

Knowledge Discovery in Reaction Databases: Landscaping Organic Reactions by a Self-Organizing Neural Network

Lingran Chen and Johann Gasteiger*

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

Received January 3, 1996. Revised Manuscript Received February 3, 1997[®]

Abstract: Chemists have always derived their knowledge about chemical reactions by inductive learning from observations on a series of individual chemical reactions. Predictions of the products of chemical reactions are made by analogy. With the availability of large reaction databases this process can be automated. In this paper a new method based on a Kohonen neural network and physicochemical variables for describing reaction centers is developed for this purpose. The results with two reaction datasets show how a set of chemical reactions with the same reaction center can automatically be classified, clearly revealing different levels of similarities of the reactions under investigation. The relative positions of reactions and clusters in the two-dimensional Kohonen map offer extra chemical information. A third reaction dataset is used to show how a trained Kohonen network can be used to predict reaction types for organic reactions.

Introduction

The advent of computers in chemical laboratories is increasingly influencing the work of organic chemists. Several computerized reaction databases containing hundreds of thousands, even several millions, of reactions are available,^{1–3} offering an easy way to access a cornucopia of individual reaction instances. It is the user's responsibility to analyze the retrieved reaction data and thus discover the essential features of a reaction type. However, often a single search can lead to a hit list of several hundred reactions, and thus the manual analysis is both laborious and time-consuming.

Another achievement in the application of computers in organic chemistry is the development of a number of computer-assisted synthesis design systems, such as LHASA,⁴ WODCA,^{5–7} and SYNGEN,⁸ and reaction prediction systems, such as EROS⁹ and CAMEO.¹⁰ In contrast to reaction database systems, both synthesis design and reaction prediction systems are knowledge-based; i.e., they must work with general representations of reaction types rather than with individual reactions. Unfortunately, up to now, the knowledge bases used in many of these systems still have to be built largely manually, which is both time-consuming and error-prone, strongly restricting the size of the knowledge bases and thus the applicability of the existing systems.

Thus, we see that the organization of hit lists from reaction retrieval systems as well as the building of knowledge bases for synthesis design and reaction prediction systems asks for categorizing reaction instances and for making generalizations about the resulting categories. One exploration of this problem has led to the development of the HORACE system (hierarchical organization of reactions through attribute and condition education).^{11–14} The importance of the problem of automatically extracting chemical knowledge from reaction databases warrants the exploration of other techniques.

Since the very beginning, chemists have derived their knowledge about chemical reactions by comparison of a series of individual reactions. Inductive learning has allowed them to draw conclusions and to make predictions of the products of chemical reactions by analogy.

In recent years neural networks, computer models of the information processing in the human brain, have gained prominence.^{15,16} Neural networks acquire knowledge about a certain task or problem from studying a training set of data; this is an inductive learning process. There are two basic ways of learning: supervised and unsupervised learning. In the first case, the neural network system is presented with a set of input patterns together with the corresponding correct answers. In unsupervised learning, the network automatically finds the distinguishing features between the different categories of patterns and organizes the output in a way that shows the relationships among the input patterns.

As neural networks mimic inductive learning, a process chemists have been so successful with in deriving their knowledge about chemical reactions, it is tempting to use neural networks for learning from the information contained in reaction

[®] Abstract published in *Advance ACS Abstracts*, April 15, 1997.

(1) Blake, J. E.; Dana, R. C. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 394–399.

(2) ChemReact, available from InfoChem, Munich, Germany.

(3) Crossfire, available from Beilstein Informationssysteme GmbH, Frankfurt/Main, Germany.

(4) Corey, E. J.; Wipke, W. T.; Cramer, R. D., III; Howe, W. J. *J. Am. Chem. Soc.* **1972**, *94*, 431–439.

(5) Gasteiger, J.; Ihlenfeldt, W. D.; Röse, P. *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 270–290.

(6) Fick, R.; Ihlenfeldt, W. D.; Gasteiger, J. *Heterocycles* **1995**, *40*, 993–1007.

(7) Ihlenfeldt, W. D.; Gasteiger, J. *Angew. Chem.* **1995**, *107*, 2807–2829; *Angew. Chem., Int. Ed. Engl.* **1995**, *34*, 2613–2633.

(8) Hendrickson, J. B. *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 323–334.

(9) Röse, P.; Gasteiger, J. *Anal. Chim. Acta* **1990**, *235*, 163–168. Gasteiger, J.; Hondelmann, U.; Röse, P.; Witznichen, W. *J. Chem. Soc., Perkin Trans. 2* **1995**, 193–204.

(10) Laird, E. R.; Jorgenson, W. L. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 458–466.

(11) Rose, J. R.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 74–90.

(12) Gasteiger, J.; Rose, J. R. In *Software-Development in Chemistry 8*; Jochum, C. Ed.; Springer: Berlin, 1994, pp 29–55.

(13) Chen, L.; Gasteiger, J.; Rose, J. R. In *Software-Development in Chemistry 9*; Moll, R., Ed.; Springer: Berlin, 1995; pp 129–139.

(14) Chen, L.; Gasteiger, J.; Rose, J. R. *J. Org. Chem.* **1995**, *60*, 8002–8014.

(15) Gasteiger, J.; Zupan, J. *Angew. Chem.* **1993**, *105*, 510–536; *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503–527.

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

databases. We will show here how the automatic classification of individual reactions into classes can be achieved by an unsupervised neural network technique, a Kohonen network,^{15–18} leading to a two-dimensional map. The representation of reactions in a two-dimensional map, the landscape of reactions, has two important benefits. Different directions in such a map can represent different *types* of similarities between reactions, and different *distances* can indicate different degrees of similarities.¹⁹

Methology

Datasets of Reactions. The datasets of chemical reactions used for the classification studies were obtained with the reaction retrieval system ISIS Host²⁰ from the ChemInform RX²¹ or the Theilheimer²⁰ reaction database. In each case, a reaction substructure search was performed, retrieving all those reactions with a certain chosen reaction center, i.e., with the same types of atoms and bonds directly involved in the bond rearrangement in the reaction.

Physicochemical Influences on the Reaction Site. Structural characteristics, or more basically the electronic and energy features, determine a reaction mechanism, and thereby the course of a reaction. We therefore calculated a variety of electronic effects at the reaction site, the set of atoms and bonds directly involved in the bond rearrangement during a reaction. In order to be able to deal with large datasets of reactions involving fairly sizable molecules, rapid empirical methods for the calculation of physicochemical effects such as partial atomic charges^{22,23} and inductive,²⁴ resonance,²³ and polarizability effects²⁵ were used, which are collected in the PETRA (parameter estimation of the treatment of reactivity applications) program package. For instance, the calculation of the physicochemical effects in the 74 product molecules of a dataset of 74 reactions (see the Results and Discussion) took only about 5 min of cpu time on a SUN Sparc 10 workstation. This process can be performed automatically by transferring the RFiles obtained from ISIS Host to the PETRA package. Thus, no redrawing of the chemical structures is necessary. In fact, these physicochemical variables can be stored in the reaction database once and for all and thus would not have to be redone in the classification of reactions or in the knowledge extraction process. The physicochemical factors obtained by these empirical methods have successfully been used for solving various problems, including the derivation of a reaction mechanism.²⁶

The number of different physicochemical properties concerning only the reaction center can be quite large. Some of the physicochemical features contain low information or high redundancy with respect to other features. For example, many bond physicochemical features (such as charge difference) are closely related to the corresponding atom physicochemical features (partial atomic charges). The distributed storage schemes of neural networks have an important advantage: The information representation can be redundant. A Kohonen network is not sensitive to linear dependencies in the input variables, quite in contrast to methods like multilinear regression analysis. This means that one can use all of the available physicochemical factors as input to the neural network; this both is simple and can avoid losing information. However, the presence of too many input features can heavily burden the training process and can produce a neural network

