

11 The Reactivity of Chemical Bonds and the Classification of Chemical Reactions

learning objectives:

- the electronic and energetic effects that determine the polar reactivity of a chemical bond
- the possibility of using a neural network to classify bonds by reactivity (susceptibility to heterolysis)
- the importance of properly choosing the training data
- a classical method of data selection: “experimental (factorial) design”
- using a Kohonen network to select data to be used to train a **different** network
- how the Kohonen classifying map can be interpreted to reveal (possibly unsuspected) relationships within the data

11.1 The Problems and the Data

The prediction of the course and outcome of a chemical reaction is one of the fundamental tasks in organic chemistry. Since chemical reactions are initiated by the breaking of one or more bonds in a molecule, a knowledge about reactive bonds, that is bonds that will easily break, is indispensable for the prediction of chemical reactions. This is the theme of the first study reported in Sections 11.1 – 11.6.

Chemists have derived their knowledge about chemical reactions largely from individual observations, have ordered these individual reactions, have generalized their observations by building models or by making predictions by analogy. In this inductive learning process, the classification of reactions into reaction types plays a major role.

With the availability of large reaction databases with some of them comprising millions of reactions, the automatic classification of reactions becomes of major interest. For, this will allow the mining of knowledge from reaction databases, knowledge that can be used for reaction prediction systems and for computer-assisted synthesis design. The classification of reactions is the theme of the second study reported in Sections 11.7 – 11.8.

Organic reactions are largely governed by polar processes, which break a bond in such a way as to generate a positive charge on one atom and a negative charge on the other. Such a polar bond breaking, also called *heterolysis*, can occur for each bond in two ways (Figure 11-1); both possibilities were investigated in the first study.

This example follows the work of V. Simon and coworkers of the Model Laboratory for Computer Chemistry at the Technical University of Munich, who have trained a neural network for predicting the polar breaking of bonds (Reference 11-6). Given any single bond in an aliphatic organic compound, such a neural network should be able to predict whether this bond will break easily, and how the charges will be shifted onto the atoms of the bond (Figure 11-1).

A dataset of 29 molecules is chosen so as to cover the diverse structural variations of aliphatic molecules; these molecules contain 385 bonds capable of 770 potential polar bond breaking modes. Considering only unique single bonds (e. g., only one C–H bond of a methyl group) leaves 373 chemically different polar breaking modes.

From among these 373, a series of 149 breaking modes are selected that can rather unequivocally be classified by chemists into 43 reactive and 106 non-reactive ones.

Figure 11-2 shows four of the 29 molecules and the 11 single functional group bonds they contain; The arrow in each bond points to the atom that obtains the negative charge. Plain bent arrows indicate reactive bonds, i.e., bonds that can be broken easily, while arrows with X's indicate non-reactive bonds. All other bonds are unclassified.

The breaking of a chemical bond is influenced by a variety of energetic, electronic or steric effects, for example: charge distribution, the inductive, resonance, and polarizability effects, bond dissociation energies, etc.

Organic chemists often discuss these effects only in a qualitative manner, but in recent years a number of empirical quantitative methods has been developed for such factors. We choose seven of these to describe a chemical bond in this study:

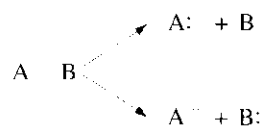


Figure 11-1: Two choices for the heterolysis of a chemical bond.

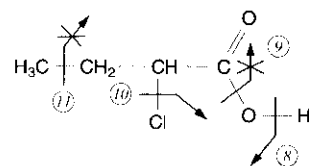
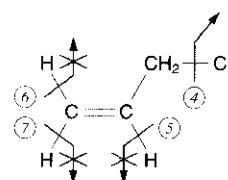
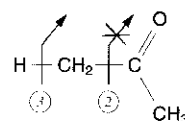
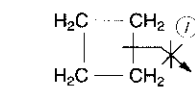


Figure 11-2: Four of the molecules and eleven of the bonds considered in this example. Plain bent arrows indicate reactive bonds whereas arrows that are crossed (X) show polar breaking modes that are difficult to achieve.

- the difference in total charge, Δq_{tot} ,
- the difference in π -charge, Δq_{π} ,
- the difference in σ -electronegativity, $\Delta \chi_{\sigma}$,
- a measure of bond polarity, Q_{σ} ,
- the amount of resonance stabilization, R^{\pm} , available for the charges generated upon heterolysis,
- the bond polarizability, α_b , and
- the bond dissociation energy, BDE .

Values of these variables are calculated and assigned to the bonds using a program package called PETRA (Parameter Estimation for the Treatment of Reactivity Applications; see Section 11.7, References 11-1 to 11-5).

