Locating Biologically Active Compounds in Medium-Sized Heterogeneous Datasets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists

Henri Bauknecht,[†] Andreas Zell,[†] Harald Bayer,[†] Paul Levi,[†] Markus Wagener,[‡]
Jens Sadowski,[‡] and Johann Gasteiger*[,‡]

Institut für Parallele und Verteilte Höchstleistungsrechner (IPVR), Universität Stuttgart, D-70565 Stuttgart,
Germany, and Computer-Chemie-Centrum, Institut für Organische Chemie, Universität Erlangen-Nürnberg,
D-91052 Erlangen, Germany

Received June 14, 1996[⊗]

Electronic properties located on the atoms of a molecule such as partial atomic charges as well as electronegativity and polarizability values are encoded by an autocorrelation vector accounting for the constitution of a molecule. This encoding procedure is able to distinguish between compounds being dopamine agonists and those being benzodiazepine receptor agonists even after projection into a two-dimensional self-organizing network. The two types of compounds can still be distinguished if they are buried in a dataset of 8323 compounds of a chemical supplier catalog comprising a wide structural variety. The maps obtained by this sequence of events, calculation of empirical physicochemical effects, encoding in a topological autocorrelation vector, and projection by a self-organizing neural network, can thus be used for searching for structural similarity, and, in particular, for finding new lead structures with biological activity.

INTRODUCTION

The development of new drugs or plant protection compounds is a laborious and expensive process. It is estimated that for each new drug that comes to the market about 25 000−35 000 new compounds have to be synthesized and screened. Clearly, this asks for massive amounts of investments. The development of new compounds having a desired biological activity usually starts from a structure with this activity and then searches for other structures, new lead compounds, that also have this activity but are sufficiently different from the initial structure in order to avoid patent infringements. Considering the present situation with so many compounds to be synthesized, any attempt to make this search for new lead structures more efficient must have great importance.

The advent of databases of chemical structures has provided a basis for the development of automatic methods for the search for new lead structures. The structure of the initial active compound, or a pharmacophore model, is taken and used for scanning a database of compounds, typically containing several hundreds of thousands of structures.

Several problems have to be solved in order to make this approach successful and efficient:

1. The chemical structures, both of the query compound and of those stored in the database, have to be coded in a manner that contains information responsible for biological activity. The binding of a substrate to its receptor is dependent on the shape of the substrate and on a variety of effects such as the molecular electrostatic potential, polarizability, hydrophobicity, and lipophilicity. The coding scheme must somehow, either explicitly or implicitly, account for these physicochemical effects. Furthermore, usually molecules of different size with different numbers of atoms have to be compared. The chosen representation of structures must allow this.

2. A similarity measure has to be defined that selects structures from the database as hits for the query structure, the initial search structure having biological activity. This similarity criterion must ensure that the compounds selected are those also having biological activity—or at least promise to have activity. Clearly, this also depends on the chosen structure representation, and, therefore, structure representation and similarity measure are closely tied together.

3. And, finally, the molecular encoding scheme must be sufficiently concise, and the search method sufficiently rapid, to allow the processing of files of 100 000−500 000 structures, or even more, within reasonable time.

We report here on a structure representation that encodes a variety of electronic effects and allows the comparison of molecules of different size. We show that this structure code is able to group compounds with the same biological activity together. Similarity of structures is detected by unsupervised learning in a self-organizing neural network. It is shown that in a dataset of 10 000 structures, compounds with a given biological activity can still be found. The method is sufficiently fast, particularly if implemented on a parallel computer, to allow the processing of large datafiles of structures.

Structure Representation by Autocorrelation Vectors. Searching in databases of structures is usually performed by screens showing the presence or absence of structural fragments.[1] From the very beginning, we, however, intended to account in our structure representation for a variety of physicochemical effects, such as charge or polarizability, centered on the atoms of a molecule. Thus, we were faced with the problem of having to compare molecules with different numbers of atoms. Information having variable length can be transformed into fixed-length information by autocorrelation. Moreau and Broto[2] were the first to apply

[†] Universität Stuttgart. E-mail: {bauknecht, zell, bayer, levi}@informatik.uni-stuttgart.de.

[‡] Universitat Erlangen-Nürnberg. {wagener, sadowski, gasteiger}@torvs.ccc.uni-erlangen.de.

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

S0095-2338(96)00346-0 CCC: $12.00

an autocorrelation function to the topology of a molecular structure (eq 1).

$$A(d) = \sum_{i,j} p_i p_j \tag{1}$$

$A(d)$ is the autocorrelation coefficient referring to atom pairs i, j separated by d bonds. p_i, p_j is an atomic property such as partial atomic charge on atom i or j, respectively. For different topological distances (number of intervening bonds), d, a series of coefficients are obtained that can be gathered in an autocorrelation vector. An autocorrelation vector has several useful properties. First, a substantial reduction in data can be achieved by limiting the topological distance, d. Second, the autocorrelation coefficients are independent of the original atom numbering—they are canonical. And thirdly, the length of the vector, $A(d)$, is independent of the size of the molecule. Topological autocorrelation vectors have already been used as molecular descriptors in QSAR studies.[3,4] Recently, a spatial autocorrelation vector based on properties on molecular surfaces was introduced and shown to be able to quantitatively model the corticosteroid binding globulin activity of a series of 31 steroids and of the cytosolic Ah receptor activity of 78 polyhalogenated aromatic compounds.[5]

Clearly, biological activity depends on the 3D-structure of molecules and, in particular, on physicochemical properties on molecular surfaces. Thus, the spatial autocorrelation vector derived from surface properties[5] would be a logical starting point for searching for new lead structures having a desired biological activity. However, in this investigation we wanted to keep things simple and explore how far one can come with the more readily available topological autocorrelation vectors, only accounting for the constitution of molecules.

The following properties were calculated by previously published empirical methods for all atoms of a molecule:

–σ charge, q_σ,[6]
–total charge, q_{tot},
–σ electronegativity, χ_σ,[7]
–π-electronegativity, χ_π,[8]
–lone pair-electronegativity, χ_{LP},
–atom polarizability, α[9]

In addition to these six electronic variables, the identity function, i. e., each atom being represented by the number 1, was used in eq 1.

