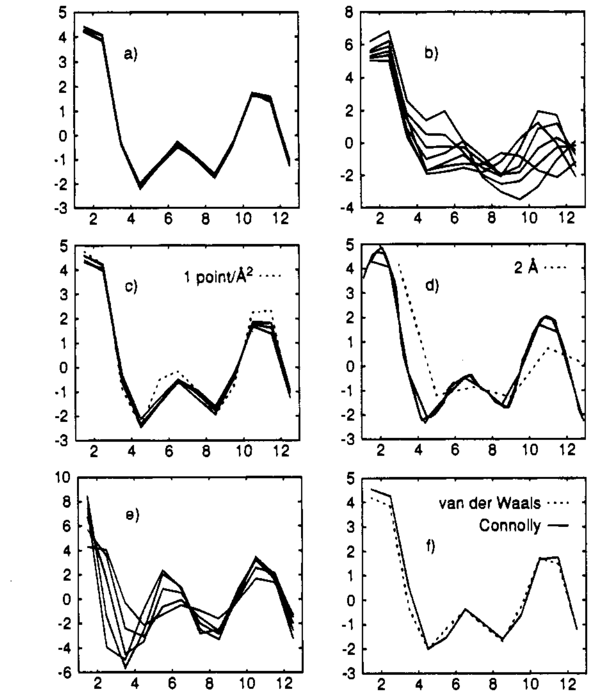
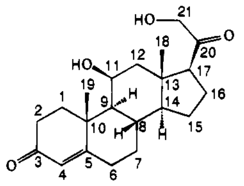


Figure 3. The dependence of the autocorrelation vector of corticosterone on six different parameters of the calculation scheme (eq 4): (a) six different spatial orientations; (b) seven different conformations of the side chain at position 17; (c) five different point densities; (d) four different distance intervals d_{ij}; (e) five different sets of atomic radii; (f) comparison of the Connolly surface with the van der Waals surface. See text for the values of the different parameters and their default values.

Chart 2



model of a molecule and its partial atomic charges, the electrostatic potential or another appropriate property is calculated for points on the molecular surface. For the steroid molecule shown in the top of Figure 2, about 3500 points are obtained which are characterized by their Cartesian coordinates and the value of the electrostatic potential. After applying the autocorrelation function, the autocorrelation vector is obtained. Considering distances from 1 to 13 Å with a step width of 1 Å, twelve autocorrelation coefficients are calculated and displayed at the centers of the distance intervals at 1.5, 2.5, etc. in the lower part of Figure 2. This transformation produces a unique fingerprint of each molecule under consideration.

Properties of the Surface Autocorrelation Vector. The autocorrelation vectors exhibit some interesting properties. First, they are unique for a given molecular geometry. Second, they are invariant to translation and rotation since only spatial distances instead of Cartesian coordinates are used. Third, a substantial reduction of the input information can be achieved. A disadvantage of the condensed description of the molecular surface by an autocorrelation vector might be that the original information cannot be reconstructed. Thus, conclusions on the nature of the pharmacophore are not evident. In the following, the dependence of the autocorrelation vector on six parameters that can be changed in the calculation scheme is presented (Figure 3). All calculations refer to corticosterone which is shown in Chart 2 with the IUPAC numbering of the steroid skeleton. The following default values were used for those parameters remaining unchanged: A point density of 10 points/Å² on the surface, a distance interval of 1 Å, 100% of the van der Waals radii, and the van der Waals surface.

(a) Spatial Orientation of the Molecule. Figure 3a shows the autocorrelation vectors obtained for six different spatial orientations. There is no significant difference between them. The observed small deviations are due to the different locations of the points on the molecular surface. Thus, the autocorrelation vector is translationally and rotationally invariant as stated above.

(b) Conformational Flexibility. Figure 3b shows the autocorrelation vectors of seven different conformations of the side chain at position 17 (see Chart 2 for the numbering). The $O{=}C_{20}{-}C_{17}{-}C_{16}$ torsional angle was varied from 0° to −180° in steps of −30°. A substantial variance in the autocorrelation vectors can be observed. Thus, the autocorrelation vector is sensitive to changes in the conformation.

(c) Point Density on the Molecular Surface. In Figure 3c, the autocorrelation vectors for five different point densities (1, 5, 10, 15, and 20 points/Å²) are compared. Only the vector obtained for a density of 1 point/Å² (dotted line) differs significantly from the average autocorrelation vector. Thus, for point densities equal to or greater than 5 points/Å² the continuous surface can be modeled with good accuracy.

