Analysis of the Reactivity of Single Bonds in Aliphatic Molecules by Statistical and Pattern Recognition Methods

J. Gasteiger* and K.-P. Schulz

Organisch-Chemisches Institut, Technische Universität München, Lichtenbergstrasse 4, D-8046 Garching, Germany

C. Kredler

Institut für Mathematik und Statistik, Technische Universität München, Arcisstrasse 21, D-8000 München 2, Germany

Received July 6, 1992

The polar breaking of a bond is the initiating step in many organic reactions. The factors influencing the heterolysis of a bond are investigated by a series of statistical and pattern recognition methods. The power of a variety of methods—including principal component analysis, linear discriminant analysis, cluster analysis, *k*-nearest-neighbor analysis, and logistic regression analysis—for understanding and predicting the reactivity of bonds based on electronic and energy parameters is elaborated. In particular, we show how information on whether a series of bonds will undergo heterolysis or not can be used to derive a function *quantifying* polar reactivity. In the end, an equation is developed that allows the prediction of the polar breaking of bonds for a wide range of aliphatic structures.

The prediction of the course of chemical reactions is a fundamental task in organic chemistry. It asks for the search for reactive bonds in a molecule and then for decisions about which of the potentially reactive bonds are reacting preferentially. In such a way, reaction mechanisms are derived by successively pinpointing the bonds that are broken and made in a reaction.

The task would be greatly facilitated if functions were available that allowed calculation of the reactivity of each bond in a molecule. Then, the bond with the highest reactivity value would be the one to be selected as being broken or made. Such reactivity functions could be derived from quantitative experimental reactivity data such as reaction rates, activation energies, or enthalpies of activation, by statistical methods.

We have taken such an approach for fundamental gas-phase reactions.[1-3] Modern experimental techniques like ion cyclotron resonance measurements provide data on proton affinities or on gas-phase acidities of high numerical accuracy. These values can be used for the derivation of equations by such simple statistical methods like multilinear regressions analysis (MLRA).

However, the situation is quite often not as favorable. Usually, for a reaction type, not enough numerical data that quantify chemical reactivity and vary all influencing variables independently (substrate, reagent, solvent, reaction, temperature, etc.) are available to provide a statistically balanced set that allows a MLRA. In most cases, even for such fundamental organic reactions like nucleophilic aliphatic substitution, too many variables are changed simultaneously. These many variables would ask for extra data points which have not been measured by experiments.

This lack of numerical data makes any endeavour for further understanding of chemical reactivity so difficult. On the other hand, an organic chemist can quite often predict by inspection which bonds in an organic molecule are reactive.

In this paper we explore how information on whether a bond is reactive or not can be used to derive methods for predicting chemical reactivity. In particular, we investigate whether this *qualitative* information can be used to derive functions that can *quantify* reactivity.

In other words, we want to express the relationship between structure and chemical reactivity in *explicit* mathematical functions. Thus, this study complements attempts to store this relationship in an *implicit* manner in an associative memory system (AMS).[4] These two investigations help the comparison of the results of an associative memory system with statistical and pattern recognition methods.

Other methods for the implicit storage of relationships are neural networks that have recently become prominent. Analysis by more traditional pattern recognition methods of the data set chosen here lays the foundation for an investigation by neural networks.[5]

THE DATA SET

The aim of the work was to study the reactivity of single bonds in aliphatic compounds. In particular, a reactivity function should be developed that is able to quantify the prospensity of a single bond for heterolysis, a process that generates a positively and a negatively charged species.

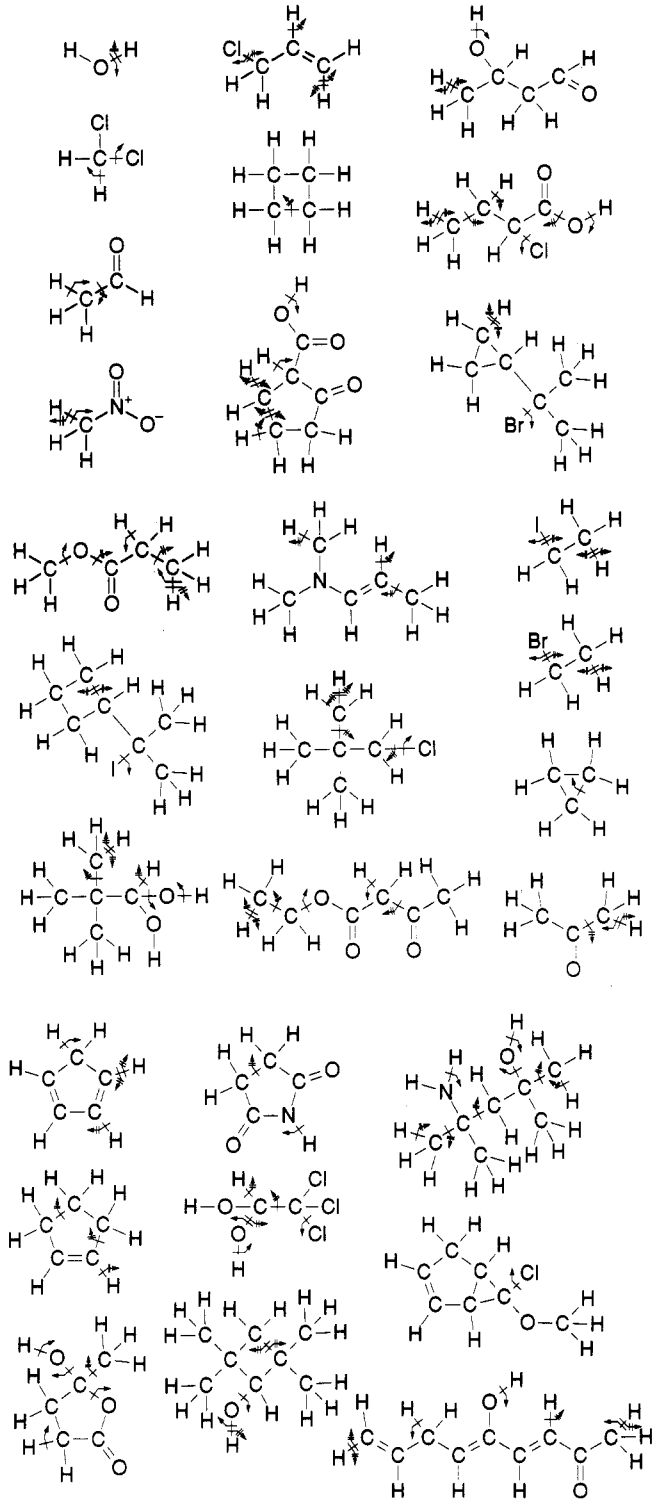
A data set of 29 aliphatic structures was selected containing bonds that comprise 724 polar bond breakings. From among these 116 were selected and classified either as breakable (reactive, 42) or nonbreakable (non-reactive, 74) (Scheme I).

Any organic chemist will agree with this classification of these bonds. Breaking the bond in cyclopropane was intentionally put into the category of reactive bonds, although it is not a very reactive bond (cf. the classification of the corresponding bond in cyclobutane), to see how this false classification will show up in the study with the various statistical methods.