The autocorrelation of these seven variables was calculated for seven topological distances (number of intervening bonds) from two to eight. The basic assumption thus was that the interaction of atoms beyond eight bonds can be neglected. Thus, the descriptor for representing molecular structures is given by eq 2

$$A(p_k,n) = \sum_{i,j \in M(n)} p_k(i) p_k(j) \tag{2}$$

with

$$\mathrm{M(n)} = \{\mathrm{(i,j)} | \# \text{ bonds(i, j)} = \mathrm{n}\}$$

$p_k(i)$ is the kth property on atom i, and # bonds (i, j) is the minimum number of bonds between atoms i and j.

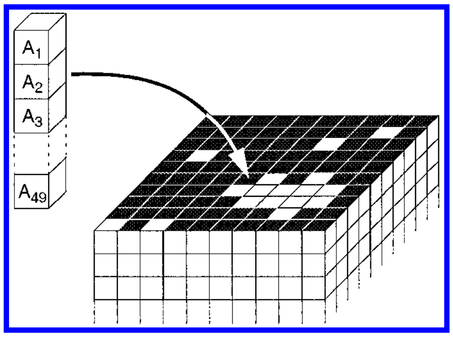


Figure 1. Projection of an m-dimensional object vector into a two-dimensional Kohonen network.

With seven variables and seven distances an autocorrelation vector of dimension 49 was obtained for each molecule irrespective of its size or number of atoms. The hydrogen atoms were not considered in the calculation of the autocorrelation vector.

Self-Organizing Neural Networks. Kohonen[10,11] introduced a neural network model that generates self-organizing maps. The basic processing elements, the artificial neurons, are arranged in an n-dimensional, usually two-dimensional, network. Objects, characterized by m variables are projected into this network. With $m > n$, a Kohonen network can be used to project a higher-dimensional space into a lower-dimensional space (Figure 1).[12−14] It has been shown that molecular surface properties such as the molecular electrostatic potential (MEP) can be projected into a two-dimensional Kohonen network.[15,16] The maps of surface properties thus obtained can be used for comparing biologically active compounds. Thus, it was shown that muscarinic and nicotinic antagonists can be distinguished by such maps of the MEP.[17]

In our application, a 49-dimensional space is projected into a two-dimensional space. The projection is such that the topology of the information is preserved as good as possible, i.e., points that are close together in the high-dimensional space will end up in the same or closely adjacent neurons.

Learning in a Kohonen network is an unsupervised and competitive process. An object, s, characterized by m variables, x_{si}, will be projected into that (central) neuron, c_s, that has weights, w_{ji}, most similar to the input variables (eq 3).

$$\mathrm{out}_{c_s} \Leftarrow \min\left[\sum_{i=1}^{m} (x_{si} - w_{ji})^2\right] \tag{3}$$

In the learning process, the weights of the neurons in the network are changed to make them even more similar to the input variables. The weights of all neurons are adjusted but to an extent that decreases with increasing distance from the central, winning neuron, c_s. In the end, a molecule is projected into that neuron of the network with weights that come closest to the description of the molecule by the autocorrelation vector.

It should be realized that the criterion embedded in eq 3 for determining the winning neuron for an object basically constitutes the measure determining the similarity of molecular structures. Molecules with similar autocorrelation vectors, X_s, are projected into the same or closely adjacent neurons.

Chart 1. Seventeen Dopamine Agonists (DPA) Taken from Ref 18

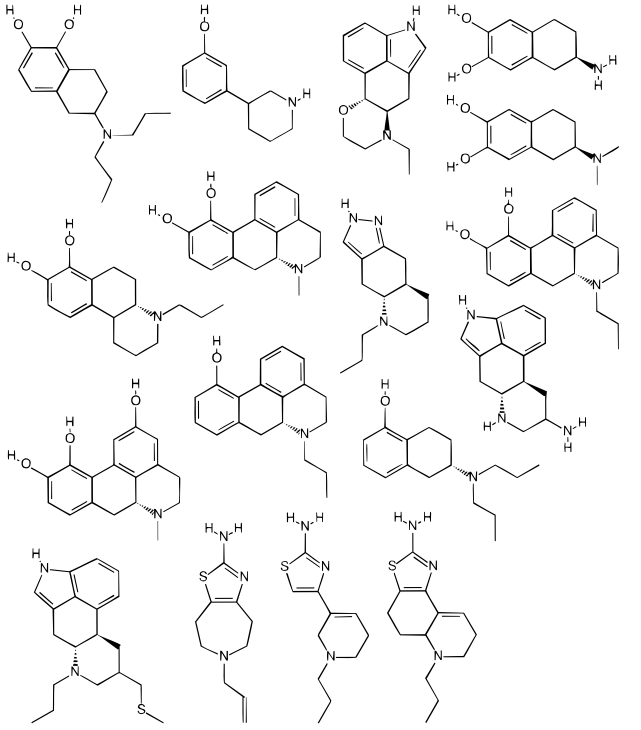
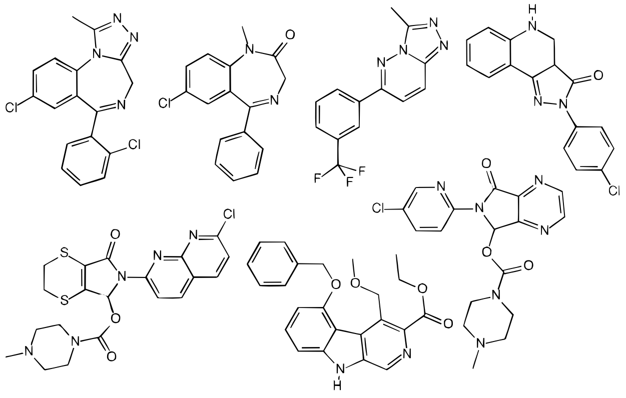


Chart 2. Seven Benzodiazepine Receptor Agonists (BDA) Taken from Ref 18



Datasets. A sequence of datasets of increasing size was used in this investigation.

Dataset I: First, a dataset with 17 dopamine agonists (DPA) (Chart 1) and seven benzodiazepine receptor agonists (BDA) (Chart 2) was taken from literature.[18] This small dataset was used to determine whether the autocorrelation vector is able to discriminate dopamine agonists from benzodiazepine receptors agonists in a Kohonen network.

Dataset II: Next, this dataset was extended by structures obtained in a search in the MDDR-3D database[19] arriving at a dataset of 112 dopamine agonists and 60 benzodiazepine receptor agonists. This dataset was used to further investigate the discriminating power of the autocorrelation vectors with a wider variety of structures.