(d) Distance Intervals. Figure 3d shows the vectors obtained for four different distance intervals (0.25, 0.5, 1.0, and 2.0 Å). Larger intervals tend to flatten the vectors. The vector obtained for an interval of 2 Å (dotted line) differs significantly from the average. Thus, intervals equal to or less than 1 Å should be used.

(e) Atomic Radii. Figure 3e shows the autocorrelation vectors obtained for five different sets of atomic radii (60, 70, 80, 90, and 100% of the van der Waals radii). The shapes of the vectors depend strongly on this parameter.

(f) Surface Type. In Figure 3f, the vectors for Connolly's solvent accessible surface[19] with a solvent radius of 2.0 Å and for the van der Waals surface are compared. There is no significant difference. Thus, the simpler van der Waals surface can be used with good accuracy to model the solvent accessible surface.

On the basis of this comparison the following values of the parameters for the calculation of the autocorrelation vectors were chosen for further investigations: a point density of 10 points/Å² on the van der Waals surface, a distance interval of 1 Å, and 100% of the van der Waals radii.

Results and Discussion

Dataset of 31 Steroids Binding to the Corticosteroid Binding Globulin (CBG) Receptor. The first example for applying the surface autocorrelation vector is presented using a well-known dataset: 31 steroids binding to the *corticosteroid binding globulin* (CBG). This dataset was first compiled by Cramer et al.[4] for the presentation of the CoMFA method. A subset of the data containing the first 21 steroids is distributed with the Sybyl program package.[20] Richards et al.[3] used the same data for similarity calculations; it is distributed as an example file with the ASP program.[21] Comparison of all these printed or computer-coded versions of the dataset shows several discrepancies in structure coding. Thus, we returned to the original literature[22,23] and carefully recompiled the dataset (Chart 3).

Comparison of the original data to those in the publications[3,4] and to the distributed datasets[20,21] indicates a number of errors in coding the topology and/or the stereochemistry of the molecules in all four secondary sources. In detail, these were seven errors in Cramer et al.[4] (**5**, **13**, **14**, **15**, **16**, **21**, **28**), one error in the Sybyl dataset[20] (**2**), six errors in Richards et al.[3] (**5**, **14**, **16**, **21**, **28**, **31**), and six errors in the ASP dataset[21] (**2**, **5**,

[19] Connolly, M. L. *J. Appl. Crystallogr.* **1983**, *16*, 548−558.

[20] Sybyl, Tripos Associates Inc.: St. Louis, MO.

[21] Automated Similarity Package, Oxford Molecular Ltd: Oxford, UK.

[22] Dunn, J. F.; Nisula, B. C.; Rodbard, D. *J. Crin. Endocrin. Metab.* **1981**, *53*, 58−68.

[23] Mickelson, K. E.; Forsthoefel, J.; Westphal, U. *Biochemistry* **1981**, *20*, 6211−6218.