The next problem is, then, how to describe the characteristics of all these bonds. The success of the endeavor will critically depend on the quality of the parameters chosen for the characterization of the bonds. The reactivity of bonds in organic structures depends on a variety of electronic, steric,

0095-2338/93/1633-0385$04.00/0

Scheme I. Data Set Molecules with Reactive (Bent Arrows) and Nonreactive Bonds (Crossed Arrows) As Indicated



and energy effects. All these factors have to be accounted for if a realistic picture of chemical reactivity is to be developed.

The following set of electronic and energy parameters were calculated by empirical methods: charge distribution,[6,7] inductive,[8] resonance, and polarizability effect,[1] as well as bond dissociation energies.[9] Each bond was characterized by the following parameters: difference in total charge, Δq_{tot}, difference in π-charge, Δq_π, σ-electronegativity difference, $\Delta\chi_\sigma$, the amount of charge, Q_σ, shifted in the PEOE method[6] across a bond as a measure of bond polarity, a measure of the resonance effect, R, for stabilizing the positive and negative charges generated by heterolysis, bond polarizability, α_b, and bond dissociation energy, BDE.

Each bond can be broken in a polar manner in two ways, depending on which atom obtains the positive charge and which the negative charge. The first four parameters mentioned above have different signs but the same absolute magnitude for the two choices of heterolysis of a bond. The resonance stabilization factor, in general, changes its value when the polar breaking of a bond is reversed, because the stabilization usually is different for a positive or a negative charge on an atom. The other two factors, bond polarizability and bond dissociation energy, are specific for a bond but do not depend on the direction of heterolysis.

INVESTIGATIONS BY PATTERN RECOGNITION METHODS

The seven parameters used to characterize a bond can be taken as coordinates of a space, the reactivity space. Reactivity spaces with three dimensions have been investigated by computer graphics methods, giving important insights into the factors controlling the course of chemical reactions.[10]

Reactivity spaces of such a high dimensionality as given here (seven), however, have to be studied by statistical and pattern recognition methods. An initial study of this data set has briefly been reported.[11] More details on the study of this data set will now be given to fully evaluate the performance of the various statistical methods and to allow a comparison with the performance of an associative memory system. The results are slightly different from those of the previous study[11] as an extra parameter, the difference in the π-charges on the atoms of a bond, Δq_π, has been included here and some small changes in the values of the resonance effect, R, and the bond dissociation energies BDE, have occurred through improvements in the calculation methods.

First, we will investigate by different pattern recognition methods whether the various electronic and energy parameters are able to reproduce the chosen classification of reactive and nonreactive bonds. Then, investigations are made on how many factors, consisting of linear combinations of the various parameters, are, in fact, necessary to reproduce the reactive/nonreactive classification of bonds. Based on the experience gained in these investigations, methods (like partial least squares analysis and logistic regression analysis) are used that produce equations to describe the reactivity of bonds in a general and explicit manner.

ANALYSIS OF THE PARAMETERS

The correlation among the seven parameters used in this reactivity data set is shown in Table I. Appreciable correlations are found between the difference in the total charge, Δq_{tot}, and the σ-bond polarity measure Q_σ (0.87), of the difference in the σ-electronegativity, $\Delta\chi_\sigma$, and the σ-polarity, Q_σ (–0.74), and of bond polarizability, α_b, with the bond dissociation energy, BDE (–0.75). These correlations have to be taken into account in the interpretation of the results. Table I also gives the correlation coefficients of the individual parameters with the reactivity classification of the bonds. The highest values are observed for those parameters that express polarity effects in the σ-skeleton: difference in σ-electronegativity, $\Delta\chi_\sigma$, difference in total charge, Δq_{tot}, and σ-polarity, Q_σ. This result is to be expected as these factors should have a large effect on the polar breaking of single bonds in aliphatic compounds.

The two variables with the highest Fisher quotients, Q_σ and $\Delta\chi_\sigma$, are plotted against each other in Figure 1. This picture shows that none of these two variables suffices to separate

Table I. Correlation Matrix and Fisher Quotients of the Parameters, Including the Correlation Coefficient of Each Individual Parameter with the Reactivity Classification

parameter	$\Delta\chi_\sigma$	Δq_π	R	α_b	Δq_{tot}	Q_σ	BDE
$\Delta\chi_\sigma$	1.00	0.25	0.00	0.04	−0.33	−0.74	−0.05
Δq_π	0.25	1.00	−0.21	0.10	−0.23	−0.37	0.01
R	0.00	−0.21	1.00	0.38	0.11	0.06	−0.29
α_b	0.04	0.10	0.38	1.00	−0.03	−0.06	−0.75
Δq_{tot}	−0.33	−0.23	0.11	−0.03	1.00	0.87	0.04
Q_σ	−0.74	−0.37	0.06	−0.06	0.87	1.00	0.06
BDE	−0.05	0.01	−0.29	−0.75	0.04	0.06	1.00
reactivity classification	−0.66	−0.30	0.28	−0.03	0.62	0.77	−0.02
Fisher quotient	1.76	0.18	0.17	0.00	1.21	2.92	0.00

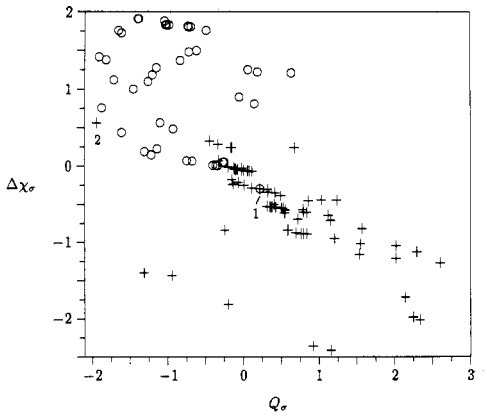


Figure 1. Plot of σ-polarity, Q_σ, against σ-electronegativity difference, $\Delta\chi_\sigma$. Both variables are autoscaled: O, reactive bonds; +, nonreactive bonds. Point no. 2 is discussed in conjunction with Figures 2 and 3. Point no. 1 refers to the heterolysis of cyclopropane which was intentionally given the wrong classification of being reactive.

Table II. Loadings of the Factors with the Parameters in a Principal Component Analysis before Varimax Rotation

	factors of PCA		
	1	2	3
$\Delta\chi_\sigma$	0.47	−0.01	−0.24
Δq_π	0.32	−0.07	0.72
R	−0.06	0.45	−0.53
α_b	0.09	0.64	0.24
Δq_{tot}	−0.52	0.06	0.15
Q_σ	−0.62	0.04	0.18
BDE	−0.09	−0.62	−0.19
% variance	35.9	28.3	13.7
Σ (% variance)	35.9	64.2	77.9
Fisher quotient	2.93	0.05	0.00

reactive from nonreactive bonds. However, the difference in σ-electronegativity, $\Delta\chi_\sigma$, comes close to that ideal.

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is a standard method for obtaining the essential effects that control the variation in the parameters in describing a data set. This unsupervised learning technique therefore leads to a reduction in the number of dimensions needed to characterize a data set. The new parameters (called factors) are calculated as linear combinations of the original ones.

The results of a principal component analysis are given in Table II. The first factor comprises the polar effects in the σ-skeleton ($\Delta\chi_\sigma$, Δq_{tot}, and Q_σ); the second factor is dominated by resonance stabilization, the bond polarizability, and the bond dissociation energy. π-charge difference and resonance stabilization control the third factor. Together they comprise 77.9% of the variance in the data set.