Table 11-1 gives the values of these seven parameters for the eleven bonds shown in Figure 11-2.

bond	reac- tivity	Δq_{tot} [e]	Δq_{π} [e]	$\Delta \chi_{\sigma}$ [eV]	Q_{σ} [e]	R^{\pm} [1/eV]	α_b [Å ³]	BDE [kJ/mol]
1	–	0.00	0.00	0.00	0.00	4.02	5.46	236
2	–	–0.16	–0.03	–2.08	0.17	0.00	4.89	240
3	+	0.04	0.00	–0.56	0.06	8.09	3.85	412
4	+	0.24	0.00	–1.96	0.33	5.16	5.83	338
5	–	–0.13	–0.01	0.61	–0.12	3.43	5.06	461
6	–	–0.15	0.01	0.37	–0.11	0.00	4.22	456
7	–	–0.15	0.01	0.37	–0.11	0.00	4.22	456
8	+	0.45	–0.10	–1.56	0.44	7.35	3.60	437
9	–	–0.54	0.08	–1.48	–0.22	0.00	5.35	445
10	+	0.30	0.00	–1.32	0.31	7.48	6.93	336
11	–	–0.04	0.00	–0.34	0.01	0.00	6.16	362

Table 11-1: Values of the effects influencing reactivity for the chemical bonds shown in Figure 11-2.

Our task is now to relate these seven variables to the reactivity classification of a particular bond. As can be seen from Table 11-1, no single parameter suffices to separate reactive (+) and nonreactive (–) breaking modes. Classifying the reactivity of chemical bonds is clearly a multivariate problem.

11.2 Architecture of the Network for Back-Propagation Learning

We first approach the problem of classifying breaking modes as reactive or nonreactive by a two-layer neural network employing back-propagation learning. The number of input units is set to seven, the number of reactivity-controlling effects. These have different ranges, e.g., one from -0.2 to $+0.2$, another from 200 to 500; since the input units expect values between 0 and 1, each input value has to be scaled separately between its minimum and its maximum value. The output classification is coded as 0 for nonreactive breaking modes, or 1 for reactive ones.

A two-category (binary) classification can be achieved either by **two output neurons**, one for each class, or by **one output neuron** which is set to zero for one class, or to one for the other (Figure 11-3).

In the two-neuron case, the sum of the two output values is always 1.0, and the weights in the output layer have, in pairs, the same value with opposite signs (Figure 11-4).

We decide to work with a one-output neural network because an additional output does not provide any advantages.

Finally, we have to decide how many hidden neurons to use. We must always remember that having too many weights relative to the number of training data will probably lead to overtraining (Section 8.6.3); hence, the aim is always to work with as few neurons as possible.

A network with three hidden neurons turns out to be appropriate; our final architecture is $(7 \times 3 \times 1)$, with 28 weights including the connections to the bias (Figure 11-5).

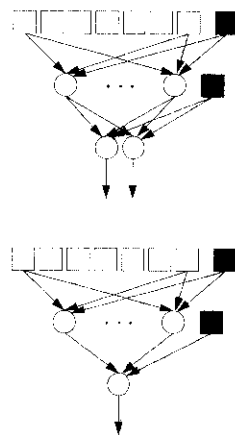


Figure 11-3: Two possible schemes for outputting a binary value.

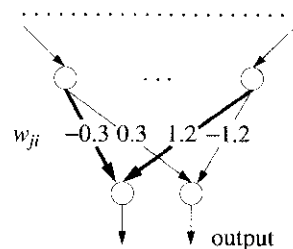


Figure 11-4: Two output neurons cause all pairs of weights connecting them to the level above to have the same value with opposite sign.

11.3 Using an Experimental Design Technique to Select the Training Set

Two measures are used to determine the quality of a network. The *recall* gives the percentage of correctly recognized objects from the training set after training is over. The ability for *prediction* gives the percentage of correctly classified objects from a test set which was not considered during training.

Next, we therefore have to divide the dataset into a training and a test set. If no clear criterion exists, you can choose the training data randomly; that's what we did initially. However, we will show in this section that random selection is not a very good strategy.

The initial dataset contained 116 bond breaking modes; we divided it into 58 for training and 58 for testing. However, to cover the measurement space better and to allow a joint comparison of the different methods for selecting the training set, we found it necessary to **extend** the dataset to 149 modes: 64 for training and 85 for testing. Only the results obtained from this larger dataset are discussed in this chapter.

Table 11-2 shows the results of the initial trial; we randomly selected ten training sets (64 modes each) and trained the $(7 \times 3 \times 1)$ neural network to recognize and predict reactivities of breaking modes with each of the ten sets separately.

The most important figure of merit for a multilayer neural network is the number of errors made in the test set. In predicted reactivity (yes or no), the number of wrong answers (out of 85) ranged from 3 to 12, with an average of 7.5.

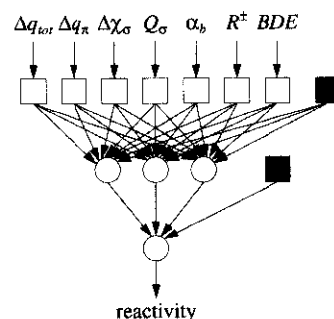


Figure 11-5: The network architecture for the prediction of the polar breaking of a bond.

training set	recall errors	prediction errors on a test set
1	1	7
2	0	3
3	0	12
4	0	10
5	0	6
6	0	7
7	1	5
8	0	9
9	0	7
10	0	5

Table 11-2: Recall and prediction ability obtained on the neural network after training with 10 **randomly** selected datasets.

The selection of the training set is so important that it is practically a separate area of investigation, not only for neural networks, but for any approach on extracting knowledge from data, and even more so when the data come from (expensive and lengthy) experiments. Hence, a field of specialization called “experimental design” or “factorial design” has been developed.