Chart 3

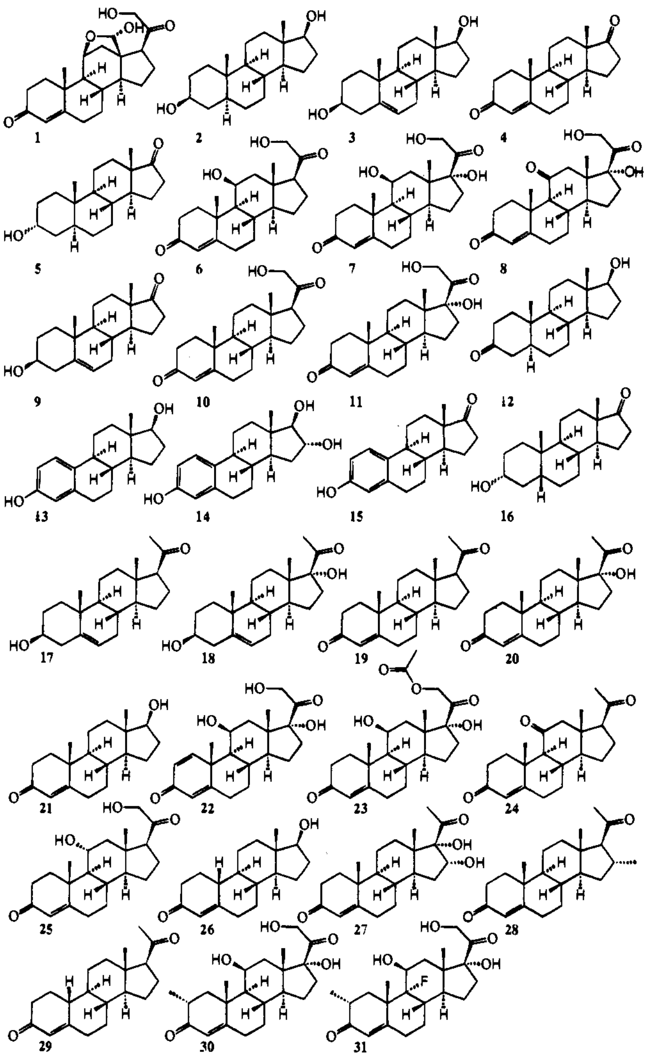


Table 1. CBG Binding Affinity Data from Ref 4

compd	CBG affinity (p*K*)	activity class[a]	compd	CBG affinity (p*K*)	activity class[a]
1	−6.279	2	**17**	−5.225	3
2	−5.000	3	**18**	−5.000	3
3	−5.000	3	**19**	−7.380	1
4	−5.763	3	**20**	−7.740	1
5	−5.613	3	**21**	−6.724	2
6	−7.881	1	**22**	−7.512	1
7	−7.881	1	**23**	−7.553	1
8	−6.892	2	**24**	−6.779	2
9	−5.000	3	**25**	−7.200	1
10	−7.653	1	**26**	−6.144	2
11	−7.881	1	**27**	−6.247	2
12	−5.919	2	**28**	−7.120	2
13	−5.000	3	**29**	−6.817	2
14	−5.000	3	**30**	−7.688	1
15	−5.000	3	**31**	−5.797	2
16	−5.225	3			

[a] 1, high; 2, intermediate; 3, low; this classification was obtained by dividing the dataset into three classes of comparable site.

14, 16, 21, 28), respectively. The correct dataset can be obtained upon request from the authors in the form of an MDL MOLFILE.

The corresponding CBG binding affinities from ref 4 are shown as −log *K* values in Table 1. It should be emphasized that giving the p*K* values of the CBG affinity with five digits precision indicates an accuracy in the biological data that is not supported by experiment. We have intentionally used the data as given in ref 4 to provide a basis for the direct comparison of our results with those of other investigators.

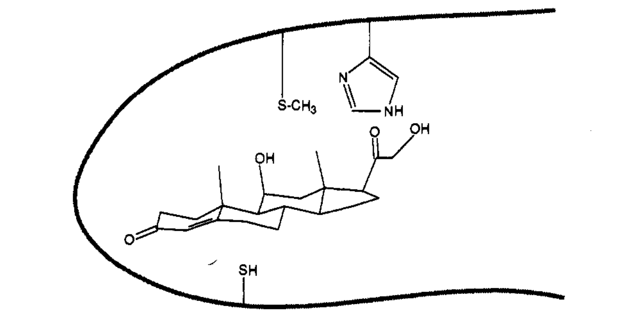


Figure 4. Visualization of the CBG binding site for binding corticosterone.[26]

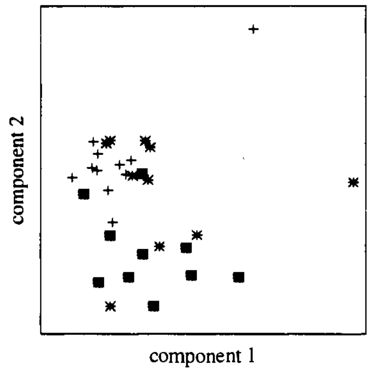


Figure 5. Principal components plot of the steroid data set: squares, high activity; asterisks, intermediate activity; crosses, low activity.

Model Building. 3D models of the structures were obtained by using the 3D structure generator Corina.[24,25] A receptor model for the CBG binding site[26] as shown in Figure 4 suggests a histidine residue above the plane of the D ring. Thus, the orientation of the C_{17} side chain was manually adjusted to an $O{=}C_{20}{-}C_{17}{-}C_{16}$ torsional angle of −60° in order to maximize the interaction of the C_{20} carbonyl function with the N−H hydrogen donor site of the histidine residue.[26] Partial atomic charges were calculated by the PEOE method[27] and its extension to conjugated systems.[28] Points were randomly distributed on the van der Waals surface and the electrostatic potential at each point was calculated by a classical Coulomb approach using a unit positive point charge and the partial charges on all atoms of the molecule. Autocorrelation vectors were calculated for each molecule for distance intervals of 1 Å from 1 to 13 Å by using the electrostatic potential as property on the van der Waals surface and a point density of 10 points/Å^2.