The principal component analysis has identified three major factors within the data set, that may influence the reactivity

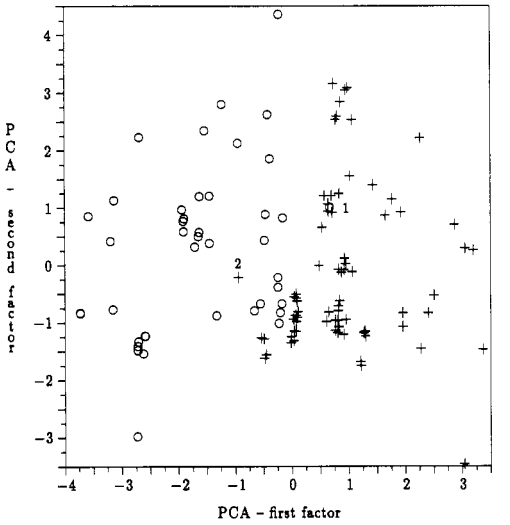
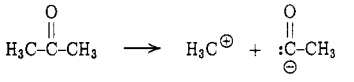


Figure 2. Plot of the second component against the first component of the PCA: O, reactive bonds; +, nonreactive bonds.

of single bonds in aliphatic compounds: (1) σ-bond polarity and inductive effect, (2) the strength of a bond (as expressed by low values of BDE or high polarizability), and (3) π-effects. Furthermore, it shows how these three factors are expressed in contributions of the seven initial electronic and energy parameters.

The first two factors of the PCA are plotted against each other in Figure 2. Although PCA is an unsupervised learning method, i.e., the information whether a bond is reactive or nonreactive is not used in PCA, the first component of the PCA can separate the two types of bonds to a large extent. This indicates that the polar effect in the σ-skeleton has a high contribution to the reactivity of a bond. Point no. 2 refers to the heterolysis of the C–C bond in acetone.



It was classified as nonreactive but is surrounded in this plot (Figure 2) by reactive bonds. The main reason for this bond to appear in the reactive region is its highly negative value for the σ-electronegativity difference (cf. Figure 1).

The plot of Figure 3 of the first and the third factors of the PCA again shows the distinguished nature of the C–C bond in acetone (no. 2). However, this plot hints that this bond lies in a space that is connected to the space of the other nonreactive bonds in the multidimensional reactivity space, a fact that does not show up in the two-dimensional projection of Figure 2.

The intentionally falsely classified bond of cyclopropane (as reactive) correctly shows up in both plots (point no. 1 in Figures 2 and 3) in the region of nonreactive bonds.

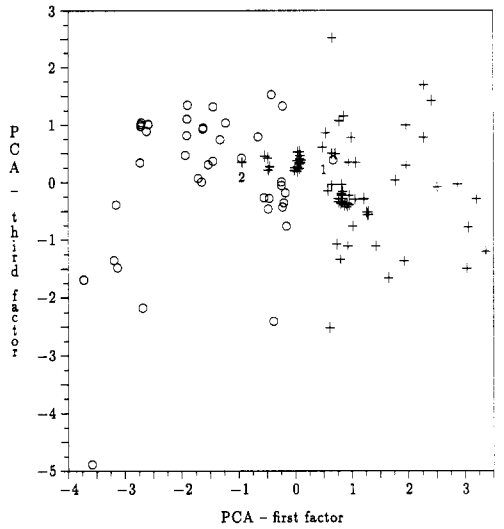


Figure 3. Plot of the third component against the first component of the PCA: O, reactive bonds; +, nonreactive bonds.

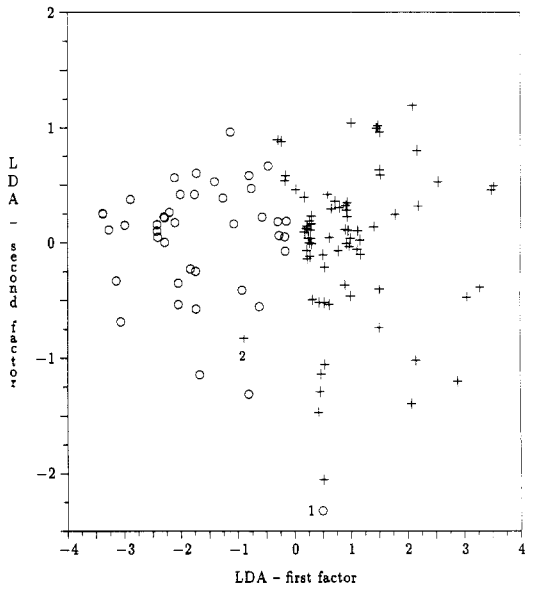


Figure 4. Results of a discriminant analysis: plot of the first and second component.

LINEAR DISCRIMINANT ANALYSIS

The PCA leads to a fairly good separation of reactive and nonreactive bonds (cf. Figure 3). The question now arises, which factor separates the two groups in an optimum manner? Linear discriminant analysis (LDA) provides an answer to that question.

In contrast to PCA, a discriminant analysis is a method of supervised learning and therefore uses the information of whether a bond is reactive or not in the statistical analysis.

The first factor again comprises the polar effects within the σ-skeleton, while the second is dominated by bond dissociation energy.

The second factor of the discriminant analysis is plotted against the first one in Figure 4. The variables have again been autoscaled. In comparison with Figure 2, the separation of the two classes of bonds has now further improved. Clearly, this was caused by the use of the reactivity information in the discriminant analysis.

Still, the C–C bond of acetone (point 2) penetrates far into the region of reactive bonds. The LDA succeeds in separating the point for the polar breaking of the C–C bond of cyclopropane from the cluster of nonreactive bonds.

However, this point is still closest to nonreactive bonds indicating that it, too, should be considered as representing a nonreactive bond. Thus, the LDA still can decipher the false classification, indicating that the results have strong chemical significance and cannot be destabilized by a wrong classification.

CLUSTER ANALYSIS

A more general method for finding groupings of similar reactions is cluster analysis. In contrast to PCA and LDA, cluster analysis is also able to show the relationships in the reactions in more than two dimensions. Therefore the analysis of the data set is not restricted to one selected plane. Cluster analysis is again a unsupervised learning method. The classification of a bond into reactive or nonreactive is not used in performing a cluster analysis. However, this classification was scrutinized in the examination of the results of a cluster analysis.

If the points of reactive or nonreactive bonds, respectively, group together, then this is an indication that the parameters used as variables to define the space for the cluster analysis do describe reactivity well.

A series of hierarchical cluster analyses were performed with different types of transformations (none, autoscaled, Mahalanobis), different metrics (Minkowski, Tanimoto, direction cosine, and absolute direction cosine), and various linkage methods (single, complete, average, weighted, median, centroid, and Ward).

The best results in terms of grouping reactive bonds together and nonreactive ones were obtained when a Mahalanobis distance measure and Ward's linkage method were used.

Figure 5 shows the dendrogram for the cluster analysis with a Mahalanobis transformation, Euclidean distance, and Ward linkage using all seven parameters as variables. It can be seen that the reactive bonds group together, indicating that they occur in clusters close together in the seven-dimensional space.

The breaking of the C–C bond of acetone falls into a cluster of reactive bonds. The peculiar nature of this bond has already been found in the principal component and the discriminant analysis. Furthermore, the point for the heterolysis of cyclopropane approaches a cluster of reactive bonds only in a late stage of the dendrogram, indicating that it should be considered nonreactive.