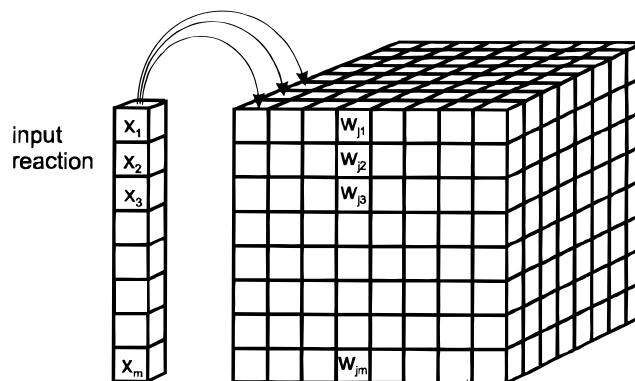


Figure 1. Basic architecture of a Kohonen network. An input pattern (vector), \mathbf{X} , consists of m elements which are presented to the corresponding input unit. Each neuron of the network is represented by a column with m weights, w_{jk} .

with many more connection weights than those required by the problem. Important work has been done and is still being made on automatic feature selection. We have examined a variety of methods such as statistical methods for the selection of those physicochemical variables containing most of the information necessary for the reactions. In fact, the Fisher t value proved to be a measure for automatically selecting the important variables.

In the discussion here, however, we intentionally give only sets of physicochemical variables selected by the user in order to keep the discussion simple and also demonstrate how a chemist can introduce his/her own models into the clustering of reactions by a Kohonen network.

Kohonen Network. The basic purpose of a Kohonen network is to construct a nonlinear projection of a high-dimensional pattern to a lower-dimensional space.^{15–18} In our case, a Kohonen network consists of a two-dimensional arrangement of neurons (Figure 1).

The dataset consists of p input reactions, and each reaction is described by m physicochemical variables and can be treated as an m -dimensional input vector:

$$\mathbf{X}_s = (x_{s1}, x_{s2}, \dots, x_{sm}) \quad s = 1, 2, \dots, p \quad (1)$$

Each neuron j has as many weights m as there are variables for describing a reaction. These m weight values constitute an m -dimensional weight vector:

$$\mathbf{W}_j = (w_{j1}, w_{j2}, \dots, w_{jm}) \quad j = 1, 2, \dots, n \quad (2)$$

where n is the number of neurons in the Kohonen network.

In a Kohonen network, an input pattern is presented to all neurons of the network and is then mapped into *one* neuron. The selection of the matching (winning) neuron for a given input reaction \mathbf{X}_s is made by comparison of the Euclidean distances, d , between the input vector, \mathbf{X}_s , and all the n weight vectors:

$$d_{sj} = \left[\sum_{i=1}^m (x_{si} - w_{ji})^2 \right]^{1/2} \quad j = 1, 2, \dots, n \quad (3)$$

The neuron, j , whose weight vector has the smallest distance value, d_{sj} , is selected as the matching neuron for the input pattern \mathbf{X}_s .

The weights of the winning neuron are adjusted in the learning process such that they become even more similar to the input pattern. Furthermore, the weights of all the other neurons are also adjusted but to an amount decreasing with increasing topological distance from the winning neuron. A Kohonen network is therefore also called a self-organizing feature map. It is necessary to present all input patterns into the network several times, in order that the weight values can gradually be adjusted in such a way that the weight vectors can approximate the input pattern vectors as good as possible.

After the learning process is completed, the entire dataset is presented again to the network one by one and each winning neuron in the competitive layer can be determined using the same method as described

(17) Kohonen, T. *Biol. Cybern.* **1982**, *43*, 59–69.

(18) Kohonen, T. *Proc. IEEE* **1990**, *78*, 1464–1480.

(19) For a preliminary report see: Chen, L.; Gasteiger J. *Angew. Chem.* **1996**, *108*, 844–846; *Angew. Chem., Int. Ed. Engl.* **1996**, *35*, 763–765.

(20) ISIS Host and the Theilheimer reaction database are available from MDL Information Systems Inc., San Leandro, CA.

(21) ChemInform RX is produced by Fachinformationszentrum Chemie, Berlin, from the information contained in the weekly abstracting service ChemInform, and marketed by MDL Information Systems Inc., San Leandro, CA.

(22) Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219–3228.

(23) Gasteiger, J.; Saller, H. *Angew. Chem.* **1985**, *97*, 699–701; *Angew. Chem., Int. Ed. Engl.* **1985**, *24*, 687–689.

(24) Hutchings, M. G.; Gasteiger, J. *Tetrahedron Lett.* **1983**, *24*, 2541–2544.

(25) Gasteiger, J.; Hutchings, M. G. *J. Chem. Soc., Perkin Trans. 2* **1984**, 559–564.

(26) Gasteiger, J.; Saller, H.; Löw, P. *Anal. Chim. Acta* **1986**, *191*, 111–123.

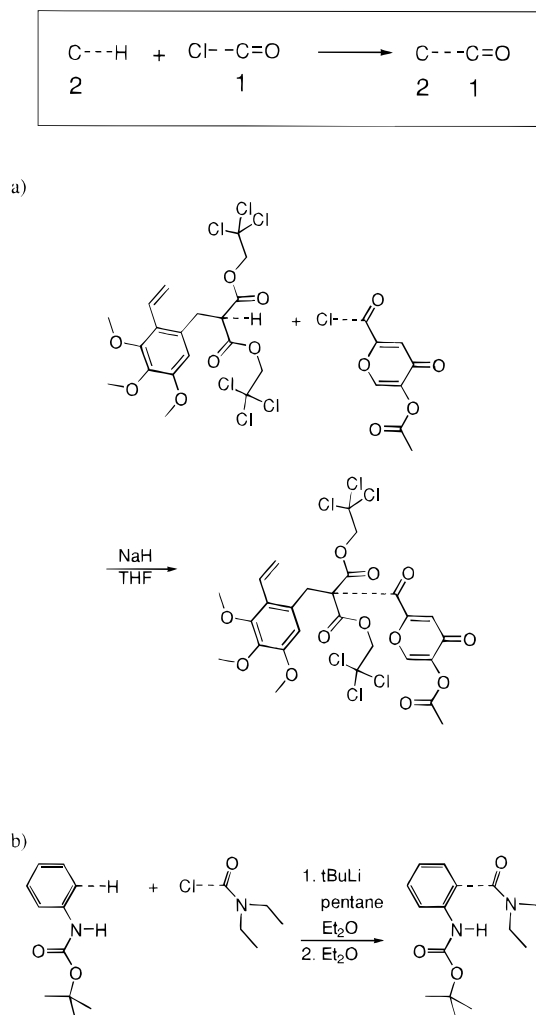


Figure 2. Reaction center common to all 74 investigated examples of reactions; the reaction center is indicated by dotted lines. (a) Nucleophilic aliphatic substitution of acyl chloride (no. 26). (b) Acylation of an arene (no. 51).

above. These winning neurons of all the input patterns are then labeled with appropriate markers characterizing the corresponding input patterns, leading to the final Kohonen feature map.

Automatic Detection of Clusters in the Kohonen Map. The unsupervised learning of a Kohonen network does not utilize any information on the relationships of the input in the learning process. A Kohonen network produces a map showing the indices of all the reactions in the corresponding occupied neurons. The problem is then how to automatically detect clusters in such a two-dimensional map.