Classification Using PCA and Kohonen Networks. As a first step toward a model for the binding affinities of the 31 steroids, the suitability of the surface property autocorrelation vectors as QSAR descriptors was investigated by a principal component analysis (PCA) to search for a relationship between autocorrelation vectors and biological activity. PCA projects multivariate data sets from a given problem space into a lower dimensionality, it thus can be used to produce two-dimensional plots. In Figure 5, a plot of the first two principal components is shown. Squares, asterisks, and crosses mark compounds with high, intermediate, and low activity, respectively. This linear

[24] Sadowski, J.; Gasteiger, J. *Chem. Rev.* **1993**, *93*, 2567−2581.

[25] Sadowski, J.; Gasteiger, J.; Klebe, G. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.

[26] Defaye, G.; Basset, M.; Monnier, N.; Chambaz, E. M. *Biochim. Biophys. Acta* **1980**, *623*, 280−294.

[27] Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219−3228.

[28] Gasteiger, J.; Saller, H. *Angew. Chem.* **1985**, *97*, 699−701; *Angew. Chem., Int. Ed. Engl.* **1985**, *24*, 687−689.

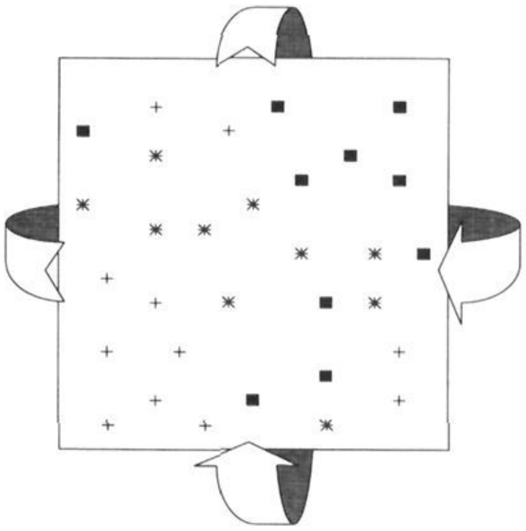


Figure 6. Kohonen map of the steroid data set: squares, high activity; asterisks, intermediate activity; crosses, low activity. The Kohonen network has a toroidal topology. Thus, the upper and lower neurons, as well as those at the left- and right-hand side, are directly connected as indicated by the arrows.

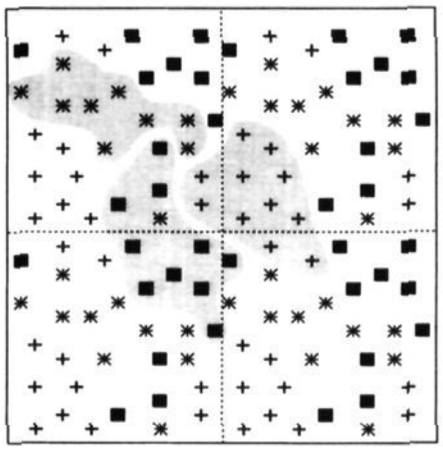


Figure 7. The 4-fold replication of the Kohonen map of Figure 6. The three different clusters of compounds with high, intermediate, and low activity are highlighted by shaded areas.

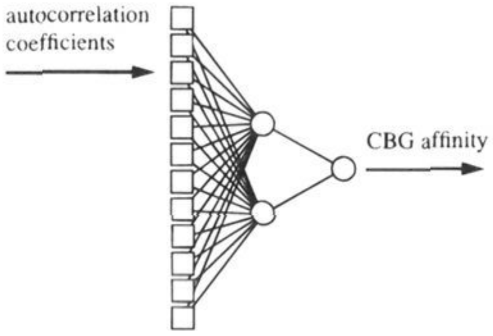


Figure 8. Multilayer neural network topology.

projection cannot sufficiently separate the three activity classes.