The picture becomes even clearer when, instead of all seven variables, only three, the σ-polarity, Q_σ, the resonance stabilization, R, and the difference in total charge, Δq_{tot}, are used. Figure 6 shows the dendrogram of the cluster analysis with these three parameters and the same Mahalanobis distance and Ward linkage method as in the study of Figure 5.

Now, the reactive bonds group together even more, the point for the C–C bond of acetone falls into the cluster of reactive bonds, and the C–C bond of cyclopropane clearly is in the region of nonreactive bonds.

Using the dendrogram of Figure 6 to partition the data set into 16 clusters (as indicated by the dashed vertical line in Figure 6) results in a correct grouping of all bonds except of the two bonds mentioned above.

K-NEAREST NEIGHBOR ANALYSIS

All methods described up to now have in common that their main use is in the analysis of a data set and not to make predictions. In contrast to this the k-nearest neighbor analysis

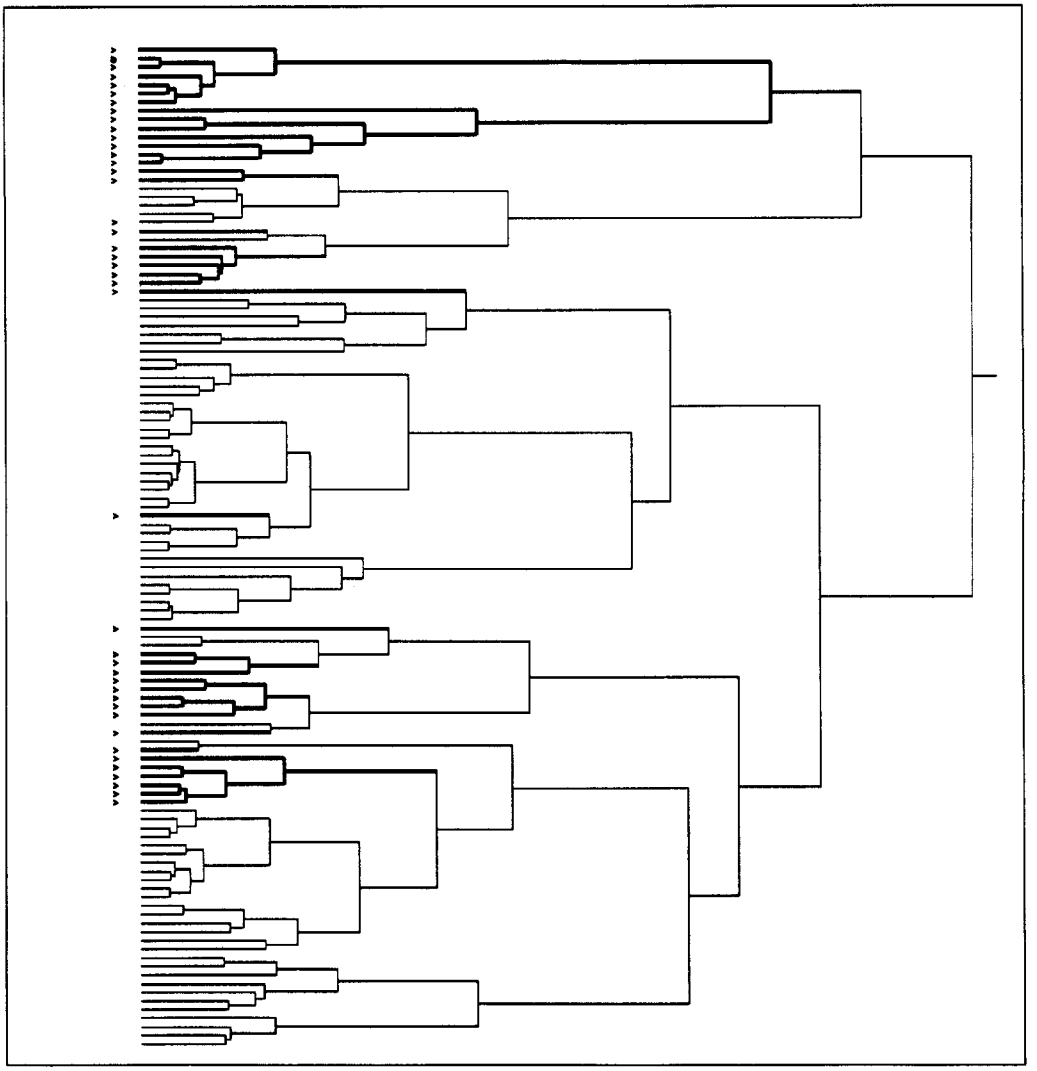


Figure 5. Dendrogram of a hierachical cluster analysis with all seven variables. A > sign marks a reactive bond.

(KNN) is an unsupervised learning method that is not only capable of finding relationships in the data set but also of predicting reactivity values for reactions not contained in the data set.

The information of whether a bond is reactive or not is only used a posteriori in analyzing the results of a KNN analysis. The classification of a bond is assumed to be the same as that of the majority of its k (k = 1, 2, 3, etc.) nearest neighbors. This deduced classification is then compared with the one initially given. If the two are different, this bond is counted as a misclassification.

Table III shows the number of misclassifications for various combinations of physicochemical parameters and with k = 1–7. (The reasons for the selection of both two-parameter and the one three-parameter combinations are given in the section on logistic regression analysis.) In these KNN analyses the distance was calculated by a Euclidean metric. Use of a Mahalanobis distance does not lead to a further improvement in the results.

The best results are obtained when the two parameters, σ-electronegativity difference and difference in total charge are used and the number of neighbors is set to 5 (or 6). From the two misclassifications one is the bond breaking in cyclopropane which has intentionally been set into the wrong category. Thus, this "misclassification" is in fact a correct result, as this wrong classification is recognized in the KNN analysis. The other misclassification is the bond breaking in acetone leading to methyl cation a reaction that has already been discussed in conjunction with the PCA, LDA, and cluster analysis. This bond, being classified as nonreactive is largely surrounded by reactive bonds. However, they are at larger distances than are usually met in this analysis: The Euclidean distance to the next neighbor is approximately twice as large as the average distance in the KNN analysis of this data set. This again underscores the unique nature of this bond breaking, indicating that the reactivity space should be filled with additional bond breakings having parameters similar to the one in acetone.

PARTIAL LEAST SQUARES

A main objective of many statistical analyses of a data set is to express the relationship between the parameters describing an object (chemical bond in our case) and the property under investigation (chemical reactivity) in a functional form. In this and the next section we will see two such attempts.

The simplest functional form is a linear relationship. Clearly, a 0/1 classification as given by the nonbreaking/breaking of a chemical bond is not very well suited for a linear approach. However, such an investigation might lead to interesting insights. The partial least squares (PLS) method[12] is an approach for the derivation of a linear functional relationship between the property of an object and the independent variables describing this object.