Dataset III: In order to see whether these two classes of compounds, dopamine and benzodiazepine receptor agonists, cannot only be separated from each other but also distinguished from compounds not having either one of the two biological activities, dataset II containing 172 active compounds was augmented with 197 structures selected from the Janssen Chimica (now: Acros Organics) catalog of available starting materials.[20] From 8323 structures contained in the Janssen Chimica catalog, all those compounds were selected, that had the same range in elemental composition and number and types of rings as observed in the 172 dopamine and benzodiazepin agonists.

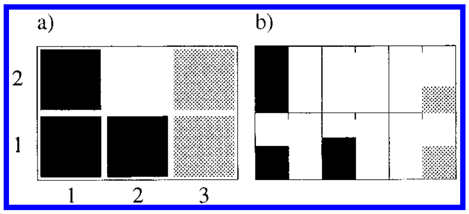


Figure 2. Kohonen map of size 3*2 neurons obtained from dataset I. (a) Neurons colored according to the type of compounds they contain; black: dopamine agonists, DPA; light gray: benzodiazepine receptor agonists, BDA. (b) Neurons indicating the type and number of compounds they contain. Colors as in (a); the size of the columns indicates the number of compounds.

Dataset IV: Finally, the 172 active compounds of dataset II were united with the full Janssen Chimica catalog of 8323 structures.

Experiments. The training of the Kohonen networks as performed with the Kohonen simulator KNet[21] implemented on a combination of a massively parallel computer and a Unix workstation. The massively parallel SIMD system MasPar MP1216 with 16 384 processors was used for the simulation of the Kohonen network performing the intensive computations. The front-end DEC station 5000/200 running under Ultrix was used to pilot the parallel computer and to graphically represent intermediate and final results under X-Windows. This set-up is—depending on the size and topology of the network—about 200−500 times faster than a standard workstation.

RESULTS AND DISCUSSION

Dataset I. Several two-dimensional Kohonen networks with different sizes were investigated. Thus, each of the compounds represented in a 49-dimensional space spanned by the autocorrelation values of eq 2 is projected into a specific neuron arranged in a two dimensional map. Structures in close vicinity in the high-dimensional space will be projected into the same or closely adjacent neurons of the Kohonen map. The smaller the network, the more structures will be projected into the same neuron. Thus, by defining the size of the network one can adjust the amount of compression of information. Figure 2 shows the results obtained with a Kohonen map of 3*2 neurons. The map obtained with this network is shown twice, on the left hand side (Figure 2a) the *types* of compounds, dopamine or benzodiazepine agonists, mapped into the individual neurons are indicated, on the right hand side (Figure 2b), in addition, the *number* of compounds projected into the various neurons is represented.

In Figure 2 and subsequent figures, dopamine agonists (DPA) are identified by black color, whereas benzodiazepine agonists (BDA) are indicated in light gray color. Figure 2a shows that the DPA are mapped into three neurons, whereas the BDA end up in two neurons; one neuron, neuron(2,2), does not receive any one of the 24 compounds, it stays empty. Figure 2b indicates this result in a more quantitative manner. The number of compounds mapped into a neuron is indicated by the height of bars, black bars on the left-hand side of a neuron identifying DPA and light gray bars

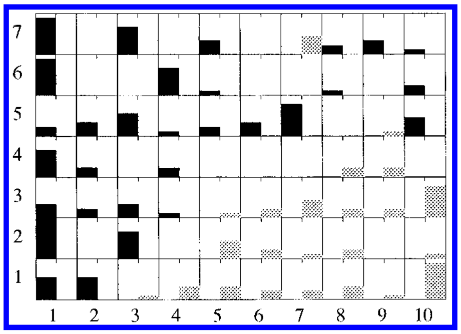
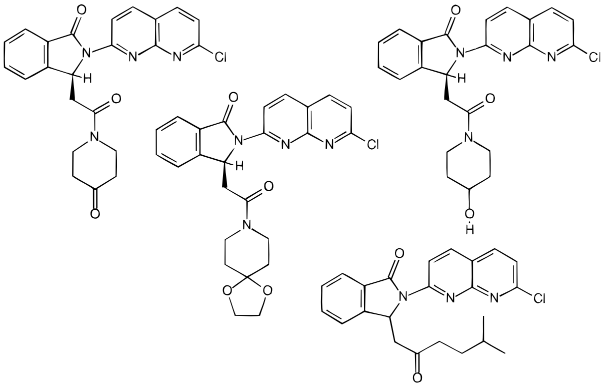


Figure 3. Kohonen map of size 10*7 neurons obtained for dataset II. The coloring of the neurons indicates the type and number of compounds they receive as explained in Figure 2b.

Chart 3. Structures Projected into Neuron(7,7) of the Map Shown in Figure 3



to the right-hand side BDA. Neuron(1,2) obtained eight DPA, neuron(1,1) four, and neuron(2,1) five DPA. Neuron(3,1) obtains four BDA and neuron(3,2) three BDA. The most important finding is that the two types of compounds, DPA and BDA, are clearly separated. No collision, the simultaneous projection of a DPA and a BDA into the same neuron, was observed. This attests to the power of the representation of these compounds by autocorrelation vectors for reflecting the biological activity of these compounds which is maintained even after the projection by the Kohonen method. The promising results in this exploration with a small dataset warranted investigation of a larger dataset.

Dataset II. This dataset consisted of 112 dopamine agonists (DPA) and 60 benzodiazepine agonists (BDA). The results obtained with a Kohonen map of size 10*7 neurons are shown in Figure 3.

The two types of compounds, DPA and BDA, separate quite well also in this larger dataset. There is only one intruder into the domain of dopamine agonists; neuron(7,7) is occupied by four benzodiazepine agonists, shown in Chart 3.

Clearly, these compounds internally show a high degree of structural similarity by the presence of both a benzopyrrolidone and a naphthyridine ring. It is not quite clear why they are projected into the domain of the DPA. The closest neuron also bearing a BDA is neuron(9,5) having a quinoline ring (Chart 4a). These two neurons are separated by neuron(8,6) that contains a structure with a tetrahydroindol ring (Chart 4b).

It is quite remarkable that the transition area between the region of DPA and that of BDA is indicated by a series of empty neurons or those that contain only a few DBA. This underlines the good separation capability of our approach to distinguish between DPA and BDA. This potential is observed here although the projection of 172 compounds into 70 neurons must result in quite some compression of information.