In recent years, neural networks have gained prominence for finding nonlinear relationships. A textbook is given in ref 29 and a review on the application of neural networks in chemistry in ref 30. In particular, Kohonen networks[31] can be used for the projection of multidimensional data into two-dimensional plots.[29,30] Thus, it had been shown that Kohonen networks can be successfully employed for the projection of reactivity data from a seven-dimensional space into two dimensions.[32]

A Kohonen network was used for the nonlinear mapping of the data from the twelve-dimensional space spanned by the autocorrelation vectors into two dimensions. Training of a 15 × 15 network with the dataset of 31 steroids on a Sun Sparc 10/512 took 71 s. Figure 6 shows the resulting network. Squares, asterisks, and crosses again mark compounds with high, intermediate, and low activity, respectively. A Kohonen network with a toroidal topology was used. Thus, the upper neurons are connected to the lower ones, and the neurons at the left-hand side are connected to those at the right-hand side, as indicated by the arrows in Figure 6.

It has been shown that the toroidal topology can nicely be indicated by replicating a Kohonen map several times and putting these maps together like tiles.[29,33] Figure 7 shows a 4-fold replication of Figure 6. Here, the neighborhood relationships become much more clear and a significant separation of the data according to the activity classes can be seen as highlighted by the three shaded areas. Only one misclassification occurs: a point referring to a compound with intermediate activity (asterisk) is surrounded by highly active compounds (squares). This is compound **31**, the only steroid being substituted at position 9; in this case with a fluorine atom.

The good classification of the data by a Kohonen map according to their binding affinity demonstrates the suitability of the autocorrelation vector based on the molecular electrostatic potential for modeling biological activity. In other words, if the variables contained in the autocorrelation vector are quite successful in separating the steroids into three activity classes, an investigation of their usefulness in modeling the real values of biological activity quantitatively (see Table 1) seems worthwhile. Comparison of the results of the principal component analysis with those of the Kohonen network indicates the nonlinear relationship between the autocorrelation vector and the biological activity. Thus, it seemed necessary to again use a nonlinear method for modeling the biological activity data. Such a method is provided by multilayer neural networks trained by the backpropagation algorithm.[29,34]

Predicting Biological Activity Using a Multilayer Neural Network. It should be emphasized that a Kohonen network is based upon an unsupervised learning method: the property in question, in our case biological activity, is **not** used in the analysis of the data. A multilayer network trained by the backpropagation algorithm,[34] on the other hand, is based on supervised learning: the values of biological activity are used in deriving a model for expressing the relationship between the independent variables (autocorrelation vector) and the biological activity.

A feedforward multilayer neural network was used to obtain a predictive model of the biological activity of the 31 steroid molecules based on their autocorrelation vectors. Figure 8 shows the topology of the network used. It has twelve input units, two neurons in the hidden layer, and one output neuron. The input units correspond to the autocorrelation vector and the output neuron to the biological activity. The multilayer neural network was simulated using a standard program package.[35]

The network was trained with the data of all 31 steroids following the "backpropagation with momentum" procedure until the error converged. The training on a Sun Sparc 10/512 required 41 s. The trained network was then used to predict the biological activities of the 31 steroid molecules. Figure 9 shows the ability of the trained network to reproduce the data used for training. The experimental pK values are plotted

[29] Zupan, J.; Gasteiger, J. *Neural Networks for Chemists—An Introduction*; VCH: Weinheim, 1993.

[30] Gasteiger, J.; Zupan, J. *Angew. Chem.* **1993**, *105*, 510–536; *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503–527.

[31] Kohonen, T. *Self-Organization and Associative Memory*, 3rd ed.; Springer: Berlin, 1989.

[32] Simon, V.; Gasteiger, J.; Zupan, J. *J. Am. Chem. Soc.* **1993**, *115*, 9148–9159.

[33] Gasteiger, J.; Li, X.; Uschold, A. *J. Mol. Graphics* **1994**, *12*, 90–97.

[34] Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*; MIT Press: Cambridge, MA, 1986.

[35] SNNS: Stuttgart Neural Network Simulator; Version 3.0, University of Stuttgart, 1993.

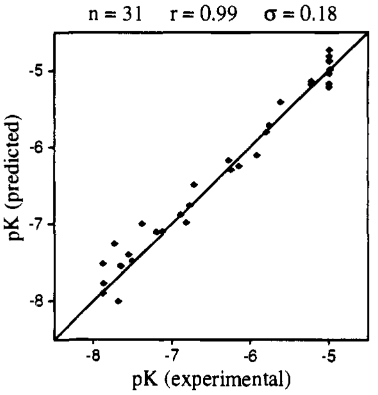


Figure 9. Plot of the experimental p*K* values against the p*K* values reproduced by the trained network.