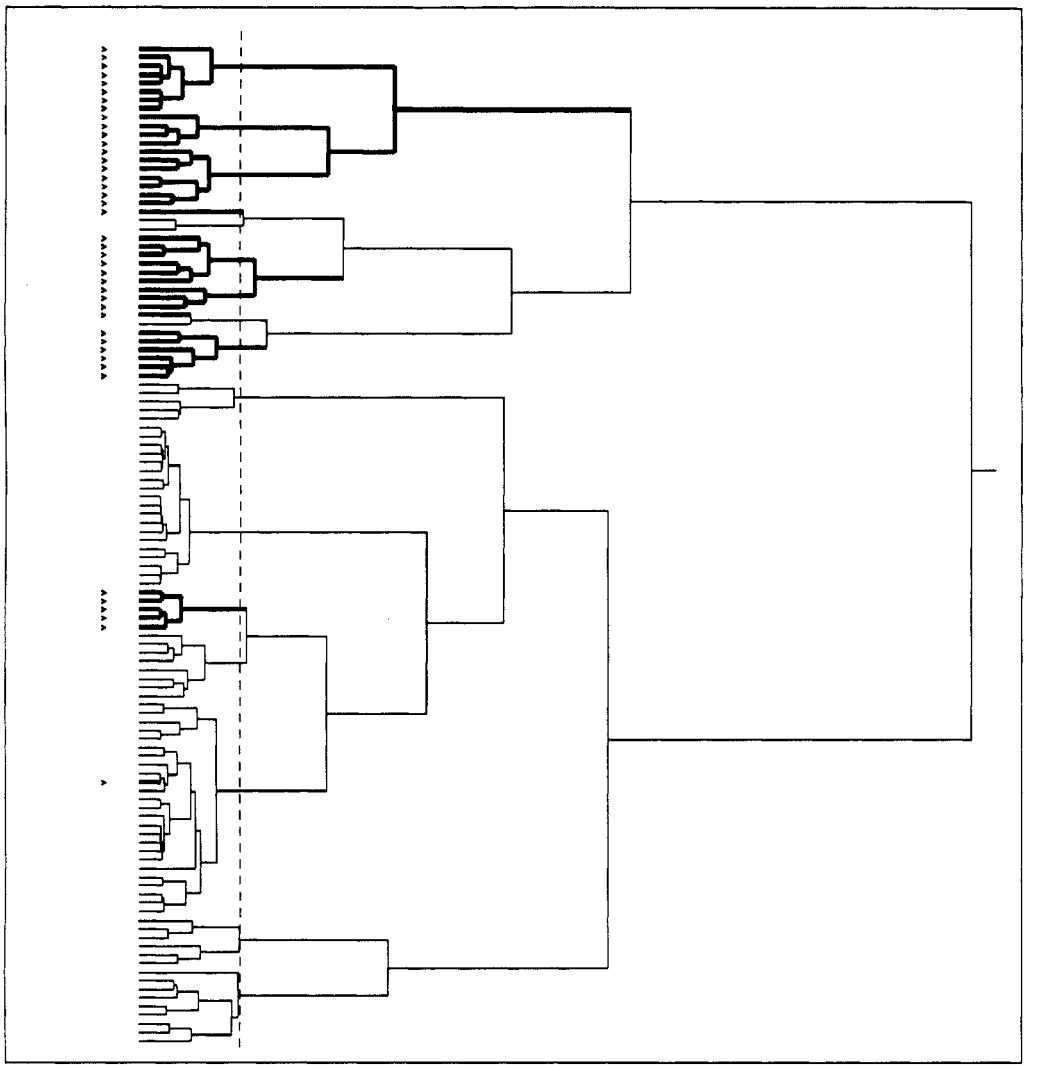


Figure 6. Dendrogram of a cluster analysis with Q_σ, R, and Δq_{tot} as variables. The dashed vertical line marks the partition into 16 groups. Reactive bonds are indicated by > signs.

Table III. Number of Misclassifications in a KNN Analysis with Different Combinations of Parameters

combination	parameters used	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$
1	all seven	5	5	4	4	6	6	6
2	$\Delta\chi_\sigma$, Δ_{qtot}, BDE	4	4	7	5	4	4	5
3	R, Q_σ	6	6	4	4	5	5	5
4	$\Delta\chi_\sigma$, Δq_{tot}	3	3	3	3	2	2	3

Table IV. Loadings of the Two Significant Factors Computed by the PLS Method

	factors of PLS			factors of PLS	
	1	2		1	2
$\Delta\chi_\sigma$	−0.484	−0.188	Δq_{tot}	0.517	−0.172
Δq_π	−0.308	0.436	Q_σ	0.618	−0.097
R	0.145	0.688	BDE	0.008	−0.334
α_b	−0.009	0.387			

The investigation with the PLS method leads to two significant factors which are shown in Table IV.

The first factor represents the polarity of the bond as expressed by the σ-electronegativity, $\Delta\chi_\sigma$, the difference in total charge, Δq_{tot}, and the bond polarity, Q_σ, as these parameters have the highest loadings. The resonance stabilization, R, dominates the second factor with additional contributions from the difference in π-charges, bond polarizability, and bond dissociation energy.

Taking the values of these two factors to calculate the ease of breaking a bond gives a linear equation with a correlation coefficient of 0.822 and a standard deviation of 0.275. Considering the crude nature of the dependent variable—it can only take 0 or 1—this is a fairly good correlation. If all calculated reactivities less than 0.5 are considered as nonreactive, then the two-dimensional PLS model gives rise to nine misclassifications as compared to the classifications of Scheme I (including the heterolytic opening of the cyclopropane ring). This result looks worse than it is because for most of the misclassifications the PLS model computes reactivity values in the range between 0.4 and 0.6. If the strict threshold of 0.5 is modified in such a way that all computed reactivities in the interval from 0.4 to 0.6 are regarded as not classified, then the PLS model leads to one misclassification only—the polar breaking of the C–C bond in cyclopropane. This is a satisfying results, as this bond breaking was intentionally put in the wrong category. The number of reactions for which no classification can be obtained then amounts to ten. This means that besides the eight misclassifications in the case of the strict boundary of 0.5, there are only two additional reactions whose predictions fall into the interval from 0.4 to 0.6.

Table V. Best Combinations of Two, Three, Four, Five, and Six Parameters Found by Logistic Regression Analysis

combination	parameters used	wrong classifications		
		total	reactive	nonreactive
1	$\Delta\chi_\sigma$, Δq_{tot}	10	6	4
2	Q_σ, R	8	6	2
3	$\Delta\chi_\sigma$, Δq_{tot}, BDE[a]	8	6	2
4	$\Delta\chi_\sigma$, Q_σ, BDE[a]	8	6	2
5	Q_σ, R, $\Delta\chi_\pi$[a]	6	4	2
6	Δq_{tot}, R, $\Delta\chi_\pi$[a]	6	4	2
7	$\Delta\chi_\sigma$, Δq_{tot}, R, BDE[a]	6	4	2
8	Q_σ, R, $\Delta\chi_\pi$,[a] BDE[a]	6	4	2
9	$\Delta\chi_\sigma$, Δq_{tot}, R, α_b[b]	2	1	1
10	Q_σ, R, $\Delta\chi_\pi$,[a] α_b[b]	2	1	1
11	R, $\Delta\chi_\pi$,[a] Δq_{tot},[a] α_b,[b] Q_σ[a,b]	2	1	1
12	$\Delta\chi_\sigma$, Δq_{tot}, R, BDE, α_b,[b] Q_σ[b]	0	0	0

[a] Variables that were found to be insignificant by the (χ^2) test. [b] Variables that have a coefficient with a sign contrary to chemical expectations.

LOGISTIC REGRESSION ANALYSIS

Logistic regression analysis (LoRA) has been found to be a particularly promising tool for investigating problems as met here. In LoRA, a data set with a binary classification of objects is used to derive a function that is able to reproduce this classification as good as possible. In our context, the classification of a bond as reactive or nonreactive is used to derive a function that quantifies chemical reactivity based on the electronic and energy parameters.