Most importantly, the training set must cover the variable space in the most representative possible manner. A standard method is to select variables that are thought to influence the system under investigation, and divide the range of values of each variable into two or three fixed levels or intervals; for example, in a three-level design, we set up a low, a medium, and a high level. If exact values of the variables are difficult to obtain, the entire range of each variable is divided, into, let's say, three fixed intervals (low, medium and high), which can be called levels just as well. After the levels are determined, the data are selected so that the set contains values for all variables, representing all **combinations** of levels.

For example, if you have two variables for which three levels (intervals) of possible values are determined, the data should be selected so that each of the $3^2 = 9$ subspaces is filled with an experiment or a data point (Figure 11-6).

In our present example, **seven** variables require $3^7 = 2,187$ bond breaking modes for a complete three-level experimental design. This is a bit large, so we must look for ways to reduce the number of data

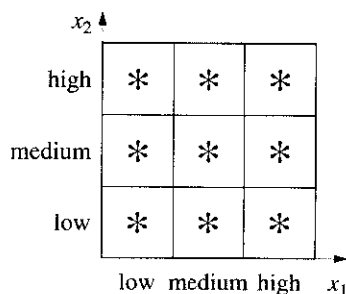


Figure 11-6: A complete three-level experimental design for a system with two variables (x_1 and x_2).

while still covering the information space as comprehensively as possible.

One way to do this is to reduce the number of variables to those representing the largest amount of **new** information. The correlation among the variables helps us make that choice: highly correlated variables carry similar information and are thus (more or less) redundant.

Based on an analysis of the correlation matrix shown in Table 11-3, four variables can be retained:

- the difference in σ -electronegativity, $\Delta\chi_{\sigma}$,
- the resonance stabilization of charges, R^{\pm}
- bond polarizability, α_b , and
- bond polarity, Q_{σ} .

	Δq_{tot} [e]	Δq_{π} [e]	$\Delta\chi_{\sigma}$ [eV]	Q_{σ} [e]	R^{\pm} [1/eV]	α_b [Å ³]	BDE [kJ/mol]
Δq_{tot}	1.00	-0.31	-0.11	0.84	0.12	-0.03	0.02
Δq_{π}		1.00	0.19	-0.41	-0.14	0.03	0.01
$\Delta\chi_{\sigma}$			1.00	-0.74	0.07	0.00	-0.01
Q_{σ}				1.00	0.06	-0.02	0.04
R^{\pm}					1.00	0.30	-0.28
α_b						1.00	-0.72
BDE							1.00

Table 11-3: Correlation matrix of seven variables for the breakability of bonds (based on 373 data).

With **four** variables, a three-level experimental design requires $3^4 = 81$ data on bond breaking modes.

Now, if we sort our 116 bonds into these 81 boxes, how many boxes stay empty? The answer is 41. If we add the remainder of the original 382, 20 more boxes are filled (with as-yet-unclassified bonds), leaving 21 of the subspaces completely empty. That is, none of the 373 breaking modes has a combination of the four variable values that would allow it to fall into any of these 21 subspaces.

We mentioned above that the 116-bond set was extended to 149; this was done to put data points into the 20 subspaces that can be filled with these data; as regards the 21 that can **not** be filled, we must assume that these subspaces do not have any chemical significance.

For the training set, we chose **one** breaking mode **from each** of the 60 occupied subspaces. Afterwards, we noted that four subspaces contained both reactive and nonreactive modes. Apparently, these subspaces are borderline cases, indicating regions where reactivity changes. In order to account for that, both a reactive and a nonreactive mode were selected from these four subspaces, increasing the training set to 64.

This, then, is the “best” training set for the back-propagation (7 x 3 x 1) neural network given the original set of 29 molecules.

Now, as we did in the random selection, we need ten different datasets for comparison, all selected according to the same criteria, ensuring that all of the 64 bond breakings cover all 60 subspaces. The results of ten identical learning procedures in the same (7 x 3 x 1) network show much better results than those obtained from the random selection; see Table 11-4 and Table 11-2.

no.	random selection		experimental design	
	recall	prediction	recall	prediction
1	1	7	0	1
2	0	3	0	3
3	0	12	0	4
4	0	10	0	3
5	0	6	0	4
6	0	7	0	2
7	1	5	0	3
8	0	9	0	5
9	0	7	0	6
10	0	5	0	3

Table 11-4: Errors in recall and prediction obtained after training with **random** training data, compared with results when the training data were determined by **experimental design**.

In results for the prediction set of 85 bond breaking modes, there are from one to six falsely classified modes, averaging 3.4. This is a remarkable improvement over the average 7.5 errors produced when a randomly selected dataset was used. This shows the importance and merit of experimental design techniques.

11.4 Application of the Kohonen Learning

The experimental design technique described above has two major disadvantages. First, we have to reduce the number of variables from seven to four; this is a rather tedious procedure, and necessarily leads to loss of information. Second, the choice of boundaries between the low, medium, and high intervals of the variable values is arbitrary and thus subject to user bias.

Is there another method that does not suffer from these problems? In this section we will show that a Kohonen neural network offers such an alternative.