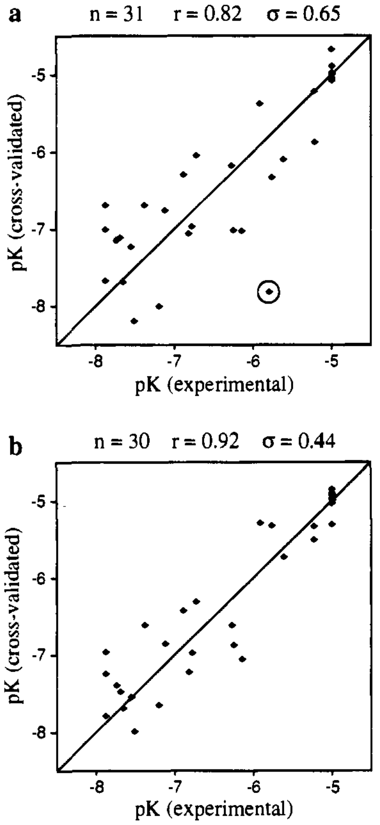


Figure 10. Plot of the experimental p*K* values against the cross-validated values: (a) entire dataset (molecule **31** marked by a circle); (b) dataset of 30 molecules without **31**.

against the predicted values. A rather high correlation (r = 0.99) with a low standard deviation (σ = 0.18) is obtained.

In order to estimate the predictive power of the model, cross-validation following the leave-one-out scheme was performed. In 31 independent experiments, the data of 30 steroids were used to train the network. The trained network was then used to predict the biological activity of the 31st molecule. This procedure was repeated 31 times, each time leaving out one molecule and then predicting its activity from the model obtained with the other 30 steroids. Figure 10 shows two plots of the experimental p*K* values against the cross-validated values. Figure 10a shows the results of the cross-validation procedure for the entire dataset. A significantly lower correlation (r = 0.82, σ = 0.65) with a cross-validated r^2 of 0.63 is obtained—a rather low predictivity. One outlier with an outstanding deviation can be identified: molecule **31** (marked by a circle). This is the same molecule which was already misclassified by the Kohonen net, the only steroid of the dataset having a substituent at position 9, in this case a fluorine atom. In two other studies on these data, molecule **31** also was found to lead to problems.[3,4]

Chart 4

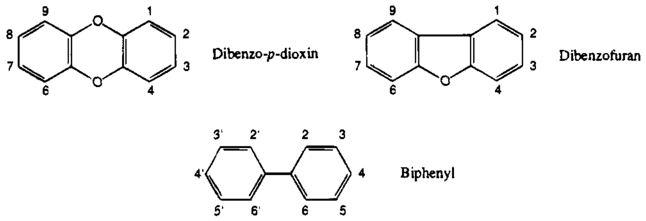


Table 2. Binding Affinities of Polychlorinated and Polybrominated Dibenzo-*p*-dioxins (for the Numbering See Chart 4)

substitution positions	pEC$_{50}$	substitution positions	pEC$_{50}$
2,3,7,8-Cl$_4$	8.000	1-Cl	4.000
1,2,3,7,8-Cl$_5$	7.102	2,3,7,8-Br$_4$	8.824
2,3,6,7-Cl$_4$	6.796	7,8-Cl$_2$-2,3-Br$_2$	8.830
2,3,6-Cl$_3$	6.658	3,7-Cl$_2$-2,8-Br$_2$	9.350
1,2,3,4,7,8-Cl$_6$	6.553	3,7,8-Cl$_3$-2-Br	7.939
1,3,7,8-Cl$_4$	6.102	1,3,7,8,9-Br$_5$	7.032
1,2,4,7,8-Cl$_5$	5.959	1,3,7,8-Br$_4$	8.699
1,2,3,4-Cl$_4$	5.886	1,2,4,7,8-Br$_5$	7.770
2,3,7-Cl$_3$	7.149	1,2,3,7,8-Br$_5$	8.180
2,8-Cl$_2$	5.495	2,3,7-Br$_3$	8.932
1,2,3,4,7-Cl$_5$	5.194	2,7-Br$_2$	7.810
1,2,4-Cl$_3$	4.886	2-Br	6.530
1,2,3,4,6,7,8,9-Cl$_8$	5.000		

Thus, it can be assumed that this structural type cannot be modeled correctly within this dataset. After deleting this molecule from the dataset, the cross-validation test was repeated. Figure 10b shows the result of the cross-validation for the reduced dataset. Now a rather high value for the predictive power is obtained (r = 0.92, σ = 0.44) with a high cross-validated r^2 of 0.84. For comparison, Cramer et al.[4] obtained for the first 21 steroid molecules of the same dataset a CoMFA model with a cross-validated r^2 of 0.66 (applying a random leave-*n*-out scheme). The best PLS model reported by Richards et al.[3] for the same 21 steroids has a cross-validated r^2 of 0.76.