The original classification is considered as an input probability P_o (reactive = 1.0, nonreactive = 0.0). This classification is modeled by a logistic function (eq 1) as a calculated probability, P, where the exponent, f, is expanded as a linear function in the parameters, C_i, used (eq 2). The

$$P = \frac{1}{1 + e^{-f}} \tag{1}$$

$$f = c_0 + c_1x_1 + c_2x_2 + \cdots \tag{2}$$

coefficients, c_i, in eq 2 are determined so as to minimize the error between the initial classification, P_o, and the calculated probability, P.

Table V gives the combinations of two, three, four, five, and six parameters that were found to be the best ones by LoRA. Only both two-parameter combinations (nos. 1 and 2) had neither variables that were established to be asymptotically insignificant by the (χ^2) criterion nor coefficients in eq 2 with a sign contrary to chemical expectations (e.g., an increase in bond polarizability is expected to increase chemical reactivity and therefore should have a coefficient in eq 2 with a positive sign). However, both these two-parameter equations led to the rather large number of eight or ten wrong classifications, respectively.

The results of the logistic regression analyses provided the reason for selecting these two combinations of parameters (Table V, lines 1 and 2), as well as the first three-parameter combination (line 3) in studies by k-nearest neighbor analysis as well as in the investigation with an associated memory system.[4]

At the other end of the scale is the six-parameter combination (line 12) which "correctly" classifies all bonds. However, this means that the intentionally misclassified cyclopropane bond breaking is not perceived. In addition, the peculiar nature of the C–C bond breaking in acetone (see PCA, LDA, and KNN) is not perceived.

However, the five-parameter combination (line 11) and the last two four-parameter combinations (lines 9 and 10) do recognize those two bonds. They are the ones accounting for the two "misclassifications", correctly putting the cyclopropane bond into the category of nonreactive bonds and classifying the C–C bond of acetone as reactive as the other pattern recognition methods (PCA, LDA, and KNN) do. Of those three combinations of parameters, the one (line 9) using the four variables σ-electronegativity difference, difference in total charge, resonance effect, and bond polarizability seems to be the best. The only drawback in this correlation (eq 2) is the wrong sign for the coefficient of bond polarizability. Apparently the numbers for the polarizability effect reproduce an effect that is statistically significant and more important than the polarizability effect. With this caveat in mind, we give eq 3 using this combination no. 9 of four parameters.

$$f = -2.72 - 5.26\Delta\chi_\sigma + 19.1\Delta q_{tot} + 0.354R - 0.722\alpha_b \tag{3}$$

Putting the results of this equation into eq 1 gives probabilities for bond breaking that correctly classify all but one of the 116 bond breakings as reactive or nonreactive (taking the perception of the misclassification of the cyclopropane bond as a correct answer). The only exception is the C–C bond of acetone that is distinguished by values of the parameters that put this bond breaking into a region that has no other bond breakings very similar to it, but reactive bonds in the farther distance.

PREDICTION OF CHEMICAL REACTIVITY

In natural science the use of a theory or model depends on its predictive power. To check the usefulness of the models computed in the previous section, it is neccessary to predict the reactivities of some molecules not contained in the data set. The merit of the different models emerges when comparing the predictions with chemical experience.

The following test includes five different models stemming from the investigations with PLS, principal component analysis, linear discriminant analysis, KNN, and logistic regression analysis. The predictions are made with those models of the various methods found to give the best results. This means that KNN predictions are based on the parameter combination 4—$\Delta\chi_\sigma$ and Δq_{tot}—and the logistic function uses eq 3. The results of PCA predictions are founded on the first and third factors only, while in the case of LDA the first and second factors were selected.

It is obvious how to calculate predictions using PLS, KNN, or LoRA, but the procedure in the case of PCA and LDA needs some explanation. One standard method for projection techniques like PCA or LDA consists of imbedding by drawing the data point that is to be predicted into the plot and classify it by inspection. This fairly efficient method has one important drawback: the human pattern recognizer will get tired if a large number of predictions is to be made.

There is a simple procedure to calculate a prediction from the distribution of data points in any n-dimensional space: the predicted reactivity value is taken as the weighted arithmetic mean of the reactivities of the bonds in the data set. The weight of each reaction of the data set with known reactivities depends on its distance to the reaction to be predicted—the greater the distance the lower its weight. The reaction to be predicted is characterized by the coordinates x, whereas the reaction i of the original data set has the parameters x_i. Equation 4 gives the general formula (with the predicted reactivity $P(x)$, the weight function, w, the

$$P(x) = \frac{\sum_{i=1}^{n} w(d(x,x_i))\,P_i}{\sum_{i=1}^{n} w(d(x,x_i))} \quad (4)$$

distance measure, d, and the reactivity of the data point in the data set, P_i). Appropriate weight functions may take the following forms:

$$w(r) = e^{-ar^n}; \quad (5)$$

$$w(r) = (1 + ar^n)^{-1}; \quad (6)$$

$$w(r) = r^{-n}; \quad (7)$$

with $a, n > 0$; $r \leftarrow d(x, x_i)$.

A further refinement of this computation procedure consists of taking only those data points into account that lie in a sphere (circle) of a given radius r_s around the considered reaction. In this local variant it is neccessary to specify the minimum number of data points k_m that the sphere must contain to avoid predictions that are based on no known reaction at all.

The same molecules as in the investigation with the associative memory system[4] form the base for testing the predictive power of the different models. The detailed discussion of every prediction of all models would require too much space; therefore the results are presented in a relatively compact form. First, all polar bond breakings are shown that are predicted by all five models as reactive. Then follows a discussion of the reactivity values of 23 selected bonds. All other bonds in these molecules are predicted as nonreactive.

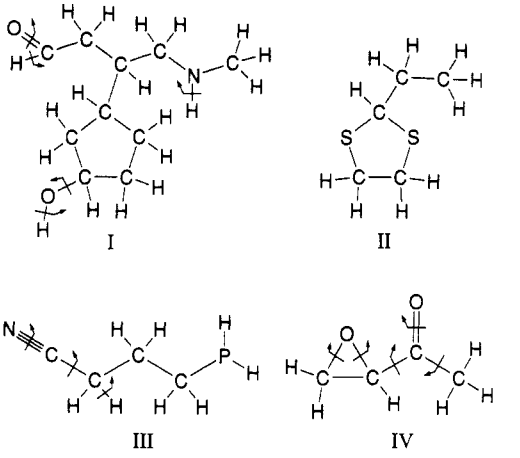
Scheme II depicts the four molecules and the reactions that are predicted by all five models as reactive. Independent of the chosen model, 13 different polar bond breakings appear as reactive. All carbonyl groups exhibit a preference for the nucleophilic attack at the carbonyl carbon atom, indicated by a breaking of the C=O double bond to give the mesomeric structure $C^+—O^-$. Moreover all carbon–oxygen single bonds can be broken in the direction that shifts the bonding electron pair to the oxygen. These predicted reactivities correspond with chemical experience. That is also true for the predicted loss of a cyanide ion or the abstraction of a proton from the amine nitrogen or an alcohol oxygen. The deprotonation of the methylene group adjacent to one nitrile function represents a reaction that is easy to achieve. The heterolysis of one C–N bond in the nitrile function corresponds to the polar breaking in a carbonyl group. The hydrolysis of the nitrile under basic conditions shows that the nucleophilic attack at the nitrile carbon atom happens in reality.