The decisive factor which determines the final distribution of the input patterns in the map is the weight values of the neurons of the trained network (see eq 3). In order to perceive relationships and clusters in the map, one has to compute and compare the Euclidean distances of two weight vectors of each directly neighboring neuron pair.²⁷ Neurons which have small weight distances between each other are quite similar and may form one cluster. Large weight distances separate different clusters. Single neurons having large weight distances to all its neighbor neurons indicate outliers. We will show that such an automatic classification corresponds quite well to an intellectual assignment of reaction types.

Results and Discussion

A. Classification of 74 Reactions. The first dataset was obtained by retrieving all reactions with the reaction center shown in Figure 2 from the reaction database of the ChemInform RX 1994.²¹ Seventy-four reactions were found.

Table 1. Five Physicochemical Parameters Used To Characterize each Reaction Center of the Reactions in the First Dataset of 74 Reactions

electronic variable ^a	Cl-C=O 1	+ H-C 2	→	C-C=O 2 1
q_{tot}				×
χ_{σ}				×
χ_{π}				×
α_i				×
a_{ar}				×

^a q_{tot} = total charge. χ_{σ} = σ -electronegativity. χ_{π} = π -electronegativity. α_i = effective atom polarizability. a_{ar} = aromaticity indicator.

The reaction center shown in Figure 2 is common to several different reaction types, such as acylations of arenes, acylations of C=C bonds, and nucleophilic aliphatic substitution of acyl chlorides. These reactions show quite different features around the reaction site and can occur under quite different reaction conditions.

With this first dataset we wanted to explore the use of a Kohonen network for classifying chemical reactions in order to build a knowledge base for *synthesis design*. In synthesis design, one starts with the reaction *product*, the target molecule, and considers reactions in a retrosynthetic manner. Thus, only physicochemical variables of the *products* were chosen for representing a chemical reaction. In order to keep the study transparent, variables were to be selected by the user and the number of variables should be as small as possible. From all the atoms of the reaction center, carbon atom 2 (see Figure 2) shows the widest structural variety, and therefore it was decided to consider physicochemical properties of this atom, only. In order to ensure that the more important physicochemical effects, charge distributions and inductive, resonance, and polarizability effects, operative at this atom of the reaction site are considered, the following variables were chosen: the total charge, q_{tot} , the σ - and π -electronegativities, χ_{σ} and χ_{π} , the effective atom polarizability, α_i , and an indicator variable for aromaticity, a_{ar} , as shown in Table 1. Values of these variables were calculated and assigned to the atoms using the PETRA package of empirical methods.²²⁻²⁵

The next important task is to determine the size of the Kohonen map. With five physicochemical properties for describing the reaction center of the product of each reaction (see Table 1), the chosen Kohonen network needs five units in the input layer. Choosing a small number of neurons increases the danger that conflicts arise; i.e., reactions belonging to different types will end up in the same neuron. With a large Kohonen network the number of input data becomes too small for the adjustment of the weights to be of significance.

Systematic tests with Kohonen networks ranging in size from 3×3 to 50×50 showed that networks having between 1 and 3 times as many neurons as input reactions perform quite well. Specifically, we have chosen for the classification of 74 reactions 144 neurons arranged in a 12×12 grid. Figure 3 shows the Kohonen map obtained. Each small square represents a neuron; the number within a neuron is the index of the reaction mapped into it. The squares containing no numbers are called empty neurons.

Another form of the Kohonen map obtained for the same dataset is shown in Figure 4. In this map, the weight distance values between two neighboring neurons are calculated by eq 3 and are represented by vertical walls between two neurons. The height of the wall indicates the magnitude of the weight distance; only distances larger than 0.85 are indicated.

Three large clusters can be identified on the right-hand side,

(27) Ultsch, A.; Guimaraes, G.; Korus, D.; Li, H. *Proc. Transputer Anwender Treffen/World Transputer Congress TAT/WTC 93 Aachen*; Springer-Verlag: New York, 1993; pp 194-203.

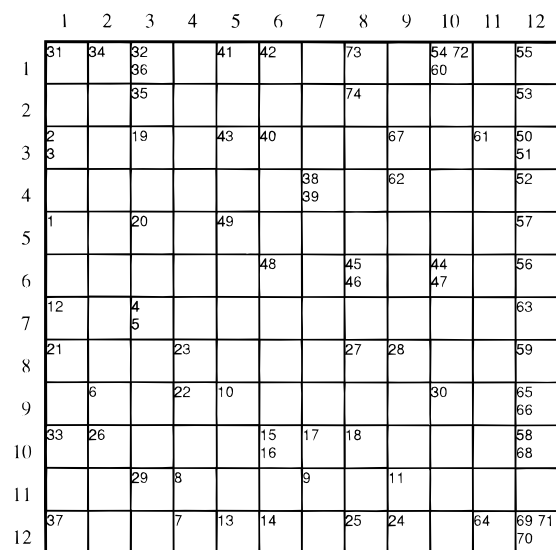


Figure 3. Kohonen map obtained for the classification of 74 reactions. The numbers within the map are reactions indices.

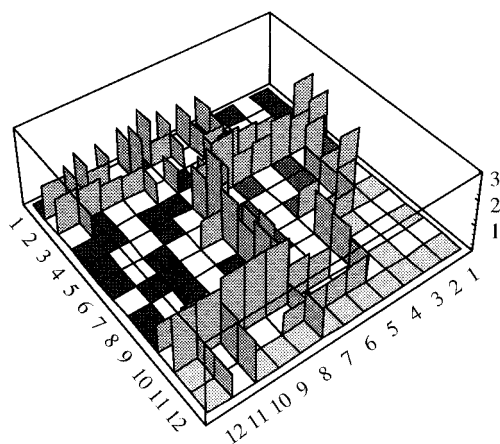


Figure 4. Same map as in Figure 3 showing the extra information on clusters detected using weight distance information. The wall indicates where the weight distance between adjacent neurons is larger than 0.85.

in the central part, and on the left-hand side of Figure 4. Several small clusters, or subclusters, are found at the upper part of the map.

The individual 74 reactions were intellectually assigned by chemists to different reaction types to confirm the correctness of the clusters detected on the basis of the above "weight distance rule". Three general reaction types were found, and they were marked in different levels of gray in identifying the mapping shown in Figure 5a: dark gray indicates nucleophilic aliphatic substitution of acyl chlorides, medium gray stands for acylation of C=C bonds, and light gray stands for acylation of arenes. In Figure 5a the weight distances between adjacent neurons are indicated by lines with thickness increasing with the weight difference.

An attempt was made to also assign empty neurons, neurons not having received reactions in order to make the separation into clusters, the grouping into reaction types, by the Kohonen network even clearer. To this effect a K-nearest neighbor technique²⁸ was used: all neurons in Figure 5a that had neighboring neurons assigned to only one class were also assigned to this class. This led to Figure 5b.

An even more completely assigned map can be obtained when also neurons are colored that have neighbors of different classes but with the number of neighbors of one class dominating. As this can be easily verified, we refrain from giving results.