In order to make the results comparable to those obtained in the previous Section, a (9 × 9) Kohonen network was chosen, containing 81 neurons that can be considered as equivalents of the “subspaces”; see Figure 11-7. When the dataset contains bonds with similar dependences on the seven controlling variables, they will map to the same neuron.

The network stabilizes after 30 training cycles, i.e., after all 373 polar bond breaking modes have been sent into the network 30 times. Six neurons are empty, 56 contain classified modes and 19 are occupied by nonclassified modes (Figure 11-8).

Remember, however, that a Kohonen network uses an unsupervised learning technique; since it does not use the class information when learning, it is remarkable that bonds of a particular classification end up in the same neuron. In all, 12 neurons carry only reactive bonds (and some unclassified ones), and 42 have only nonreactive (and some unclassified) bonds.

Only two neurons have conflicts, carrying both reactive and nonreactive bonds.

Also, the Kohonen map contains quite a lot of additional information that lends itself to chemical interpretation. This is further explained in Section 11.6.

Thus, the Kohonen network produces a basis for the selection of a training set. Again the (7 × 3 × 1) architecture is chosen and 10 different datasets are selected. This time only 56 bond breaking modes are necessary to cover the information space: one mode each from the neurons containing classified bonds.

One bond is chosen from each of the 56 neurons, plus two from those where conflicts occur. The only difference among the 10

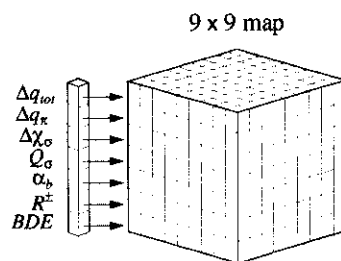


Figure 11-7: Kohonen network for mapping data on bond reactivity.

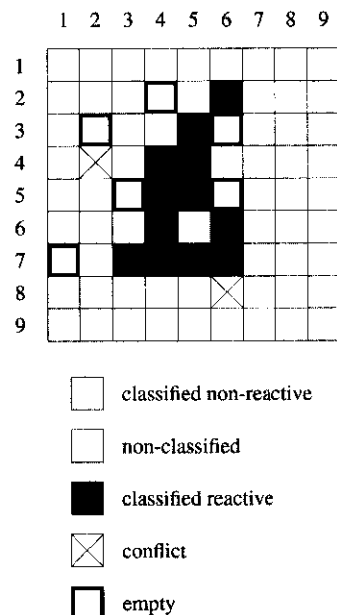


Figure 11-8: Kohonen map of the 373 polar bond breaking modes.

datasets is that when a neuron carries several modes of the same type, a different one may be selected in the next training set.

The results of this study, combined with the contents of Tables 11-2 and 11-4, are shown in Table 11-5.

There is one important difference between the selection of the training set by experimental design and by Kohonen mapping. The Kohonen network was trained by using **all seven** controlling variables, whereas in the experimental design study we had to cut the variables down to four in order to reduce the number of subspaces (which depends exponentially on the number of variables). Thus, one does not have to go through the tedious and time-consuming procedure for reducing the number of variables.

no.	random selection		experimental design		Kohonen network	
	recall	prediction	recall	prediction	recall	prediction
1	1	7	0	1	0	2
2	0	3	0	3	0	2
3	0	12	0	4	0	1
4	0	10	0	3	0	0
5	0	6	0	4	0	2
6	0	7	0	2	0	0
7	1	5	0	3	0	0
8	0	9	0	5	0	0
9	0	7	0	6	0	0
10	0	5	0	3	0	2

Table 11-5: Errors in recall and prediction obtained after training with random data, with those determined by experimental design, and by Kohonen mapping.

11.5 Application of the Trained Multilayer Network

The chemical importance of this example is that, by applying a **combination** of neural network techniques (a one layer (7 x 9 x 9) Kohonen network for selecting the training set, and (7 x 3 x 1) multilayer neural network with back-propagation learning), we end up with a trained network that is able to predict which single bonds will preferentially break in a polar manner and which will not **in any** aliphatic molecule. In addition, it even indicates in which direction the charges are shifted upon heterolysis.

Let's test this with a relatively complex molecule containing several functional groups. For all the bonds in this molecule, the seven controlling parameters were calculated for the two polar bond breaking modes. With 32 bonds in the molecule, this amounted to 448 ($= 32 \times 2 \times 7$) variables, which are input into the trained ($7 \times 3 \times 1$) network. From among the 64 bond breaking modes (two for each of the 32 bonds), only nine are found to be reactive; these are indicated in Figure 11-9. The corresponding unscaled values of the controlling parameters are shown in Table 11-6.

The network correctly predicts a high reactivity for the deprotonation of hydroxyl (9), of $-NH$, (7) and of methylene in α -position to the aldehyde group (3 and 4); the loss of a hydroxyl anion (8) is correctly predicted to be easy. It also predicts loss of the proton at the aldehyde group (1); this is usually not observed, since most bases are also strong nucleophiles, and would rather make a nucleophilic attack at the carbonyl. However, the predicted deprotonation does occur in formic esters, which contain the $H-C=O$ group.