The remaining three bonds which all five models predict as reactive have one feature in common: the bonding electron pair is shifted to the carbon atom of a carbonyl group. The deprotonation of the aldehyde function is not an important process in chemical experiments. This is due to competitive reactions that are faster (e.g., nucleophilic attack at the carbonyl group, Cannizarro reaction, etc.). The similarity of the last two reactions to the heterolysis of the C–C bond in acetone, a reaction in the data set which plays a special role in all investigations (including AMS[4]), is even more evident. In (2-oxoethyl)oxiran both C–C bonds containing the carbonyl carbon atom are predicted to be reactive in such a way that the carbonyl carbon atom receives the negative charge.

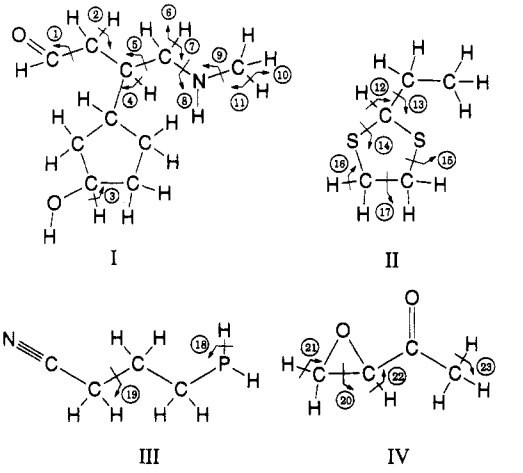
Scheme III shows 23 selected reactions of the five molecules that will be discussed subsequently. These 23 reactions cover all reactions that are predicted as reactive by at least one model, but not by all five.

Table VI lists the predictions on the reactions shown in Scheme III. The obtained values stem from the investigations

Scheme II. Test Set of Molecules with Those Reactions That All Five Models Predict as Reactive



Scheme III. 23 Reactions of the Test Set Chosen for Detailed Discussion



with PCA, LDA, KNN, PLS, LoRA, and the AMS.[4] This column is given here to permit a comparison of the results with those obtained in an accompanying study.[4] Table VI also contains the estimate of the reactivity by a chemist.

A look at the columns of Table VI shows that all models except PLS predict in most cases either very high or very low reactivities while PLS computes only values in the range from 0.258 to 0.606. Another interesting feature that distinguishes the models emerges if the number of bonds in Table VI that are predicted as reactive are counted. The number of bonds predicted reactive leads to the following order of the models: PLS (6), PCA (6), KNN (9), LDA (11), and LoRA (17). These two features point to disadvantages of the models using explicit functions: the predictions of PLS are too close together (cf. Scheme III: reactions 6 and 2), and the world of the LoRA model seems to be too reactive.

All models except PCA predict the formation of the formyl anion (1) as a polar bond breaking with a high reactivity. This reaction resembles the cleavage of the C–C bond in acetone (cf. the discussion above).

A chemist will classify reaction 2 of Scheme II as reactive. In contrast to this, the predictions of the five models show wide variations. Only KNN and LoRA estimate the C–H bond as reactive. The result of PLS lies somewhat near to the class boundary while the prediction of the PCA depends on the parameters chosen (the restriction to a sphere with r_s = 0.05 and k_m = 2 leads to a reactivity value of 0.503). This dependence indicates that the reaction lying closest to reaction 2 is reactive while the second closest reaction is nonreactive.

Table VI. Reactivities of 23 Selected Bonds Calculated by According to Different Models

reaction	PCA	LDA	KNN	PLS	LoRA	AMS[c]	chem[d]
1	0.212	1.0	1	0.606	0.976	0 (80)	0
2	0.376[a]	0.0	1	0.458	0.911	98 (93)	1
3	0.145	1.0	1	0.390	0.687	100 (100)	1
4	0.065	0.0	1	0.258	0.090	86 (100)	0
5	0.0	0.075	0	0.499	0.663	3 (100)	0
6	0.0	0.0	0	0.455	0.673	0 (100)	0
7	0.0	0.153	1	0.309	0.266	90 (93)	0
8	0.762	1.0	0	0.526	0.970		0
9	1.0	0.715	0	0.491	0.973		0
10	0.0	0.0	0	0.413	0.780	0 (100)	0
11	0.0	0.113	0	0.310	0.573	1 (100)	0
12	0.658[b]	0.0	1	0.335	0.187	100 (100)	1
13	0.0	0.0	0	0.309	0.034	100 (80)	0
14	0.096	1.0	0	0.605	0.941	0 (40)	0
15	1.0[b]	0.830	0	0.524	0.856	0 (53)	0
16	0.0	0.092	1	0.286	0.191	97 (100)	0
17	0.250[a]	0.0	0	0.267	0.026	6 (100)	0
18	0.641	0.677	0	0.360	0.625	0 (46)	0
19	0.129	0.523	0	0.471	0.764	100 (60)	0
20	0.0	0.356	0	0.571	0.956	100 (46)	1
21	0.0	0.556	0	0.379	0.649	73 (100)	1
22	0.303	1.0	1	0.605	0.990	100 (100)	1
23	1.0	1.0	1	0.468	0.977	97 (100)	1

[a] Cases in which a steep weight function $w(r)$ leads to a change in the predicted category. [b] Cases in which a less steep weighting function $w(r)$ leads to a change in the predicted category of reactivity. [c] The first value gives the reactivity (in percent); The value in parentheses is the reliability of the reactivity value. [d] Classification by a chemist.

Only LDA estimates a greater reactivity for reaction 3 as for reaction 2 whereas chemical experience agrees with the opposite ranking that is predicted by PLS, PCA, and LoRA. Disregarding competitive reactions, reaction 3 should be considered as reactive due to the inductive effect of the alcohol group—a classification that is given by LDA, KNN, and LoRA.

KNN is the only model besides the AMS that predicts the deprotonation of the tertiary carbon (4) as reactive. Both models—KNN and the AMS—do not make use of the parameter resonance stabilization and are therefore tempted by the coincidence of the inductive parameters of reaction 4 with those of, e.g., the deprotonation of the methyl group of ethanal (for further discussion see ref 4). This is also true for reaction 7 which the AMS and KNN put in the wrong class.

The three reactions 5, 6, and 10 have two things in common. First only LoRA estimates them as reactive with ratings between 0.663 and 0.780. Second, the nitrogen stabilizes the resulting positive charge using its lone electron pair. The influence of the mesomeric stabilization is overestimated as these reactions never occur in experiments.

The predictions for the reactions cleaving the carbon nitrogen single bond (8 and 9) show only minor differences between the different methods. In the case of PLS this leads to different classifications. While KNN predicts the two bonds as nonbreakable, the other models put them into the reactive class. In this case chemical experience supports the results of KNN.

The deprotonation of the methyl group adjacent to the nitrogen, reaction 11, resembles reaction 7—the deprotonation of the methylene group adjacent to the same nitrogen—both in chemical experience as well in the predictions of the PLS, PCA, and LDA models. The reason why KNN leads to the correct result in the case of reaction 11 in contrast to reaction 7 is that there are small but not negligible differences in $\Delta\chi_\sigma$. This increased polarity of bond 11 leads in the LoRA model to a classification as reactive.