Chart 4. Structures Mapped into (a) Neuron(9,5) and (b) Neuron(8,6) of Figure 3

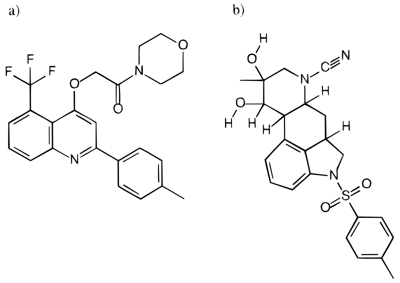
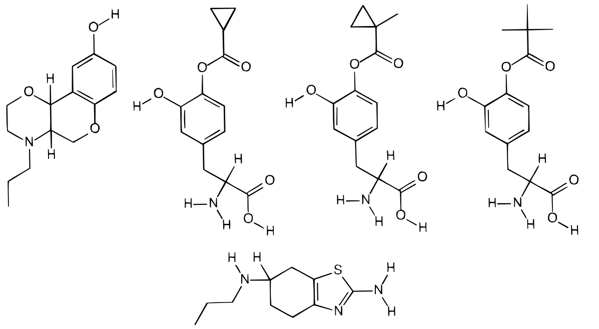


Chart 5. Structures Projected into Neuron(1,1) of the Map Shown in Figure 3



The power of our approach to perceive structural similarity that may be derived either easily or with difficulty from a visual inspection of the structural formula can be seen from the contents of many neurons. As an illustration, the molecules mapped into neuron(1,1) are given in Chart 5.

Neuron(1,1) contains five DPA, three are derivatives of 3,4-dihydroxyphenylalanine. Clearly, this structural similarity can easily be extracted from direct inspection of the structural formula. However, our approach also notes that the first other compound, and the last having quite different ring systems, are structurally similar to the three compounds, a similarity that cannot directly be deduced by a visual inspection of the structural formulas. However, this similarity must exist because all five compounds bind to the dopamine receptor. It is gratifying that this similarity is perceived by our encoding scheme.

Having established that the chosen electronic properties and their encoding in topological autocorrelation vectors are able to clearly separate DPA and BDA compounds within the Kohonen learning process, we addressed the problem whether these two types of compounds can still be recognized and separated when contained in a sea of other compounds. This is the theme of the investigations on datasets III and IV.

Dataset III. The dataset of 112 DPA and 60 BDA was extended with 197 compounds of the Janssen Chimica catalog of available starting materials. These additional compounds were chosen on the basis of criteria that should make their structure quite similar to DPA and BDA (see above). The activities of these additional 197 compounds are not known.

Figure 4 shows the results obtained for a Kohonen map of size 20*15 neurons. Most of the DPA compounds are collected in the upper left-hand corner of the map; only five

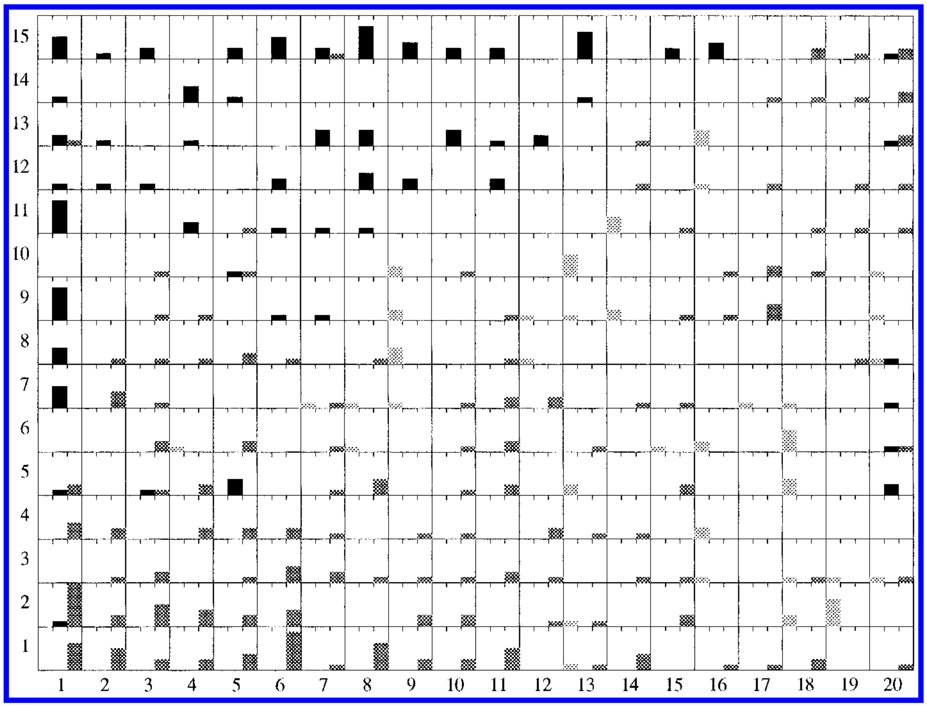


Figure 4. Kohonen map of size 20*15 neurons obtained for dataset III. The coloring of the neurons indicates the types and number of compounds that are mapped into them (cf. Figure 2b). Black: DPA; light gray: BDA; dark gray: compounds of unknown activity.

DPA are remarkably separated (neuron(20,5) to neuron(20,8)) from the rest of the DPA. Only four compounds of unknown activity are found in the large coherent area of DPA in the upper left-hand corner of the map. Most of the DPA can quite well be separated both from the BDA and the compounds of unknown activity. The BDA occur in several clusters within the area of compounds of unknown activity, with a clear tendency to find the BDA in the space from the border to the DPA area to the lower right corner of the map. There is only a single neuron, neuron(20,8), where a collision between a DPA and BDA occurs. Furthermore, one BDA, the one projected into neuron(4,6), comes quite close to the domain of DPA.

Clearly, there is not one small island of DPA and one of BDA in the sea of compounds of unknown activity. However, this cannot be expected, for both DPA and BDA encompass quite a variety of structures, and the structures of unknown activity had intentionally been chosen to make them structurally quite similar to DPA and BDA. However, even after addition of “noise”, of structures of unknown activity, DPA and BDA compounds can still remarkably well be separated from each other and also fairly well from the compounds of unknown activity. It may even well be that the structures of unknown activity comprise DPA and BDA.

Dataset IV. The promising results with the exploratory dataset III made us address the same kind of questions—(1) can DPA and BDA be identified among a sizable number of diverse structures, and, (2) can DPA and BDA then still be separated from each other—with a much larger dataset, containing the 172 active compounds, DPA and BDA, in the bulk of 8323 structures of unknown activity. These structures, comprising the entire set of compounds offered by a fine chemicals supplier, is of great structural variety, all the way from acetic acid and thiophosgene to triphenylmethane dyestuffs and dipeptides.