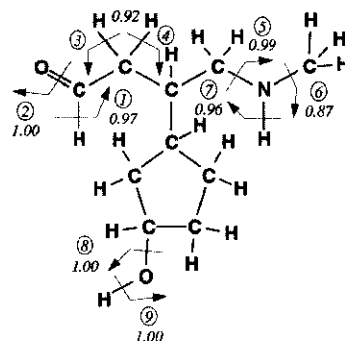


Figure 11-9: Predictions of bond breaking made by the ($7 \times 3 \times 1$) neural network trained as described above.

bond	Δq_{tot} [e]	Δq_{π} [e]	$\Delta \chi_{\sigma}$ [eV]	Q_{σ} [e]	R^{\pm} [1/eV]	α_b [Å ³]	BDE [kJ/mol]
1	-0.04	-0.03	-2.13	0.21	0.00	4.80	356
2	0.48	0.06	-3.03	0.60	3.69	4.81	425
3	0.03	0.00	-0.66	0.07	8.09	6.28	397
4	0.03	0.00	-0.66	0.07	8.09	6.28	397
5	0.32	0.00	-0.26	0.19	3.81	7.39	331
6	0.31	0.00	-0.39	0.20	0.00	5.97	343
7	0.44	0.00	-0.31	0.24	0.00	5.37	382
8	0.45	0.00	-0.84	0.36	7.68	6.19	384
9	0.60	0.00	-0.88	0.42	0.00	3.93	437

Table 11-6: Values of the controlling parameters for the nine reactive bond breaking modes in the test molecule of Section 11.5. The modes are identified in Figure 11-9.

Although the network has been trained only for single bonds, it is able to further **generalize** and also assign the correct reactivity to the $C=O$ double bond (2). The breaking of the two $C-N$ bonds (5 and 6) is also considered feasible. These bonds are quite polar, but would need further activation in order to react.

The study of a series of other molecules has shown the overall correctness of most of the predictions on chemical reactivity. It is

tempting to take the values output by the network as **probabilities** for bond breaking – as a quantitative measure of reactivity. However, this would certainly amount to an overinterpretation of the results; the network has been trained for classification and not for modeling. However, it shows that the border between classification and modeling is not hard and fast; had we used quantitative values for the bond reactivity instead of a mere binary classification, we could have come up with a model that is able to predict quantitative reactivity values.

11.6 Chemical Significance of the Kohonen Map

We mentioned in Section 11.4 that the Kohonen map for 373 bond breaking modes contains additional information capable of an interesting chemical interpretation.

Figure 11-10 is just a duplicate of Figure 11-8. Since not all the bonds were classified as reactive or nonreactive, unclassified bonds are spread throughout the network; they are indicated only in those neurons that contain neither reactive nor nonreactive modes, nor conflicts.

It can be seen that the reactive modes form a cluster in the center of the map (Figure 11-10). This is an indication that the self-organization that occurs during Kohonen learning perceives the similarity of **certain types** of modes and puts them into the same neurons. The Kohonen network even goes beyond that by recognizing the similarity of all reactive modes and putting them into neighboring neurons, thus forming the observed clustering of neurons with reactive modes.

There is one intruder: a bond classified unreactive by chemists occurs in the cluster of reactive modes (neuron at column 5, row 6 in Figure 11-10). This bond breaking mode is shown in Figure 11-11. The polar breaking of this bond is not observed in this molecule, and thus its classification as nonreactive is justified **for this molecule**.

However, in compounds having a bulkier group instead of C_2H_5 , the breaking of this bond is observed. Thus, the bond breaking shown in Figure 11-11 can occur when it is not superseded by other types of reactions. In effect, this has to be considered as a **potentially** reactive

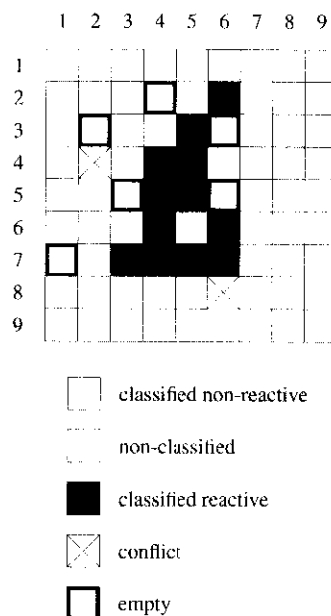


Figure 11-10: Kohonen map of the 373 polar bond breaking modes (cf. Figure 11-8).

bond, which makes reasonable its occurrence in the cluster of reactive bond breaking modes.

Neurons activated by nonreactive bond breaking modes and those activated by reactive ones touch each other in only a few places. This is a further indication of the success of Kohonen learning in differentiating reactive from nonreactive modes. The cluster of reactive bonds is surrounded by neurons mapping conflicts or nonclassified modes. The conflicts are a consequence of the transition from reactive to nonreactive; the nonclassified modes indicate how cautious chemists are about classifying bonds in doubtful cases.

After discussing the general features of the Kohonen map, let's take a closer look at the bond breaking modes ending up in individual neurons. In order to show some features in more detail, we have shifted the map of Figure 11-10 two columns left and one row down (which represents cutting the toroidal mapping surface at different places).

First, let's look at the reactive bonds (shaded cluster). All carbon-heteroatom breaking modes are at the right-hand side of this cluster, starting at the top with carbon-iodine and carbon-bromine bonds, then carbon-chlorine, carbon-oxygen and carbon-nitrogen bonds. This "top-to-bottom" sequence shows a clear tendency of decreasing polar reactivity.