The reactions of molecule II (12–17) represent bonds that are not contained in the data set (cf. Scheme I). The aim of this molecule is to test the scope of the different models.

An example of a well-known reaction is bond 12, the anion being stabilized by the two sulfur atoms. Correct predictions are made by KNN and PCA. All five models predict bond 13 as nonreactive. This result corresponds to chemical experience.

The difference between reactions 14 and 15 stems from the possible resonance stabilization of the positive charge generated in reaction 14 by the adjacent sulfur atom. The predictions of PLS, LDA, and LoRA express this preference. The disadvantage of these models is that they estimate the sulfur as a good nucleofuge. From the viewpoint of chemical experience the predictions of KNN should be preferred.

In contrast to reaction 12 the negative charge in the product of reaction 16 can be stabilized by only one sulfur atom. Therefore this reaction must be considered as nonreactive. All models except KNN and the AMS predict the correct behavior.

The last reaction of molecule II consists of the heterolytic opening of the ring via breaking of the C–C bond. All models classify this reaction as nonreactive due to the apolar nature of the bond.

Phosphines are weaker acids and also weaker bases than amines. The predictions for reaction 18 show that all models agree with this lower acidity of the P–H bond, keeping in mind that all five models predict the deprotonation of the amine group in molecule I (Scheme II) as feasible. PLS and KNN put reaction 18 into the nonreactive class, while the three other models estimate a reactivity slightly above the threshold of 0.5.

Reaction 19 has some similarity to the deprotonation of the methylene group adjacent to the nitrile function. The nitrile group stabilizes the resulting negative charge on carbon both by inductive and mesomeric effects. This explains the estimations of LoRA and LDA (and the rather high value of the PLS model). Nearly all chemists will classify this reaction as nonreactive, taking into account that the solvation of the formed cation is worse than in the case of the proton.

Reaction 20 shows some interesting features. The inductive effect of the oxygen atom as well as that of the carbonyl group

stabilizes the formed negative charge while the positive charge is located adjacent to an electronegative atom. The resonance effect plays a more important part: Both charges can be stabilized, the positive by the oxygen and the negative due to conjugation with the carbonyl group. The prediction of the reactivity of this bond constitutes no easy task even for an experienced chemist because of the reactivity of the C–O bonds in the oxiran ring. Almost every nucleophilic attack at the methylene group will lead to a cleavage of the C–O bond and not of the C–C bond. The suppression of reaction 20 by fast competing reactions (the heterolysis of the C–O bonds is predicted by all five models as reactive) pretends a stability of the C–C bond in the oxiran ring. Summarizing these facts, bond 20 has to be considered as reactive. And, indeed, it can be observed in 1.3-dipolar cycloaddition reactions. The predictions of PLS and LoRA come to that conclusion. The disregard of the resonance effect in the KNN and the LDA model explain their (incorrect) predictions.

The last three reactions treat the acidities of the different hydrogen atoms contained in molecule IV of Scheme III. Ranking these three bonds according to their acidities should lead to the following result: $22 \geq 23 \geq 21$. The only model that fails to predict this sequence is PCA. The predictions of the LDA and the LoRA model correspond to the chemical experience that classifies all three bonds as reactive.

CONCLUSION

The discussion of the predictions of the different models demonstrate the value of the pattern recognition methods in predictions on reactivity in organic chemistry. The benefit from the use of the pattern recognition methods is illustrated by their error rate of approximately 10%. The calculation of the error rate is based not only on the reactions discussed above but on all possible reactions of the prediction set. The prediction set (cf. Scheme III) contains 148 reactions. After removal of equivalent reactions 105 different reactions remain. The number of mismatches between chemical estimation and the predictions of the five models ranges from 8 (KNN) to 14 (LoRA). Mismatches appear only in some of the 36 discussed reactions, not in the remaining 69 nonreactive reactions.

An error rate of 10% is fairly good considering the high demands the test cases make on the models.

The second mentioned advantage of the predictions with respect to the investigations emerges in comparison of the results of the predictions with the results given for each method. The number of misclassifications in the investigations leading to the five models is much smaller than the number of mismatches obtained with the predictions on the five molecules of Scheme II. Take as an example KNN. The error rate in the investigation of the data set has a value of about 1% although the reactivity classification is obtained as a prediction (a reaction is not a neighbor of itself). The analyses lead to a model that consists only of the differences in σ-electronegativity and in total charge. The discussion of Table VI indicates that this model fails on reactions which are governed by resonance stabilization. This is due to a not optimally designed data set. The predictions as well as the application of set of different pattern recognition methods help to detect such deficiencies.

Despite the crude nature of the given reactivity values and the drawbacks—containing one (intentional) misclassification and not being optimally balanced with respect to the importance of the different effects—all developed models yield good results both in analysis as in prediction. This robustness of the methods is especially important for methods like LoRA which compute an explicit function to model the investigated feature.

Another insight that is supported by the results obtained reads: apply as many different analysis methods as possible because this reduces the probability of overlooking the importance of an effect.

ACKNOWLEDGMENT

We thank Dr. P. Löw for putting this data set together and for performing initial studies on these data. We appreciate the contribution of Dr. H. Saller for bringing logistic regression analysis to our attention. Interesting discussions with Prof. Dr. H. Kubinyi, BASF AG, are appreciated. Financial support of this work by the Bundesminister für Forschung und Technologie is gratefully acknowledged.

REFERENCES AND NOTES

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

(2) Gasteiger, J.; Hutchings, M. G. Quantitative Models of Gas-Phase Proton Transfer Reactions Involving Alcohols, Ethers, and Their Thio Analogs. Correlation Analyses Based on Residual Electronegativity and Effective Polarizability. *J. Am. Chem. Soc.* **1984**, *106*, 6489–6495.

(3) Hutchings, M. G.; Gasteiger, J. A Quantitative Description of Fundamental Polar Reaction Types. Proton and Hydride Transfer Reactions Connecting Alcohols and Carbonyl Compounds in the Gas Phase. *J. Chem. Soc., Trans. Perkin 2*, **1986**, 447–454.

(4) Gasteiger, J.; Schulz, K.-P. Elucidation of Chemical Reactivity Using an Associative Memory System. *J. Chem. Inf. Comput. Sci.*, following paper in this issue.

(5) Simon, V.; Gasteiger, J.; Zupan J. A Combined Application of Two Different Neural Network Types for the Prediction of Chemical Reactivity. *J. Am. Chem. Soc.*, submitted for publication.

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

(7) Gasteiger, J.; Saller, H. Calculation of the Charge Distribution in Conjugated Systems by a Quantification of the Resonance Concept. *Angew Chem.* **1985**, *97*, 699–701; *Angew. Chem., Int. Ed. Engl.* **1985**, *24*, 687–689.

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

(9) Gasteiger, J. Automatic Estimation of Heats of Atomization and Heats of Reaction. *Tetrahedron* **1979**, *35*, 1419–1426.

(10) Gasteiger, J.; Röse, P.; Saller, H. Multidimensional Explorations into Chemical Reactivity: The Reactivity Space. *J. Mol. Graphics* **1988**, *6*, 87–92.

(11) Gasteiger, J.; Saller, H.; Löw, P. Elucidating Chemical Reactivity by Pattern Recognition Methods. *Anal. Chim. Acta* **1986**, *191*, 111–123.

(12) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM J. Stat. Comput.* **1984**, *5*, 735–743.