A Second Example: Modeling the Toxicity of Polyhalogenated Aromatic Compounds. In a second example, the affinities of 78 polyhalogenated aromatic compounds for binding to the cytosolic *Ah* receptor were studied. The dataset consisted of 25 chlorinated and brominated dibenzo-*p*-dioxins,[8,9] 39 chlorinated dibenzofurans,[8,9] and 14 chlorinated biphenyls.[10] The *Ah* affinity is a measure for the toxicity of these compounds. Since there is an increasing public interest in avoiding these kinds of toxic compounds, it is highly desirable to have a means for predicting their activity. Consistent with this, already several attempts have been made to model the toxicity of these compounds.[18,36] Chart 4 shows the molecular skeletons of the different structural types and the numbering scheme. Tables 2–4 give the binding affinities of the compounds.

Since these compounds are highly hydrophobic, the hydrophobicity potential was used as molecular surface property. It was calculated for all 78 polyhalogenated aromatic compounds from atomic increments for log P[37] by using a distance-dependent potential function.[38] In analogy to the steroid example, 12 autocorrelation coefficients per molecule were

[36] Waller, C. L.; McKinney, J. D. *J. Med. Chem.* **1992**, *35*, 3660–3666.

[37] Ghose, A. K.; Crippen, G. M. *J. Comp. Chem.* **1986**, *7*, 565–577.

[38] Heiden, W.; Moeckel, G.; Brickmann, J. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 503–514.

Table 3. Binding Affinities of Polychlorinated Dibenzofurans (for the Numbering see Chart 4)

Cl positions	pEC_{50}	Cl positions	pEC_{50}
2	3.553	1,2,4,7,8	5.886
3	4.377	2,3,4,7,8	7.824
4	3.000	1,2,3,4,7,8	6.638
2,3	5.326	1,2,3,6,7,8	6.569
2,6	3.609	1,2,4,6,7,8	5.081
2,8	3.590	2,3,4,6,7,8	7.328
1,3,6	5.357	2,3,6,8	6.658
1,3,8	4.071	1,2,3,6	6.456
2,3,4	4.721	1,2,3,7	6.959
2,3,8	6.000	1,3,4,7,8	6.699
2,6,7	6.347	2,3,4,7,9	6.699
2,3,4,6	6.456	1,2,3,7,9	6.398
2,3,4,8	6.699		3.000
1,3,6,8	6.658	2,3,4,7	7.602
2,3,7,8	7.387	1,2,3,7	6.959
1,2,4,8	5.000	1,3,4,7,8	6.699
1,2,4,6,7	7.169	2,3,4,7,9	6.699
1,2,4,7,9	4.699	1,2,3,7,9	6.398
1,2,3,4,8	6.921	1,2,4,6,8	5.509
1,2,3,7,8	7.128		

Table 4. Binding Affinities of Polychlorinated Biphenyls (for the Numbering See Chart 4)

Cl positions	pEC_{50}	Cl positions	pEC_{50}
3,3′,4,4′	6.149	2,3,3′,4,4′,5	5.301
3,4,4′,5	4.553	2,3′,4,4′,5,5′	4.796
3,3′,4,4′,5	6.886	2,3,3′,4,4′,5′	5.149
2′,3,4,4′,5	4.854	2,2′,4,4′	3.886
2,3,3′,4,4′	5.367	2,2′,4,4′,5,5′	4.102
2,3′,4,4′,5	5.041	2,3,4,5	3.854
2,3,4,4′,5	5.387	2,3′,4,4′,5′,6	4.004

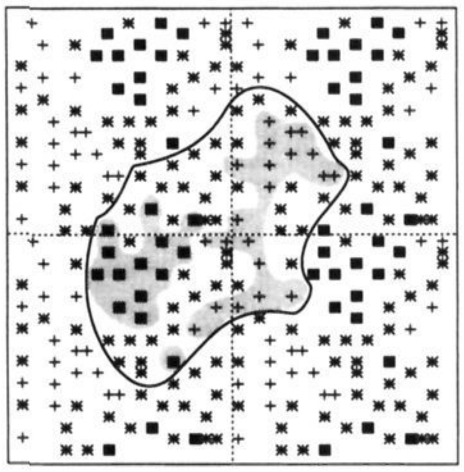


Figure 11. The 4-fold replication of the 20 × 20 Kohonen map of the polyhalogenated aromatic compounds: squares, high affinity; asterisks, medium affinity; crosses, low affinity. The area occupied by one replication of the whole dataset is enclosed by a black line. Areas of compounds with high and low affinity are shaded.