The left-hand side of the shaded cluster shows modes that correspond to the dissociation of a proton. The more acidic O-H and N-H are more to the center of the cluster, and the less acidic C-H are towards the outskirts.

Second, the nonreactive C-H- and C-C-bonds are distributed over a wide area of the Kohonen map, because in the test set of molecules they have such differing first- and second-neighbors. A discussion of these small variations goes beyond the scope of this book (cf. Reference 11-6).

One more major feature of this map should be pointed out: the reactive modes are in the **lower left-hand corner**, while in the vicinity of the **upper right-hand corner** are those cases where a polar bond is broken **against** its inherent polarity (and thus are particularly unlikely to occur).

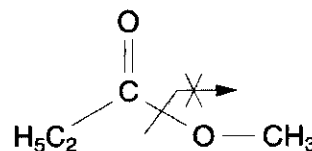


Figure 11-11: A bond considered unreactive by chemists that ends up in the cluster of reactive bond breaking modes.




	1	2	3	4	5	6	7	8	9
1	$\text{C} \times \text{C}$ (4)	$\text{C} \times \text{C}$ (1)	$\text{C} \times \text{C}$ (3)				$\text{Cl} \times \text{C}$ (1) $\text{Br} \times \text{C}$ (1)	$\text{C} \times \text{H}$ (1)	$\text{C} \times \text{H}$ (8)
2	$\text{C} \times \text{C}$ (2)			$\text{C} \times \text{C}$ (1)	$\text{C} \times \text{C}$ (2)	$\text{I} \times \text{C}$ (1)	$\text{O} \times \text{C}$ (1) $\text{N} \times \text{H}$ (1)	$\text{HO} \times \text{C}$ (2)	
3	$\text{C} \times \text{C}$ (2)			$\text{C} \rightarrow \text{I}$ (2) $\text{C} \rightarrow \text{Br}$ (1)		$\text{N} \times \text{C}$ (1)	$\text{Cl} \times \text{C}$ (1)	$\text{O} \times \text{H}$ (4)	$\text{O} \times \text{C}$ (2)
4	$\text{C} \times \text{C}$ (1)		$\text{C} \rightarrow \text{Br}$ (1)		$\text{C} \times \text{C}$ (4)	$\text{C} \times \text{C}$ (1)	$\text{N} \times \text{C}$ (2)		
5		$\text{H} \rightarrow \text{N}$ (1) $\text{C} \rightarrow \text{O}$ (2)	$\text{C} \rightarrow \text{Cl}$ (1)	$\text{C} \times \text{N}$ (1)	$\text{C} \times \text{C}$ (1)	$\text{C} \times \text{C}$ (2)	$\text{C} \times \text{C}$ (4)	$\text{C} \times \text{C}$ (1)	$\text{H} \times \text{C}$ (1) $\text{H} \times \text{CCl}_2$ (1)
6		$\text{H} \rightarrow \text{O}$ (10)	$\text{C} \rightarrow \text{OH}$ (5)		$\text{C} \times \text{C}$ (1)			$\text{C} \times \text{C}$ (1) $\text{H} \times \text{C}$ (1)	
7	$\text{H} \times \text{C}$ (4)	$\text{H} \rightarrow \text{N}$ (1)	$\text{C} \times \text{O}$ (1)	$\text{C} \rightarrow \text{Cl}$ (2)  (1)			$\text{C} \times \text{C}$ (2)	$\text{H} \times \text{C}$ (4)	$\text{H} \times \text{C}$ (12)
8	$\text{H} \rightarrow \text{C}$ (3) CO $\text{H} \rightarrow \text{C}$ (1) NO_2	$\text{H} \rightarrow \text{C}$ $\text{C} =$ (3)	$\text{C} =$ $\text{H} \rightarrow \text{C}$ (3) $\text{C} =$	$\text{C} \times \text{C}$ (2)  (1)	$\text{C} \times \text{C}$ (3)	$\text{C} \times \text{H}$ (1)			
9	$\text{C} \times \text{H}$ (4)		$\text{C} \times \text{H}$ (2)	$\text{C} \times \text{C}$ (2)  (1)	$\text{C} \times \text{C}$ (1)		$\text{C} \times \text{C}$ (1)	$\text{C} \times \text{C}$ (1)	$\text{C} \times \text{H}$ (7)

Figure 11-12: Expanded and shifted version of Figure 11-10. Reactive bond types are drawn on a shaded background; the arrow indicates the polarization of the electron pair. The numbers in parentheses are the numbers of bonds mapped onto this particular neuron.

11.7 Classification of Reactions: The Data

The results reported in Section 11.6 show that there is a lot of chemical significance in the projection of bonds, as represented by physicochemical descriptors, into a Kohonen map. This observation can be taken one step further by considering all bonds that are broken in a chemical reaction, the reaction center, and projecting them into a two-dimensional Kohonen map. A two-dimensional arrangement of reactions has great advantages for the comparison of chemical reactions (Figure 11-13): the distance between two reactions in such a map can represent the degree of similarity; different directions in such a map can express different types of similarities.