The map resulting from training a Kohonen network of 40*30 neurons with these 8495 structures is shown in Figure 5.

The map of Figure 5 shows that both, DPA and BDA, occupy only limited areas in the overall map. Furthermore, the areas of DPA and BDA are quite well separated from each other, only one neuron with BDA, neuron(4,24), intrudes into the domain of DPA and only two neurons with conflicts, obtaining both DPA and BDA, occur: neuron(2,7) and neuron(6,12). Clearly, the areas of neurons with DPA or BDA are larger than one probably has hoped them to be. For, with the results obtained here, the search for new active compounds or new lead structures in a dataset of compounds of unknown activity will have to scan a fairly large area and, correspondingly, quite a few compounds. However, compared to the overall size of the network, the areas where DPA and BDA are to be found are distinctly smaller and quite concentrated.

Let us now turn to a more detailed discussion of structural similarities that can be derived from a closer inspection of the Kohonen map of Figure 5. This analysis will further our understanding what kind of similarity is perceived in our approach and its power for clustering compounds. Clearly, a discussion of all the details distilled into this map is beyond the scope of this paper. Rather, we want to concentrate on those aspects that pertain to DPA and BDA compounds.

However, to prepare the reader for an appreciation of the similarities perceived here, let us start with the inspection of the contents of a neuron not containing DPA or BDA, such as neuron(8,30). The structures that have been mapped into this neuron are shown in Chart 6.

As can be seen, these are all high molecular weight hydrocarbons or derivatives therefrom with only one or two functional groups. Thus, overall, the nonpolar character of these compounds is clearly prevailing. Interestingly, no high molecular weight aliphatic ring compounds, although contained in the dataset, are mapped into this neuron; they are taken as sufficiently dissimilar to those shown in Chart 3. The similarity of these structures is also reflected in the

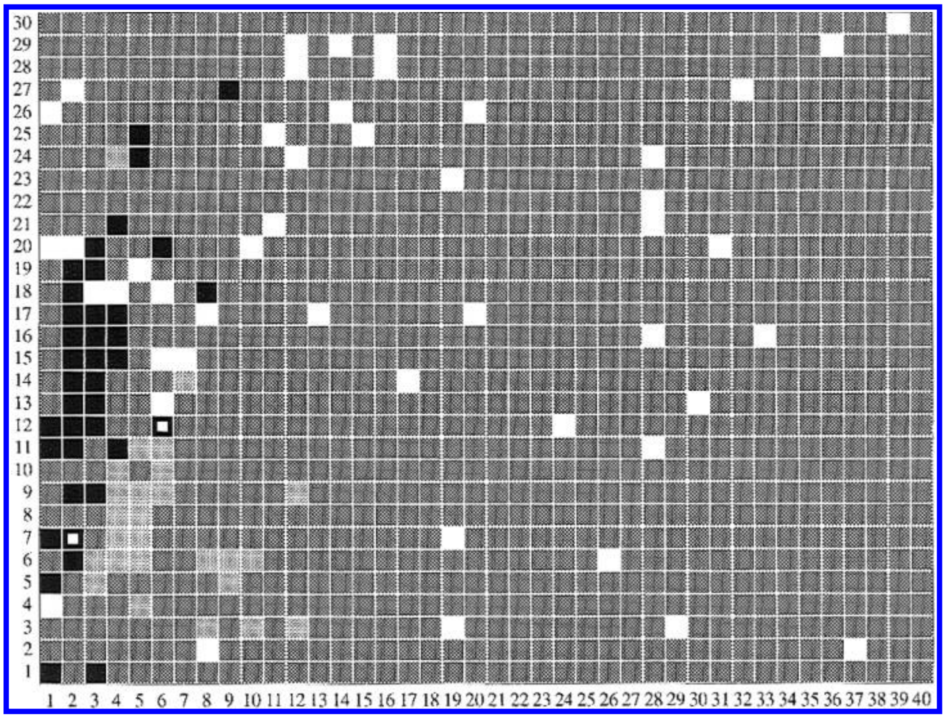


Figure 5. Kohonen map of 40*30 neurons obtained by training with dataset IV. Only the types of compounds mapped into the individual neurons is indicated. Again, black identifies DPA, light gray BDA, and dark gray compounds of unknown activity. Empty neurons are colored in white; the two neurons marked by a black frame indicate conflicts where both DPA and BDA are mapped into the same neuron.

Chart 6. Structures Projected into Neuron(8,30) of the Map Shown in Figure 5

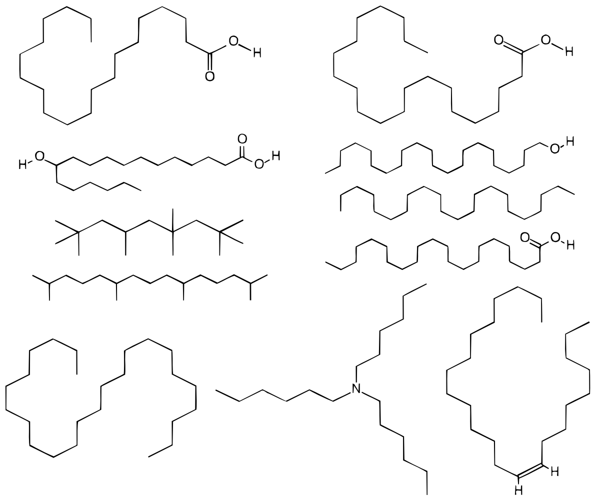


Chart 7. Structures Projected into Neuron(9,30) of the Map Shown in Figure 5

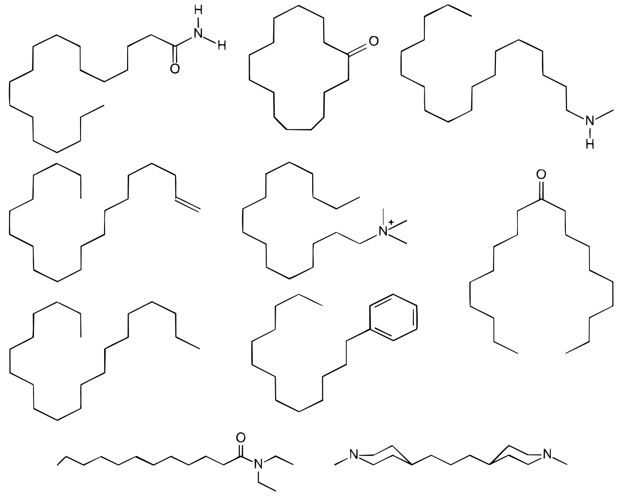


Table 1. Molecular Formulas of the Compounds of Neuron(8,30) of the Map Shown in Figure 5 (Cf. Chart 6)

		Hydrocarbons		
$C_{16}H_{34}$	$C_{19}H_{40}$	$C_{22}H_{46}$	$C_{23}H_{46}$	$C_{24}H_{50}$
		O_1 or N_1 Compounds		
$C_{22}H_{46}O$	$C_{18}H_{39}N$			
		O_2 or O_3 Compounds		
$C_{20}H_{40}O_2$	$C_{21}H_{42}O_2$		$C_{22}H_{44}O_2$	$C_{18}H_{36}O_3$

molecular formulas of these compounds as given in Table 1.