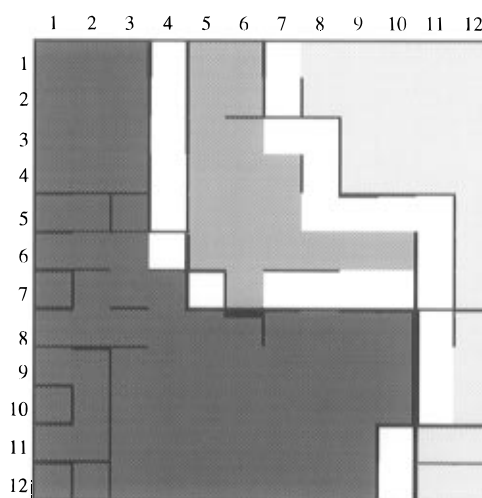
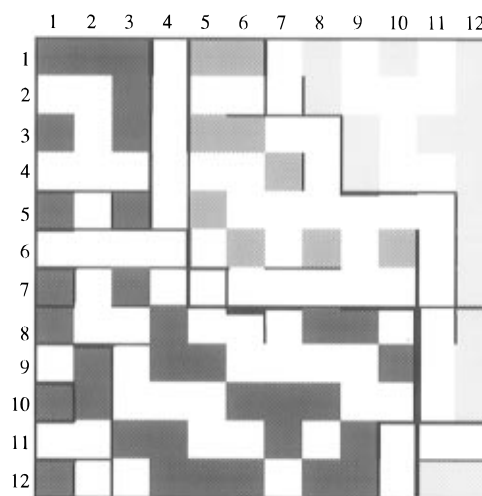


Figure 5. Indication of the weight distance between adjacent neurons by the thickness of lines separating them. The occupied neurons are marked in different gray levels: dark gray indicates nucleophilic aliphatic substitution of acyl chlorides, medium gray stands for acylation of C=C bonds, and light gray stands for acylation of arenes. (a) Only neurons with mapped reactions are marked. (b) Assignment of internal empty neurons to the corresponding clusters.

Figure 5b clearly shows how reactions belonging to the same reaction type are grouped together by the Kohonen network. Furthermore, it can also be recognized that the larger weight distances occur in the border region between two reaction types. The intellectual classification (shown in Figure 5) corresponds quite nicely to the automatic separation of the Kohonen map into regions by the weight distances as shown in Figure 4.

Thus, indeed, the weight distances of a Kohonen map offer a method for the automatic classification of reactions into chemically significant types.

Chemical Contents of the Kohonen Map. A detailed analysis of the mapping of reactions into the Kohonen network goes beyond the scope of this paper. It shows that the arrangement of the various reactions in the Kohonen map, the landscape of organic reactions, reflects a high degree of chemical information. We have already seen that mountain ranges separate different reaction types. With the next dataset we will see that lower elevations in mountain ridges, mountain passes, indicate smooth transitions between reaction types. Here, we will show that the further reactions are separated, the more different they are, even within the same reaction class. In wandering from one reaction across the landscape, a smooth

(28) See ref 16, pp 178–179.

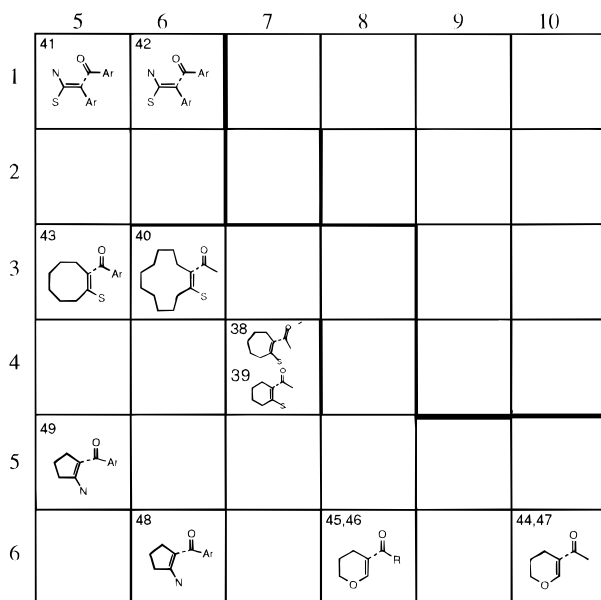


Figure 6. Expanded version of Figure 3. The bond made during the reaction is drawn as a dotted line. Ar = an aromatic ring. N = an N(III) atom. S = an S(II) atom.

change in reaction type can be observed. To limit the discussion, we only show in Figure 6 the cluster of acylations of alkenes.

In neuron (5, 1) we observe with reaction no. 41 an acylation of a double bond doubly activated by heteroatoms, a sulfur and a nitrogen atom. All reactions at the left-most part of this map, reaction nos. 41, 42, 43, 49, and 48, comprise acylations by aromatic acid chlorides.

Walking across the landscape from reaction no. 41 to the southeast, to reaction no. 39, we see that a single sulfur atom together with a carbon atom is sufficient for making this reaction occur. Progressing further southeast to reaction no. 46, we realize that an oxygen atom alone can also initiate this reaction type (however, as shown with the full reaction equation in Figure 7, a more activating catalyst is needed).

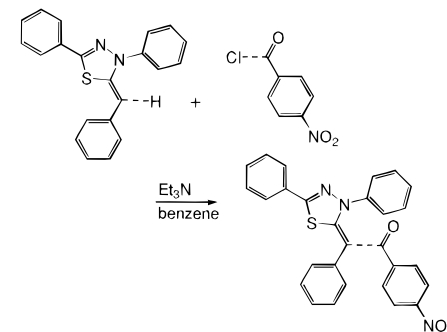
When we pursue directly south from reaction no. 41, to reaction nos. 49 and 48, it becomes clear that also a nitrogen atom (in conjunction with a carbon atom) is sufficient to allow acylation of a double bond. Figure 7 gives reaction nos. 41 and 46 as representatives of this section of the dataset in full reaction equations.

Let us now briefly turn our attention to the Friedel–Crafts acylation contained in the left-hand and upper part of the Kohonen maps of Figure 5. The weight distances indicate in this figure that there are three subclusters to be found in this reaction type, which indeed can be confirmed by closer inspection: At the bottom right-hand corner we find acylations to imidazoles and pyridinines (e.g., reaction no. 70); the cluster above contains acylations to the pyrrole system (such as reaction no. 65). The largest area within this reaction type, the one in the top right-hand corner, comprises acylations of benzene and naphthalene derivatives (e.g., reaction no. 50). A representative for each of these subclasses is given in Figure 7.

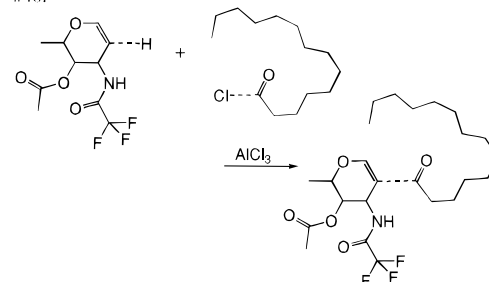
The largest number of subclusters is detected in the class of acylation at C_{sp^3} centers. This is a clear indication of the wide variety of reactions in this reaction type. For reasons of space we have to refrain from a detailed discussion. Rather, we only want to show that reactions at the outskirts of a reaction cluster constitute special classes. Reaction no. 37, mapped into the lower left-hand corner, separated by rather large weight distances from the rest of the reactions (see Figure 5), is an unusual case,

Additions to Alkenes

41:

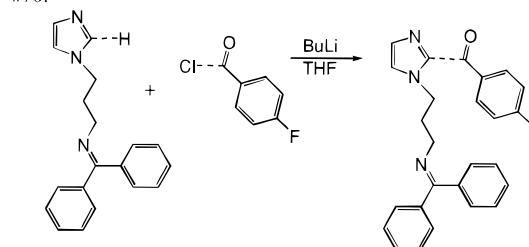


#46:

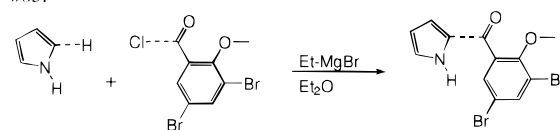


Friedel-Crafts Acylations

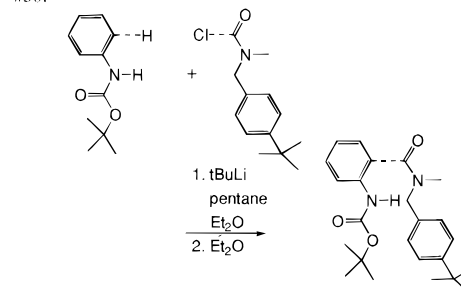
#70:



#65:



#50:



Special Reaction

#37:

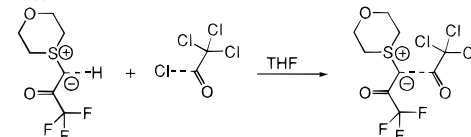


Figure 7. Some individual reactions from different (sub)clusters of the dataset of 74 reactions.

involving the acylation of a 1,4-oxathianium 3,3,3-trifluoro-2-oxopropyl ylide by trichloroacetic chloride (see Figure 7).²⁹

In summary, in this section we described how clusters, and thereby reaction types, can be detected in the Kohonen map on

(29) Wittmann, H. *Monatsh. Chem.* **1992**, *123*, 1207–1212.

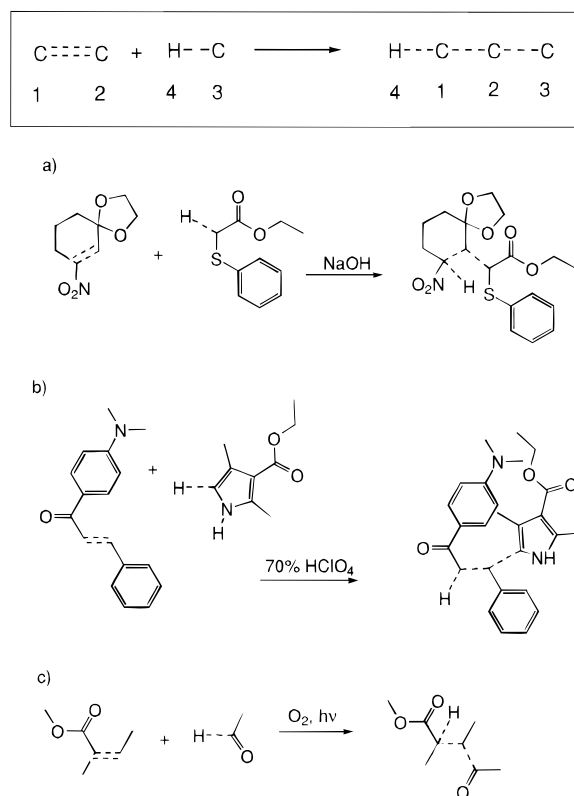


Figure 8. Reaction center common to the second dataset of 120 reactions. Some individual reactions showing different structural features around the reaction center are given. (a) Michael addition (no. 38). (b) Friedel-Crafts alkylation by alkene (no. 191). (c) Photochemical reaction (no. 94). The bonds in the reaction center are indicated by dotted lines.

the basis of an analysis of the differences in the weights of the neurons of the two-dimensional map. We showed that the classes such detected correspond to the intellectual classification and the assignment of the empty neurons by considerations of their neighbors. The arrangement of the reactions in such a two-dimensional map reflects fine details of high chemical significance. With this, we further explore the merit of this approach.

B. Classification of 120 Reactions. The second dataset was obtained by retrieving all those reactions contained in the 1992 edition of the ChemInform RX database²¹ that involve the addition of an H-C bond to a C=C bond to form a new C-C bond. One hundred twenty reactions were found. This reaction center and some individual reactions of this second dataset are shown in Figure 8.

In this study, physicochemical variables of the *educt* side where used in order to investigate the merit of our approach for building a knowledge base for *reaction prediction*. We used the same physicochemical variables as in the previous study, except the aromaticity indicator variable (α_i , q_{tot} , χ_σ , χ_π). Rather than considering these four electronic properties on all atoms of the reaction site, a selection was made by considering which effects are likely to be most important at which atoms. Specifically, σ - and π -electronegativities, χ_σ and χ_π , were considered at atoms C-1 and C-3, as were the total charges, q_{tot} , on atoms C-2 and C-3, as well as the effective polarizability, α_i , on atom C-3 (see Table 2). Here, too, the full set of variables on all atoms of the reaction site can be used, or a subset of variables can be automatically selected by statistical methods. However, we have intentionally chosen the variables on the basis of chemical intuition in order to show how a small set of variables deemed chemically significant can do the job of reaction classification.

Table 2. Seven Physicochemical Property Data Used To Characterize Each Reaction Center

electronic variable ^a	C=C		+	H-C		→	H-C-C-C			
	1	2		4	3		4	1	2	3
q_{tot}		×			×					
χ_σ	×				×					
χ_π	×				×					
α_i					×					

^a q_{tot} = total charge. χ_σ = σ -electronegativity. χ_π = π -electronegativity. α_i = effective atom polarizability.

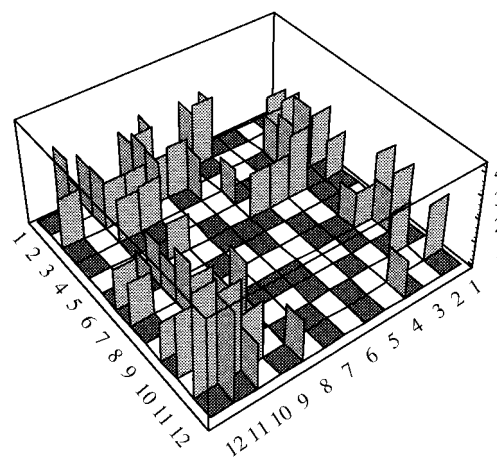
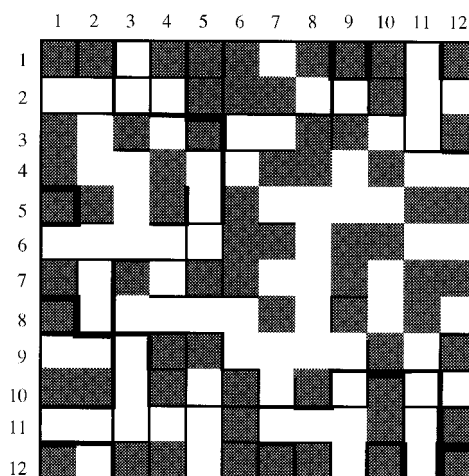


Figure 9. Simplified representation of the clusters automatically detected for the Kohonen map obtained for the classification of 120 reactions. The occupied neurons are marked in dark gray. The weight distances larger than 0.85 are indicated with black lines (a) or walls (b).

With seven physicochemical property data for describing each reaction under investigation (see Table 2), the Kohonen network needs seven units in the input layer. Again, the size of the network was chosen to be 12×12 , amounting to 144 neurons. This study with 120 reactions, more reactions than the previous investigation (74 reactions), must lead to more compression of information, forcing reactions closer together. However, we will see that the results are still reasonable, no reaction belonging to different types ending up in the same neuron. In our experience the classification results are good and quite stable when the number of neurons is between 1 and 3 times the number of reaction instances.