calculated by eq 4 and then used to train a 20 × 20 Kohonen network. Figure 11 shows the 4-fold replicated map. The replication enables one to show the entire dataset in one coherent area as indicated by the enclosing black line. Squares, asterisks, and crosses mark compounds with high, intermediate, and low toxicity, respectively. Points referring to molecules with high and low toxicity are grouped closely together and highlighted by shaded areas. The compounds with medium toxicity (stars) are in the area of transition from high to low toxicity. Note that the six compounds of medium toxicity in the domain at the right-hand side of the encircled area also are in a transition region from low to high toxicity. Only one outlier can be detected: a highly toxic compound surrounded by molecules with medium toxicity. The proper clustering of the compounds in the map gives evidence that the hydrophobicity descriptor used is well-suited for modeling the *Ah* affinity.

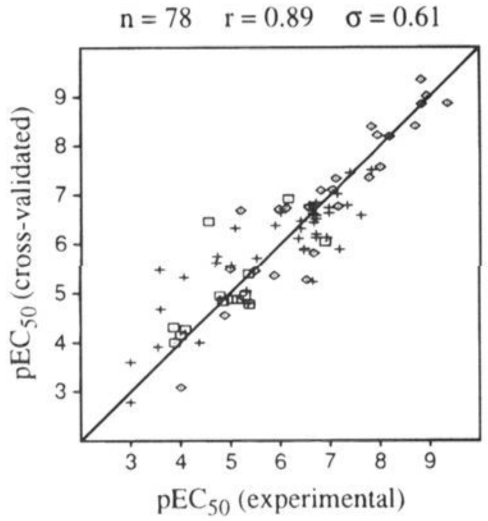


Figure 12. Plot of the experimental pEC_{50} values of the polyhalogenated dibenzo-*p*-dioxins (rhombs), dibenzofurans (crosses), and biphenyls (squares) against the cross-validated pEC_{50} values.

This result encouraged us to use the autocorrelation vectors and the pEC_{50} values to train a multilayer neural network by the backpropagation algorithm. Figure 12 shows a plot of the experimental pEC_{50} values against the cross-validated values (leave-one-out). A good correlation can be observed ($r = 0.89$, $\sigma = 0.61$) with a high cross-validated r^2 of 0.83. For comparison, Waller and McKinney obtained with a CoMFA model for the same dataset a cross-validated r^2 of 0.72.[36]

Conclusions

Autocorrelation allows the representation of molecular surface properties by a vector of fixed length, independent of the size of a molecule. The approach chosen here takes account of the magnitude of a surface property and its change over distance. It provides a dramatic reduction in the number of data—in our case approximately 3500 single point data were condensed into a vector of length 12. Furthermore, this representation of a molecular surface is invariant to rotation and translation. Thus, no special alignment of a molecule is required.

Combination of this representation of a molecule with neural network methods has a vast potential for the classification and modeling of physical, chemical, or biological activity. Unsupervised learning techniques—such as in a Kohonen network—provide a method for choosing an appropriate surface property to enter into the autocorrelation. Supervised learning—such as in a feedforward network trained by the backpropagation algorithm—can then be used for modeling the activity in question.

This approach was found to be superior to reported results on the modeling of the *corticosteroid binding globulin* (CBG) affinity of 31 steroids. In a similar manner the cytosolic *Ah* receptor affinity of 78 polyhalogenated dibenzo-*p*-dioxins, dibenzofurans, and biphenyls was modeled by the autocorrelation of the hydrophobicity potential on the van der Waals surface with a higher predictive power than that obtained by the most widely used alternative method.

Acknowledgment. We gratefully acknowledge the inspiring remarks by Dr. G. Klebe (BASF-AG) and the helpful discussions with Dr. K.-P. Schulz (Computer-Chemie-Centrum). We thank H. Bauknecht and Dr. A. Zell (Institut für Parallele und Verteilte Höchstleistungsrechner, Universität Stuttgart) for assistance in using the SNNS neural network software. This work was supported by funds from the Bundesminister für Bildung und Forschung, Germany. We acknowledge financial assistance by the Fonds der Chemischen Industrie.

JA950229S