These are compounds having at least 16 carbon atoms and at most one nitrogen or one, two, or three oxygen atoms. Clearly, the carbon and hydrogen atoms are overwhelmingly prevailing.

This example shows the power of the Kohonen learning algorithm, given a proper coding of molecular structures, for perceiving structural similarity. The topology conserving features of self-organization in a Kohonen network have the effect that similar structures can be found in closely adjacent neurons. To illustrate this, we reproduce here in the following three schemes the contents of neuron(9,30) (Chart 7), neuron(8,29) (Chart 8), and neuron(9,29) (Chart 9).

All the structures mapped into the four adjacent neurons ((8,30), (9,30), (8,29), (9,29)) contain rather high molecular weight compounds with high hydrophobicity bearing only 0−2 functional groups. This result underlines that molecules found in the same or closely adjacent neuron are indeed structurally quite similar.

Let us now turn our attention to those parts of the map shown in Figure 5, that contain active compounds, DPA or BDA. It was already said that the two types of biologically active compounds separate quite well. We will first analyze neurons that contain only either BDA or DPA (or compounds

Chart 8. Structures Projected into Neuron(8,29) of the Map Shown in Figure 5

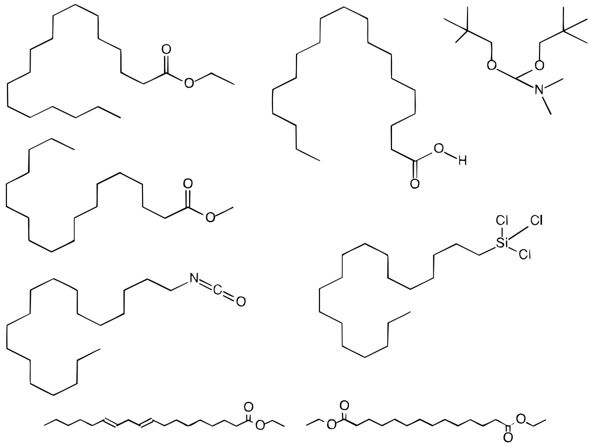


Chart 9. Structures Projected into Neuron(9,29) of the Map Shown in Figure 5

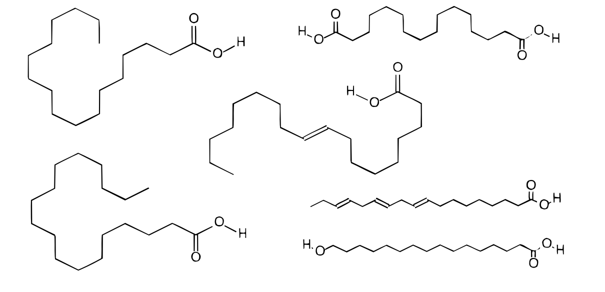
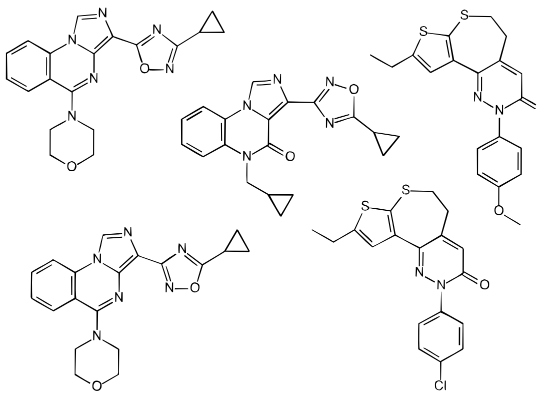


Chart 10. Structures Projected into Neuron(5,8) of Figure 5, All Being BDA



of unknown activity) and then analyze those regions of the map where conflict situations occur, situations where both DPA or BDA compounds are mapped into the same neuron or directly adjacent neurons.

With neuron(5,8), a part of the map only containing BDA is shown in Chart 10.

The structural formulas show that the five BDA mapped into this neuron fall into two classes, three molecules with a condensed tricyclic ring system consisting of two six-membered and one five-membered ring, carrying altogether three nitrogen atoms in this ring system, and two molecules with a condensed tricyclic ring system consisting of a thiophene ring, a seven-membered ring with a sulfur atom, and a six-membered ring with two nitrogen atoms. It is gratifying to see that the autocorrelation vectors perceive both the similarity within and between the two classes of compounds and that this similarity is maintained after projection into the two-dimensional self-organizing map.

Chart 11. Structures Mapped into Neuron(3,9) of Figure 5 Consisting of Three Compounds with Unknown Activity and Two DPA

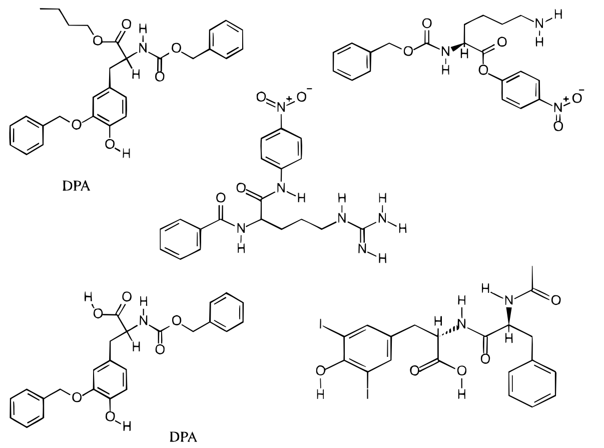
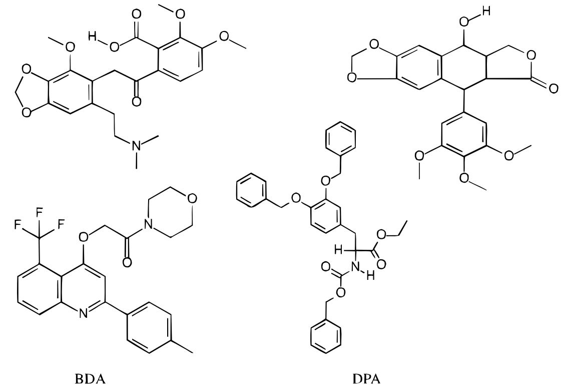


Chart 12. Structures Mapped into Neuron(2,7), a Conflict Neuron



Next, the contents of neuron(3,9) is shown in Chart 11. The set of compounds of neuron(3,9) consists of three compounds of unknown activity and two DPA. The five structures have quite a few features in common. In fact, the last structure, *N*-acetyl-L-phenylalanyl-3,5-diiodo-L-tyrosine is the most similar to the two DPA and promises to be a candidate worth testing for DPA activity.