The results of the mapping of the reactions into the Kohonen self-organizing map are shown in Figure 9, which gives the weight distances to show the grouping of reactions into types. The map is broken up into a fairly large number of segments, indicating that quite a variety of reaction types can be found















reaction type	number of reactions	index of reaction
A.  Michael addition	75	1-14,24-84
B.  wrongly assigned reaction center	1	85
C.  Friedel-Crafts alkylation by alkene	18	15-23, 86-92, 115, 116
D.  electron transfer reaction	1	93
E.  photochemical addition of acyl radical to electron-deficient alkene	1	94
F.  photoinduced alkylation reaction	4	95-97,101
G.  hydrogen atom (radical) transfer reaction	3	98-100
H.  reaction with special reaction mechanism	1	102
I.  hydride abstraction reaction	3	103-105
J.  Nazarov reaction	5	106-110
K.  allylation of a keto ester followed by anti-Markovnikov addition of HBr (2-step-reaction)	1	111
L.  palladium-catalyzed oxyhexatriene cyclization (2-step reactions)	3	112-114
M.  condensation reaction	1	117
N.  photoinitiated radical addition of a crown ether	3	118-120

Figure 10. Intellectually assigned reaction types, associated symbols, and the corresponding number of reaction instances in the second dataset of 120 reactions.

having the reaction scheme shown in Figure 8. However, a large cluster of similar reactions can be noted in the center-right part of the map.

In order to better understand and identify the mapping of the reactions into the Kohonen feature map and to compare the clusters detected on the basis of the weight distance rule, the reactions were intellectually classified by chemists. More than 10 different reaction types were identified and assigned by inspection as shown in Figure 10. These reactions comprise several reaction types of high importance in chemical synthesis like Michael additions and Friedel-Crafts alkylations by alkenes, showing a wide variety of structural features around the reaction center. However, one also finds some more uncommon reaction types like the hydride abstraction reaction.

The identification of reaction types by the symbols shown in Figure 10 is chosen to mark the corresponding occupied neurons in the map shown in Figure 11.

In all cases but one, the reactions ending up in one neuron always come from the same type. This indicates that the classification using the method developed here based on electronic variables for describing the atoms of the reaction center and using the differences of a Kohonen network agrees with the classification performed by chemists on inspection. This shows the high chemical significance of the method introduced here. There was only one conflict situation: Neuron (12, 1) contains both a condensation reaction (no. 117)³⁰ and a reaction (no. 85)³¹ that on inspection was shown to have been coded in the database with a wrong reaction center.¹⁴ Thus, the method presented here even allows one to find errors in the database,

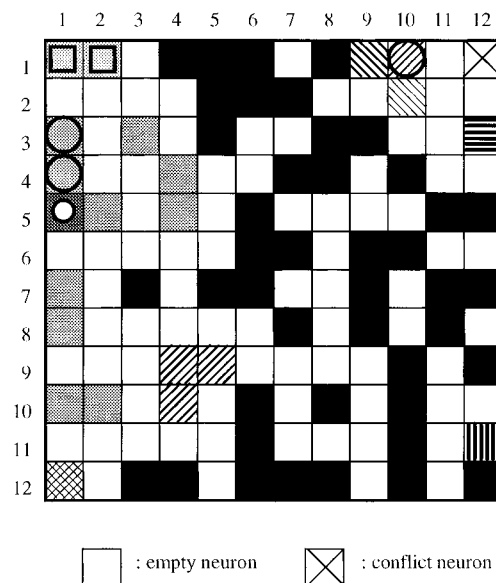


Figure 11. Kohonen map obtained for the classification of 120 reactions with the occupied neurons marked with different symbols as shown in Figure 10. (In the original screen output of the map, the different reaction types are indicated by different colors, giving a more vivid representation of the various reaction types.)

underscoring the high chemical significance of this classification method. It should be noted that the segmentation of the Kohonen map into individual clusters as indicated by the weight distances shown in Figure 9 agrees quite well with the intellectual assignment of reaction types given in Figure 11. For reasons of space a more detailed discussion is deferred to the Supporting Information.

Michael Additions. We only want to present some of the chemical contents of the map of the largest cluster, that of the Michael additions. Of 120 reactions, 75 reactions were classified by chemists as Michael additions, underlining the chemical importance of this reaction type. These 75 reactions were mapped into 43 neurons which shows, as we will now discuss, the large structural variety and scope of this reaction type.

Michael additions are activated by electron-withdrawing groups at both bonds directly involved in the reaction event, the C-H bond and the C=C double bond. We have counted the number of activating groups for each reaction at each of these bonds separately and have indicated this further differentiation of Michael additions in Figure 12. This figure shows that part of the Kohonen map with reaction instances belonging to Michael additions.

As Figure 12 shows, reactions that have only one (no marker), two (2Z), or three (3Z) electron-withdrawing groups at the reacting H-C bond are quite well separated. By the same token, those reactions that have two strongly electron-withdrawing groups (*) at the reacting C=C bond are well separated from those that have only one (no marker). Specifically, all eleven Michael additions with their reacting H-C bond activated by three electron-withdrawing groups—either three -COOEt groups or three chlorine atoms—were found at the center top perimeter of the Kohonen map.

The Kohonen network was able to realize, on the basis of physicochemical variables, that a C-H bond can be activated to undergo a Michael addition by substituents that can stabilize a carbanion through delocalization as well as by the combined inductive effects of three chlorine atoms. Furthermore, it

(30) Cheskis, B. A.; Ivanova, N. M.; Moiseenkov, A. M.; Nefedov, O. M. *Izv. Akad. Nauk SSSR, Ser. Khim.* **1990**, 9, 2025-2036.

(31) Cheskis, B. A.; Isakov, YA. I.; Novikov, A. V.; Moiseenkov, A. M.; Minachev, KH. M. *Izv. Akad. Nauk SSSR, Ser. Khim.* **1990**, 4, 902-905.

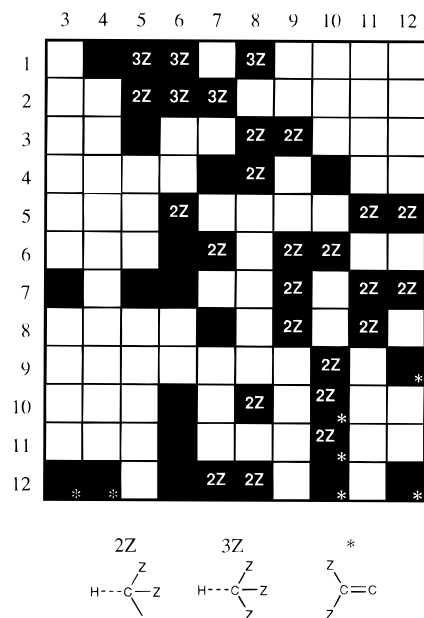


Figure 12. Functional groups at reaction sites (focus on Michael addition cluster). 2Z = two electron-withdrawing groups at H—C. 3Z = three electron-withdrawing groups at H—C. * = two electron-withdrawing groups at C=C.

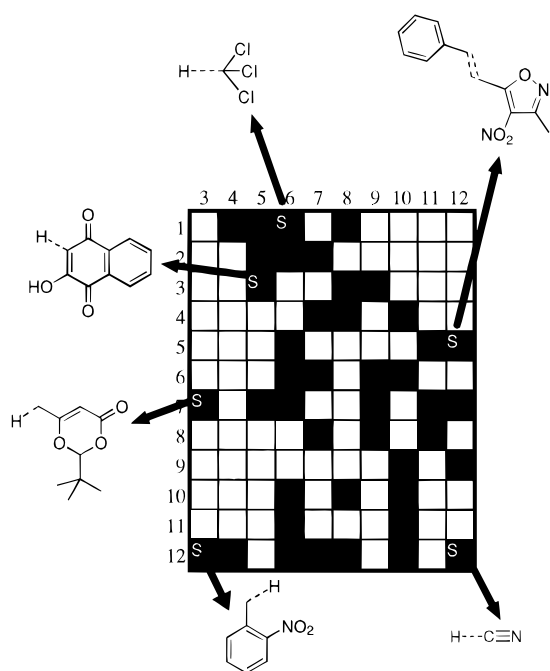


Figure 13. All special Michael additions (S) are found at the border of the Michael addition cluster.

demonstrates the inherent similarities among these structurally quite different reactions by putting them into the same region of the map.