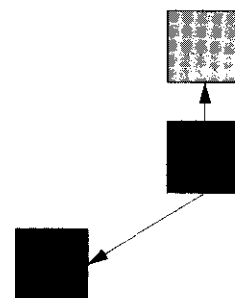
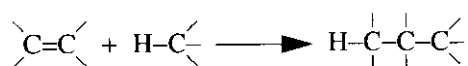


Figure 11-13: The representation of the degree and the type of similarity of reactions in a Kohonen map.

Clearly, it is rather easy to classify reactions that have different reaction centers, that have different types of atoms and bonds involved in the electron rearrangement during a reaction. But how about reactions that break and make the same type of bonds during a reaction but which chemists would classify into different reaction types because of the functional groups and substructures that are adjacent to the reaction center but are not directly involved in the electron relocation during a reaction?

In order to classify such reactions we have, as a typical example, retrieved reactions that involve the addition of a C–H bond to a C=C as indicated by the following scheme:



All those reactions, that contain this reaction center were retrieved from the 1992 edition of the ChemInform RX reaction database. Altogether, 120 reactions were obtained that a chemist would classify into different reaction types such as Michael additions, Friedel-Crafts alkylation of aromatic compounds by olefins, radical reactions, etc.

The next step is the choice how to represent these reactions. The most important driving forces of chemical reactions are electronic effects and, therefore, we performed calculations on charge distribution, inductive effects, as represented by orbital electronegativities, as well as effective polarizabilities by the empirical methods collected in the PETRA program package in order to account for the influence of functional groups onto the reaction site. Specifically, total charges, q_{tot} , σ - and π -electronegativities, χ_{σ} and χ_{π} , as well as effective polarizabilities, α_i , were calculated for all atoms of the reaction site. From all these variables chemists selected those that were deemed to be of importance for the reaction type under study, ending up with the seven physicochemical descriptors shown in Table 11-7.

11.8 Classification of Reactions: Results

The previous section has shown how the chemical reactions of the chosen data set are represented as points in a seven-dimensional space of electronic descriptors. A Kohonen network of planar topology with 12 x 12 neurons is chosen to project these 120 reactions from the seven-dimensional space into two dimensions.

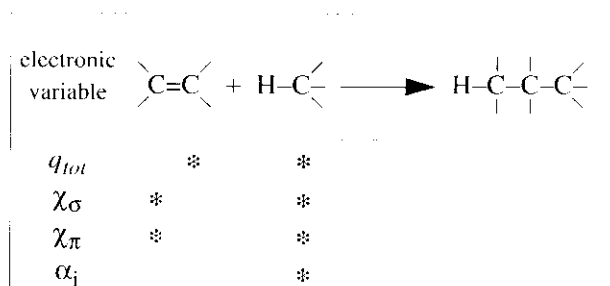


Table 11-7: Physicochemical variables used to describe the reaction site. An asterisk indicates the atom for which the corresponding descriptor was used.

How was the size of the network chosen? For studies of this type, which is basically a similarity perception, it is a good starting point to choose approximately as many neurons (here 121) as there are objects (here 120) to investigate. Reducing the number of neurons forces more objects into the same neuron, reducing somewhat the resolution of the Kohonen network.

The resulting map with the individual reactions of the data set identified by their number is shown in Figure 11-14. How can this map be interpreted? In order to evaluate the chemical significance of this mapping, all reactions of the data set were intellectually classified into reaction types by chemists. Altogether, 14 reaction types were identified containing between 1 and 75 reactions as members. This assignment of reaction types was used for coloring the Kohonen map of Figure 11-14 to give Figure 11-15.

This Figure shows that reactions of the same type form coherent areas in this map. Thus, the Kohonen learning has identified on the basis of the seven physicochemical variables reaction types much in the same manner as they were attributed by a chemist.

The next question is now, how do we know where one reaction type changes into another one in this map? This question can be addressed by an analysis of the weight differences between weights of adjacent neurons. The weights do not have a uniform distribution throughout the Kohonen network. Rather, one can identify locations in the Kohonen network where the weight differences between adjacent weights are larger than in other places.

Figure 11-16 shows the weight differences that exceed the selected threshold of 1.5 as walls, with higher walls indicating larger weight differences. It can be discerned from Figures 11-15 and 11-16 that the barriers in the weight differences between adjacent neurons coincide

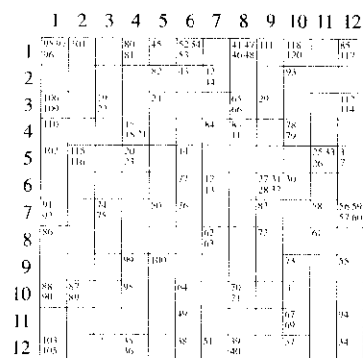


Figure 11-14: The Kohonen map for the 120 reactions containing the reaction center shown in Table 11-7 and described by seven physicochemical variables.