Next, the two neurons containing a conflict situation between DPA and BDA are analyzed. The first conflict is found in neuron(2,7); Chart 12 shows the structures mapped into this neuron.

At first sight, the dopamine and the benzodiazepine agonists look quite different. However, one should not be misled by only looking at the structural formula. Biological activity is tied to geometric and electronic factors that go beyond the information derived from visual inspection of a two-dimensional graph. We have tried to account for the electronic factors by using a series of such descriptors in the autocorrelation vectors (cf. eq 1). That we have succeeded to perceive similarity that goes beyond that contained in the two-dimensional structural formula can be seen from the structures contained in neuron(5,8) (Chart 10). The reason for the conflict occurring in neuron(2,7) cannot be explained in simple terms. It has to be attributed to the fact that the two types of compounds, DPA and BDA, are indeed not that different. This is expressed by the fact that the areas containing DPA and BDA compounds touch each other in this part of the area. In order to inspect this further we show the structures mapped into adjacent neurons. Chart 13 shows the structures contained in neuron(2,6); all are dopamine agonists.

Chart 13. Structures Mapped into Neuron(2,6), All Being DPA

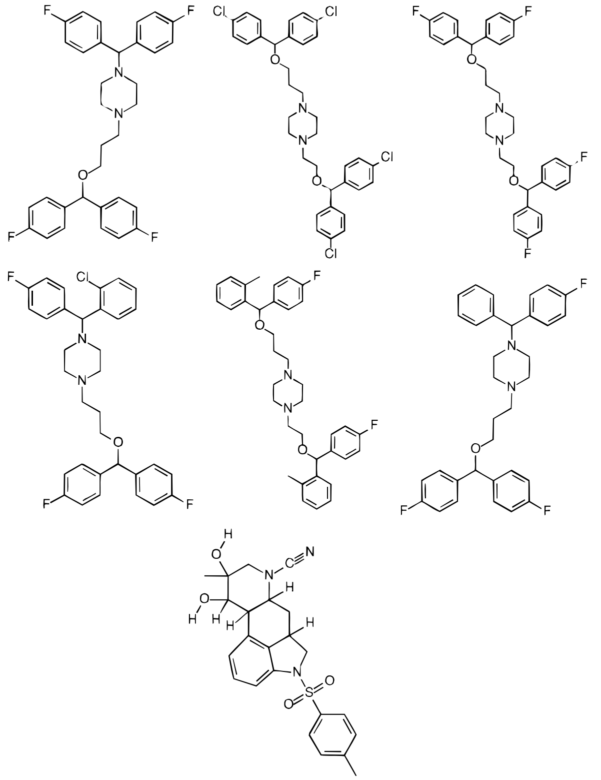


Chart 14. Structures Mapped into Neuron(3,6)

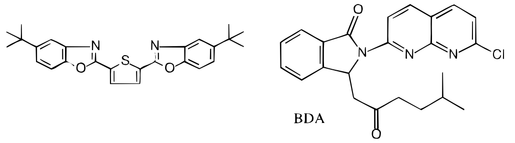


Chart 15. Structures Mapped into Neuron(6,12) of Figure 5, One of the Conflict Neurons

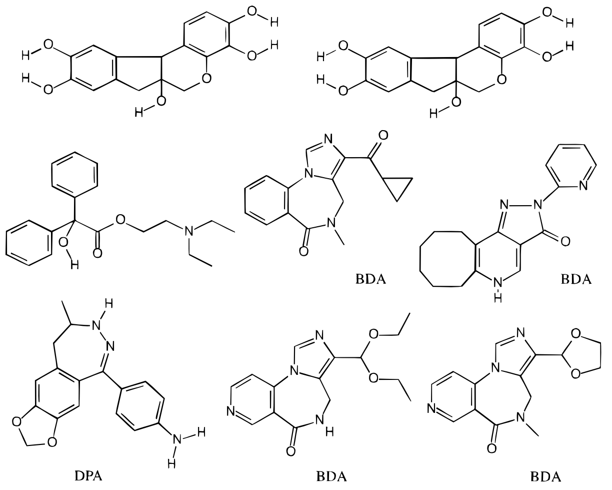


Chart 14 gives the two compounds contained in neuron(3,6); one is a benzodiazepine agonist, the other compound is from the chemical supplier and is of unknown activity.

The structures mapped into the other conflict neuron, neuron(6,12), are shown in Chart 15. The first three compounds are from the Janssen Chimica catalog and are of unknown activity. The seventh molecule is a DPA, whereas the other four compounds are BDA. The structural formulas show that the DPA has a close similarity to three BDA; all four compounds possess a seven-membered ring with two nitrogen atoms condensed to an aromatic six-membered ring. It is therefore not surprising that these four structures are mapped into the same neuron. There must, however, be fine details in the structure of molecule no. 7 that make it a DPA as against the other three compounds showing BDA activity.