The mapping of Michael additions by a Kohonen network also clusters reactions according to whether aromatic groups are activating the reacting C—H and C=C bonds. This is shown in the Supporting Information.

Next we turn our attention to those Michael additions which possess unusual activating functional groups on their reaction centers. These are marked with S in Figure 13 and further characterized by the structure of the reaction partner considered somehow unusual.

The reactions mapped into neurons (3, 7) and (3, 12) have none of the strongly electron-withdrawing groups mentioned above and necessary for a Michael addition directly bonded to

the C—H bond. However, such a group, a C=O or an NO₂ group, is contained beyond the directly bonded alkene or benzene system and can exert its influence through conjugation across the alkene or aromatic system. This is perceived by the procedure used to calculate π -electronegativities, underlining the importance of using such general physicochemical effects in reaction classification. It should be noted that a classification based on the types of atoms in α - and β -positions to the reacting bonds as presently used in commercial reaction databases such as the ones distributed by MDL cannot perceive such long-range effects.

The second type of special Michael additions is found in neuron (12, 12) in the lower-right corner of the Michael addition area (Figure 13). This is the only reaction in the entire dataset where the carbanion is formed at a C_{sp} center.³² All other Michael additions have the carbanion formed at a C_{sp}³ center. This is the reason why neuron (12, 12) shows particularly large weight distances to its neighbor neurons (Figure 9) and was not merged into the main body of the Michael addition area (Figure 11). The fact that a cyanide ion can also undergo a Michael addition thus clearly extends the knowledge so far found in the other Michael additions. It is therefore quite appropriate that this reaction was mapped into a neuron of its own, allowing easy discovery of this extra knowledge.

The reaction mapped into neuron (12, 5) has no strongly electronegative group directly bonded to the reacting double bond. However, such a group, a nitro substituent, can be found on the isoxazole system. It can exert its influence, as observed for the reaction mapped into neuron (3, 12) through conjugation. Thus, the vinylog principle can work both at the C—H and the C=C bonds.

The reaction projected into neuron (6, 1) has already been mentioned previously in conjunction with Figure 12. It shows that also inductively strongly electron-withdrawing groups can activate a C—H bond or undergo a Michael addition.

The reaction mapped into neuron (5, 3) shows that a C—H bond of hydroxynaphthoquinone can undergo a Michael addition reaction with methyl vinyl ketone (no. 24).³³ As can be seen from Figure 11, neuron (5, 3) is the only occupied neuron in the Michael addition cluster which has a direct connection to a neuron occupied by Friedel—Crafts alkylations (neuron (4, 4)). The unusual structural features of this reaction have also been detected and indicated by the thick lines surrounding neuron (5, 3) in Figure 9. It is therefore reasonable to deduce that the reaction of neuron (5, 3) might possess an unusual feature which is somehow between those of a typical Michael addition and a Friedel—Crafts alkylation reaction. This is indeed true! We know that, in a typical Michael addition, the carbon atom in the reacting H—C group has an sp³ hybridization state (or an sp hybridization carbon atom as just met with hydrogen cyanide in neuron (12, 12)). On the other hand, in a Friedel—Crafts alkylation by alkenes, the carbon atom in the reacting H—C group has an sp² hybridization state and is also a member of an aromatic ring system. Thus, on the basis of hybridization type of the carbon atom (C_{sp}²), this reaction is clearly related to Friedel—Crafts alkylation, whereas the base catalyst indicates this reaction to be a Michael addition. In fact, the reaction certainly occurs through a carbanion stabilized by two adjacent carbonyl groups and subsequent tautomerization.

It is interesting to note that all special Michael additions are mapped into outskirts of the Kohonen map of the Michael addition area. This underscores the special nature of these reactions. The analysis of the special Michael additions

(32) Griffiths, G.; Mettler, H.; Mills, L. S.; Previdoli, F. *Helv. Chim. Acta* **1991**, *74*, 309–314.

(33) Saitz, C.; Valderrama, J. A.; Tapia, R. *Synth. Commun.* **1990**, *20*, 3103–3114.

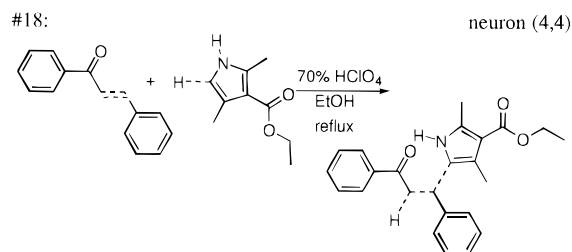


Figure 14. Reaction that can be considered as either a Michael addition or a Friedel–Crafts alkylation.

extended our knowledge on this important reaction type. Furthermore, it should be noted that the information on the fine relationships between different reactions such as those between reaction no. 24 and typical Michael additions as well as Friedel–Crafts alkylations can hardly be found by traditional clustering methods. The above facts stress the advantage of a two-dimensional representation of the relationships between different reactions.

Friedel–Crafts Alkylations. The next largest group of reactions that fall into the scheme treated in this section (see Figure 8) are Friedel–Crafts alkylations of aromatic compounds by alkenes. The dataset contained 18 reactions of this type which were mapped into 8 neurons located at the left-hand side of the Kohonen map. The weight distances (see Figure 9) split this area of Friedel–Crafts reactions into several subtypes, a fact that is supported by closer inspection of the individual reactions and discussed in the Supporting Information.

Here we only discuss the nine reactions mapped into neurons (3, 3), (4, 4), and (4, 5). One of these instances (no. 18) is shown in Figure 14; the others are variations of this theme.

It is interesting to note that these nine reactions, nos. 15–23, were carried out by the same research group,³⁴ which named them Michael additions. On the basis of the types of bonds involved (C_{sp^2} –H) and the catalyst used (acid), we prefer to call them Friedel–Crafts alkylations. It is remarkable that the Kohonen network—based on the chosen physicochemical variables for describing the reaction center—was able to map these questionable reactions into neurons which are located between the cluster of Michael additions and that of the Friedel–Crafts alkylations (see Figure 11). This indicates the characteristics of these reactions to be somehow related to both Michael addition and Friedel–Crafts alkylation. The unsupervised learning technique of the Kohonen network can stay clear of these semantic problems but rather concentrates on the inherent features of reactions.

Comparison with HORACE. The approach taken here by classifying these 120 reactions with a Kohonen neural network leads to 13 reaction classes. This can be deduced from looking at Figure 11 using the explanation by symbols contained in Figure 10.

HORACE generated 30 reaction classes for the same dataset¹¹ and thus was quite often not able to group reactions together that a chemist would consider as belonging to the same reaction type. This is largely due to the second phase of HORACE's classification, the one using the functional groups (topological features) at the reaction centers. Reaction instances are kept apart because they have different functional groups although these groups exert the same physicochemical effects. This situation is most clearly reflected by Michael additions. Among the 30 classes obtained by HORACE, 11 classes belong to Michael additions. HORACE is forced by the presence of quite a variety of functional groups around the reaction center to put these reactions into separate classes. In fact, such a limitation

is inherent in any reaction classification based on functional groups.