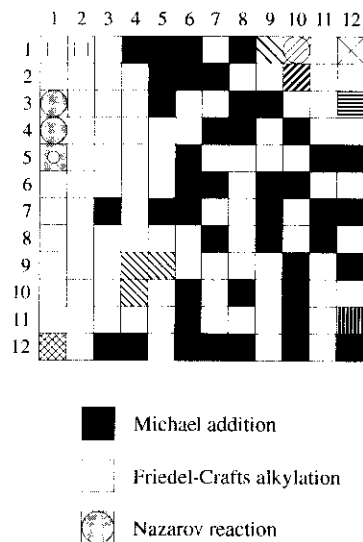


Figure 11-15: The Kohonen map of Figure 11-14, now with reactions identified by their respective reaction types. Three of the reaction types are identified by names.

with transitions from one reaction type to another. Thus, clusters of reactions can be found in the Kohonen maps of data sets of reactions that comprise reaction types.

A detailed discussion of the chemical inferences of these maps of chemical reactions and reaction types that truly form landscapes with mountains, passes, and valleys of weight differences goes beyond the scope of this book. The interested reader is suggested to contact the corresponding publications in journals.

Furthermore, methods have been developed that allow the automatic assignment of reaction types, thus superceding the time-consuming intellectual identification of reaction types.

Suffice to say that these landscapes of chemical reactions allow

- the identification of reaction types,
- the clustering of reaction databases and of hits from reaction searches,
- the location of transitions between reaction types,
- the definition of the scope of a reaction type,
- the identification of special reactions,
- the extraction of knowledge from reaction databases for reaction prediction,
- the mining of reaction databases for synthesis design.

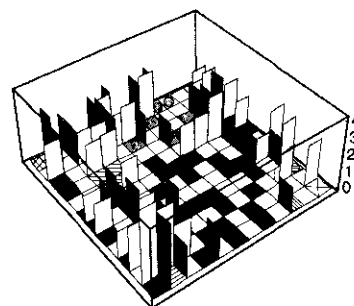


Figure 11-16: The Kohonen map of Figure 11-15 with the weight differences that exceed the threshold 1.5 shown as walls.

11.9 References and Suggested Readings

- 11-1. J. Gasteiger, M. Marsili, M. G. Hutchings, H. Saller, P. Löw, P. Röse and K. Rafeiner, "Models for the Representation of Knowledge about Chemical Reactions", *J. Chem. Inf. Comput. Sci.* **30** (1990) 467 – 476.
- 11-2. J. Gasteiger and M. Marsili, "Iterative Partial Equalization of Orbital Electronegativity – A Rapid Access to Atomic Charges", *Tetrahedron* **36** (1980) 3219 – 3228.
- 11-3. M. G. Hutchings and J. Gasteiger, "Residual Electronegativity – An Empirical Quantification of Polar Influences and its Application to the Proton Affinity of Amines", *Tetrahedron Lett.* **24** (1983) 2541 – 2544.
- 11-4. J. Gasteiger and M. G. Hutchings, "Quantification of Effective Polarisability. Applications to Studies of X-Ray Photoelectron Spectroscopy and Alkylamine Protonation", *J. Chem. Soc. Perkin 2* (1984) 559 – 564.
- 11-5. J. Gasteiger and H. Saller, "Berechnung der Ladungsverteilung in konjugierten Systemen durch eine Quantifizierung des Mesomeriekonzeptes", *Angew. Chem.* **97** (1985) 699 – 701; "Calculation of the Charge Distribution in Conjugated Systems by a Quantification of the Resonance Concept", *Angew. Chem. Int. Ed. Engl.* **24** (1985) 687 – 689.
- 11-6. V. Simon, J. Gasteiger and J. Zupan, "A Combined Application of Two Different Neural Network Types to the Prediction of Chemical Reactivity", *J. Am. Chem. Soc.* **115** (1993) 9148 – 9159.
- 11-7. L. Chen and J. Gasteiger, "Organische Reaktionen mit Hilfe neuronaler Netze klassifiziert: Michael-Additionen, Friedel-Crafts-Alkylierungen durch Alkene und verwandte Reaktionen", *Angew. Chem.* **108** (1996) 844 – 846; "Organic Reactions Classified by Neural Networks: Michael Additions, Friedel-Crafts Alkylations by Alkenes, and Related Reactions", *Angew. Chem. Int. Ed. Engl.* **35** (1996) 763 – 765.
- 11-8. L. Chen and J. Gasteiger, "Knowledge Discovery in Reaction Databases: Landscaping Organic Reactions by a Self-Organizing Neural Network", *J. Am. Chem. Soc.* **119** (1997) 4033 – 4042.

- 11-9. The ChemInform RX reaction database is produced by FIZ CHEMIE, Berlin, Germany and marketed by MDL Information Systems, Inc., San Leandro, CA, USA.
- 11-10. Information on the methods in the PETRA package can be found on: <http://www2.ccc.uni-erlangen.de/software/petra/>
- 11-11. A. Ultsch, G. Guimaraes, D. Korus and H. Li, "Proc. Transputer Anwendertreffen/World Transputer Congress TAT/WTC 93", Aachen, Springer Verlag, New York, USA, 1993, pp. 194 – 203.
- 11-12. H. Satoh, O. Sacher, T. Nakata, L. Chen, J. Gasteiger and K. Funatsu, "Classification of Organic Reactions: Similarity of Reactions Based on Changes in the Electronic Features of Oxygen Atoms at the Reaction Sites", *J. Chem. Inf. Comput. Sci.* **38** (1998) 210 – 219.