Chart 16. Structures Contained in Neuron(4,24)

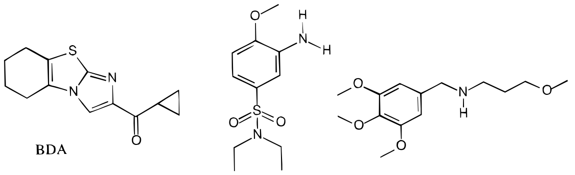


Chart 17. Structures Contained in Neuron(5,24)

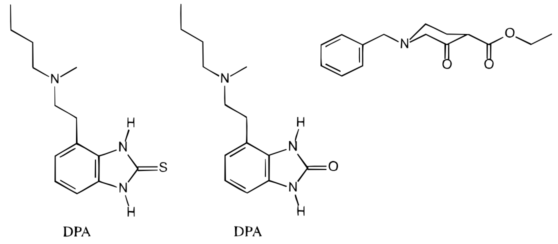
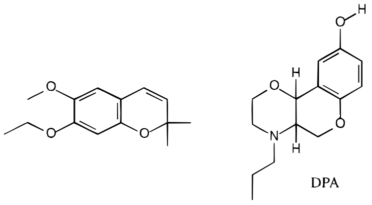


Chart 18. Structures Contained in Neuron(5,25)



Finally, there is an area where a BDA compound penetrates into the domain of DPA compounds. Neuron(4,24) (Chart 16) contains a BDA, whereas neuron(5,24) (Chart 17) and neuron(5,25) (Chart 18) both contain DPA compounds.

Inspection of the structural formula of the BDA of neuron(4,24) and those of the two DPA of neuron(5,24) show some distinct similarities, and thus their closeness in the Kohonen map is understandable. However, we have already cautioned to interpret too strongly the two-dimensional structural formula. Biological activity is dependent on the three-dimensional structure of a compound and electronic effects such as electrostatic potentials, hydrophobicity, and hydrogen bridging potentials. We have tried to encompass such effects to a certain extent by the various atomic properties used in the autocorrelation vectors. That we have been somehow successful in perceiving structural similarity that goes beyond that contained in the two-dimensional graph can be seen from a comparison of the two DPA in neuron(5,24), with the DPA in neuron(5,25). Although the structural formulas have quite distinct differences the projection of these compounds into adjacent neurons points to their structural similarity.

In trying to rationalize the closeness of BDA and DPA in neurons(4,24), (5,24), and (5,25) one has to recall that these molecules are represented in a 49-dimensional space (by autocorrelation vectors). Projection of such a high dimensional space into two-dimensional necessarily must lead to substantial distortions. In view of this, it is quite remarkable that the areas of the two-dimensional map where BDA and where DPA compounds can be found only slightly overlap at three points, neuron(2,7), neuron(6,12), and neuron(4,24).

CONCLUSIONS

The six electronic factors and the connectivity of a molecule, their encoding into topological autocorrelation vectors, and their projection into a two-dimensional map by the self-organizing capability of a Kohonen network provide a powerful means for the detection of similarity in the structure of organic molecules.

Dopamine agonists can be separated from benzodiazepine receptor agonists and this separation is maintained when these two types of compounds are embedded in a larger set of structures.

This opens the way for searching for compounds with a desired biological activity and for discovering new lead structures in large databases of compounds. The implementation of a Kohonen network on a massively parallel computer promises to make the processing of datasets of several hundreds of thousands of structures a feasible endeavor.

Clearly, the more compounds of the same activity that are concentrated into a fewer numbers of neurons, the fewer number of structures have to be searched for new lead structures. We are continuing our efforts to achieve this.

ACKNOWLEDGMENT

We gratefully acknowledge support of our collaboration by the Bundesminister für Bildung, Wissenschaft, Forschung und Technologie (BMBF) within the scheme “Molekulare Bioinformatik”. We thank Acros Organics (formerly Janssen Chimica) for providing us with their dataset of available starting materials in computer-readable form. Furthermore, we thank MDL Information Systems for giving us access to the MDDR data base.

REFERENCES AND NOTES

(1) Willett, P. A review of chemical structure retrieval systems. *J. Chemometrics* **1987**, *1*, 139−155.

(2) Moreau, G.; Broto, P. The Autocorrelation of a Topological Structure: A New Molecular Descriptor. *Nouv. J. Chim.* **1980**, *4*, 359−360.

(3) Moreau, G.; Broto, P. Autocorrelation of Molecular Structures: Application to SAR Studies. *Nouv. J. Chim.* **1980**, *4*, 757−764.

(4) Zakarya, D.; Tiyal, F.; Chastrette, M. Use of the Multifunctional Autocorrelation Method to Estimate Molar Volumes of Alkanes and Oxygenated Compounds. Comparison between Components of Autocorrelation Vectors and Topological Indices. *J. Phys. Org. Chem.* **1993**, *6*, 574−582.

(5) 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.

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

(7) 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.

(8) Gasteiger, J.; Saller, H. Berechnung der Ladungsverteilung in konjungierten Systemen durch eine Quantifizierung des Mesomeriekonzepts. *Angew. Chem.* **1985**, *97*, 699−701. Calculation of Charge Distribution in Conjugated Systems by a Quantification of the Resonance Concept. *Angew. Chem., Int.. Ed. Engl.* **1985**, *24*, 687−689.

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

(10) Kohonen, T. *Biol. Cybern.* **1982**, *43*, 59−69.

(11) Kohonen, T. *Self-Organization and Associative Memory*, 3rd ed; Springer, Berlin, 1989.

(12) Gasteiger, J.; Zupan, J. Neuronale Netze in der Chemie. *Angew. Chem.* **1993**, 105, 510−536. Neural Networks in Chemistry. *Angew. Chem., Int. Ed. Engl.* **1995**, *32*, 503−527.

(13) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists: An Introduction*, VCH-Verlag, Weinheim, 1993.

(14) Zell, A. *Simulation neuronaler Netze*; Addison-Wesley, 1994.

(15) Gasteiger, J.; Li, X.; Rudolph, C; Sadowski, J.; Zupan, J. Representation of Molecular Electrostatic Potentials by Topological Feature Maps. *J. Am. Chem. Soc.* **1994**, *116*, 4608−4620.

(16) Gasteiger, J.; Li, X.; Uschold, A. The beauty of molecular surfaces as revealed by self-organizing neural networks. *J. Mol. Graphics* **1994**, *12*, 90−97.

(17) Gasteiger, J.; Li, X. Abbildung elektrostatischer Potentiale muscarinischer und nicotinischer Agonisten mit künstlichen neuronalen Netzen, *Angew. Chem.* **1994,** *106*, 671−674. Mapping the Electrostatic Potential of Muscarinic and Nicotinic Agonists with Artificial Neural Networks. *Angew. Chem., Int. Ed. Engl.* **1994**, *33*, 643−646.

(18) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico I.; Pavlik, P. A. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput.-Aided Mol. Design* **1993**, *7*, 83−102.

(19) MDDR-3D database; available from MDL Information Systems, San Leandro, CA, U.S.A.

(20) Janssen Chimica catalogue (now Acros Organics); version 1989.

(21) Bayer, H. *KNet User Manual*; University of Stuttgart.

CI960346M