The Kohonen network approach developed here also perceives this large variety of Michael additions. It takes account of this fact by pushing the special Michael additions to the outskirts of the area reserved for this group of reactions. Thus, two things are gained: first these, reactions still belong to the area of Michael additions, but, second, their peculiar features are taken care of by having them at the borders of this area. This, again, underlines the advantage of a two-dimensional classification scheme, given by a Kohonen map, over a one-dimensional classification scheme, given by putting reactions into different classes.

In fact, the entire concept of classifying reactions into classes becomes questionable. Reactions are under the influence of many factors; changing some of the more important influences might lead to a gradual change in the reaction mechanism, sometimes to such a dramatic extent that a chemist would group this reaction into a different class. A two-dimensional landscape can take account of this situation by gradually shifting a reaction more and more away from its original position until it ends up in an area that belongs to a different reaction type. We have seen such a smooth connection between reaction classes in the case of certain Friedel–Crafts alkylations of aromatic compounds by alkenes that other people preferred to call Michael additions.

This makes it clear that it is more important to know at what position a reaction is located in the two-dimensional landscape rather than to assign it to a specific reaction class. We will see this when we make predictions on chemical reactions as shown in the next section.

Automatic Classification of Reactions. Once a Kohonen network has been trained, it can be used to make predictions. The weights of the Kohonen network contain in a condensed and generalized manner the information from the training set and are thus a representation of knowledge inherent in the entire dataset.

We illustrate the method and the scope and limitations for making predictions with an additional dataset having the same reaction center as the dataset of the 120 reactions (see Figure 8). Fifty-six such reactions were found in the Theilheimer reaction database.²⁰ For each of these reactions the physicochemical variables indicated in Table 2 were calculated and input into the Kohonen network previously trained with 120 reactions as discussed above. With such a map it becomes quite straightforward to assign the new reactions to their corresponding reaction types: When a test reaction falls into the area of a certain cluster, it can be assigned to the reaction type of that cluster. When, however, a reaction of the test set is mapped into an unassignable border empty neuron, the type of that reaction cannot be decided.

It was found that the trained network produces with the dataset of 56 reactions 38 correct, 12 undecided, and 6 wrong classifications. With only 68% (=38/56) correct predictions, the results appear somehow discouraging. However, this rather low prediction rate can be attributed to the deficiencies in the training set. Reactions of the test set that have to be left undecided had no counterpart in the original training set of 120 reactions.

Therefore, we can deduce that adding more representative reactions to the training set should significantly improve the prediction performance of the Kohonen network. The next experiment was designed to confirm this point.

The dataset of 56 reactions was split into 2 parts, 28 reactions with *even* indices and 28 reactions with *odd* indices. The Kohonen network used had exactly the same architecture as

(34) Löönd, R.; Neier, R. *Helv. Chim. Acta* **1991**, *74*, 91–102.

Table 3. Summary of the Prediction Results

no.	training set	test set	diction results		
			correct	undecided	wrong
I	120	56	38 (68%)	12	6
II	120 + 28 (odd)	28 (even)	24 (86%)	4	0
III	120 + 28 (even)	28 (odd)	21 (75%)	2	5

that used for the classification of the 120 reactions described above, but now it was trained with the 120 reactions plus the 28 reactions with either even or odd indices. The results are summarized in Table 3. It can be seen that adding more representative reactions into the training set significantly enhances the prediction ability of the Kohonen network. Specifically, the correct prediction rate for the test set of 28 reactions with odd indices increases from 68% to 75%; for the test set of 28 reactions with even indices, the correct prediction rate increases to 86%.

It must be stressed here again that the remaining undecided and wrongly predicted cases in this experiment still resulted from special reactions for which there are no similar reaction instances in the training sets.

According to the above discussion, we can conclude that (a) the test reactions mapped into a neuron are quite similar to the training reactions occupying the same neuron, (b) novel or unusual reactions can be found in empty neurons at the outskirts of the Kohonen maps, (c) the prediction ability of the Kohonen network can easily be enhanced through learning from new examples, (d) in order to make good predictions, it is extremely important to select a widely representative training dataset, and (e) the prediction reliability also heavily depends upon the reliability of assigning empty neurons.

Conclusions

The automatic extraction of chemical knowledge from reaction data is of great importance both in experimental and in computer chemistry. The approach developed in this paper is based on a Kohonen network and a set of physicochemical variables for the description of the reaction center. An analysis of the weight distances in the Kohonen networks leads to an automatic assignment of reactions to different classes. Application of these methods to 2 datasets consisting of 74 and 120 reactions, respectively, shows their potential for grouping a set of chemical reactions into classes, intuitively shown in a two-dimensional map. Such a map contains rich information amenable to interesting chemical interpretations as detailed in the studies reported here.

The self-organizing neural network approach to the analysis of a set of chemical reactions has several advantages.

(1) Various factors influencing chemical reactions can be selected as input into the Kohonen network. In this investigation we have chosen electronic effects exerted onto the reaction center. However, other effects coming from the starting materials or products of a reaction or even reaction conditions, such as reagents, solvent, and catalyst, can be used as input variables. This gives great flexibility to the approach and allows an analysis of such features on the grouping of reactions and their chemical significance. Calculation of the physicochemical variables used in the present study can directly be initiated with the RDFiles obtained from the ISIS Host reaction retrieval system.

(2) The computation time is short, and most of it has to be spent in the training phase; predictions with neural networks are rapid indeed. Thus, training the Kohonen network with the dataset of 120 reactions took 17.5 s on a SUN Sparc10 workstation. Predictions for the test set of 56 reactions took 0.5 s on this workstation.

(3) The size of the Kohonen network can be adjusted. For the two datasets studied here, the number of neurons (144) was about twice as large as that of the first dataset (74) or about of equal size to that of the second reaction dataset (120 reactions). This is a good compromise for making use of both the similarity perception and interpolation capabilities of the self-organizing neural network.

(4) Classification of a series of reactions into various reaction types is an important endeavor in our understanding of chemical reactions. However, a classification scheme is by its very nature one-dimensional and thus can hardly account for the richness of observations on chemical reactions. A two-dimensional landscape can much better reflect the result of the various influences on the course of a chemical reaction. It has space for showing different kinds of influences and thus different *kinds* of similarities by different *directions* in this landscape, and it has space for showing different *degrees* of similarities by giving different *distances*, smaller distances indicating stronger similarities. In this landscape, mountains—large weight distances in the map—separate different reaction types; saddles account for transitions between such reaction types.

(5) The method can be used for the automatic extraction of knowledge from reaction databases both for reaction prediction and for synthesis design. For reaction prediction, variables of the starting materials have to be input into the Kohonen network. Such a trained network can then be used for making predictions of the products of a reaction given the starting materials. This type of application was exemplified with the datasets of 120 reactions and 56 reactions. For synthesis design, variables for the products of the reactions have to be input into the Kohonen network. This approach can be extended to make predictions of strategic bonds in a molecule. This direction was illustrated with the dataset of 74 reactions.

Acknowledgment. We are grateful to the Alexander von Humboldt-Foundation for the support provided to L.C. in the form of a research fellowship. The Kohonen network simulator (KMAP) used in this work was originally developed by Dr. X. Li in this group. Thanks also go to Professor A. Ultsch for drawing our attention to the importance of the weight distances in a Kohonen network and to Dr. M. Wagener for implementing the simplified representation of weight distance information in a two-dimensional map. We thank MDL Information Systems Inc., San Leandro, CA, for providing us with an ISIS Host installation and, together with Fachinformationszentrum Chemie, Berlin, for giving access to the ChemInform RX reaction database.

Supporting Information Available: Text and Figures 15–23 describing in more detail the classification of 120 reactions (21 pages). See any current masthead page for ordering and Internet access instructions.

JA960027B