Models for the Representation of Knowledge about Chemical Reactions†

J. GASTEIGER,* M. MARSILI,‡ M. G. HUTCHINGS,§ H. SALLER,∥ P. LÖW,∥ P. RÖSE, and K. RAFEINER

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

Received July 18, 1990

A variety of approaches to the calculation of data on chemical reactivity have been explored. They rely on values obtained from empirical methods for the quantification of fundamental chemical factors such as bond dissociation energy, charge distribution, and inductive, resonance, and polarizability effects. Equations have been derived for the calculation of parameters of linear free energy relationships, of data on gas-phase reactions, and of data on the selectivity and reactivity of Diels–Alder reactions, as well as for locating reaction sites in molecules. These models are incorporated into the program system EROS for the prediction of reaction mechanisms and the automatic generation of reaction rules from reaction databases.

INTRODUCTION

Chemists have accumulated a vast amount of information on chemical reactions. Nevertheless, they are often quite puzzled when they are faced with their daily questions:

- What will be the product of a chemical reaction? (yield)
- Which reaction conditions should I choose? (conditions)
- How fast will a reaction go? (rate)

The complexity of events determining the course of a chemical reaction is too high, and the factors influencing reactivity are too many to make a purely theoretical treatment of chemical reactions routine work. At present even single calculations on model systems have to make approximations and are time-consuming.

In this situation, chemists have to resort to comparisons with known information to derive rules or draw conclusions by analogy. This is the background where reaction databases, being built for several years, are finding their place.

However, even databases with more than a million individual reactions cannot ensure that a chemist finds answers to the above questions. There will always be many reactions that have not yet been added to the database, and new reactions cannot be contained there at all.

The task is therefore to use the information gathered on chemical reactions or stored in databases to learn more about reactions, to transform individual information into general knowledge.

Several approaches seem possible:

1. To define similarities between chemical reactions that allow conclusions to be drawn by analogy.
2. To derive rules by methods of artificial intelligence.
3. To derive mathematical equations that quantify chemical reactivity and are thereby allowed to make selections and predictions.

Our group has mainly chosen the last approach and has made major inroads over the last 15 years. The policy involved a two-step strategy: to develop procedures for the quantification of all-important chemical effects (chemical models) and to use the values from these methods for the reproduction and prediction of reactivity data (statistical models). Some of the methods developed and results obtained will be summarized here.

This work has been included in the EROS (Elaboration of Reactions for Organic Synthesis) system to predict automatically the products of chemical reactions and to assist in the design of organic syntheses. With progress in our insight, more advanced versions of the program have been issued.[1]

This paper can by no means give an overview of other approaches to the quantitative prediction of reactivity data. Rather, it gives a personal view, and therefore, only references to the work of our group are included.

THE FRAMEWORK: THERMOCHEMISTRY

In a chemical reaction, the coordinates of the atoms of the species involved change so as to follow a pathway of minimum energy. Chemical reactions are events on multidimensional energy hypersurfaces. In order to fathom the course of chemical reactions, pinpointing the local minima on such energy hypersurfaces is a good starting point. These points correspond to stable chemical species, whereas saddlepoints on these surfaces amount to transition states.

Additivity schemes have been quite successful for the estimation of heats of formation of organic compounds. In such a scheme, a structure is dissected into substructures, and these are associated with parameters that are added to arrive at a value for the heat of formation of the compound being considered. The Benson group method is a typical example of such an approach.[2] The method selected in our group starts with bond parameters that are further refined with values for 1,3- and 1,4-interactions. Values for these substructural parameters have been determined through multilinear regression analyses of experimental heats of formation.

Once the parameters have been determined they are entered in a table. A program has been developed that searches for all the substructures of a molecule and retrieves the appropriate parameters from the table.[3] Computation times are very short and only increase linearly with the number of atoms in a molecule. The accuracy of the approach is within 5 kJ/mol as can be seen from the typical values of Table I.

Once parametrized, the method also allows the prediction of heats of formation of compounds which have not been determined experimentally. Thus, the data gaps in Table I can be filled by predicted values.

Heats of reaction can be determined by evaluating the heats of formation of starting materials and products of a reaction. An additivity scheme can also be used for estimating other thermodynamic properties like entropies, heat capacities, etc.

The scheme can also be extended to radicals. This allows the calculation of *bond dissociation energies* (BDE) by evaluating the heats of formation of a compound and of the two radicals formed by breaking a bond.

$$A - B \xrightarrow{BDE} A^{\bullet} + {}^{\bullet}B$$

Table II gives a comparison of experimental and calculated values of a series of bond dissociation energies.

† Dedicated to Professor Ivar Ugi on the occasion of his 60th Birthday.
‡ Present address: Università di L'Aquila, Italy.
§ Present address: Imperial Chemical Industries, plc, Manchester, U.K.
∥ Present address: CHEMODATA Computer-Chemie, Munich, Germany.

0095-2338/90/1630-0467$02.50/0

Table I. Deviations between Experimental and Calculated Heats of Formation of Compounds A–B (Standard Deviation: 3.58 kJ/mol)[a]

A/B	H	F	Cl	Br	I	OH	NH_2
H	0.00	0.00	−0.01	0.00	−0.01	0.67	
CH_3	0.57		1.69	2.27	0.03	0.68	3.66
C_2H_5	0.26		−0.40	1.40	0.34	−0.30	2.88
n-C_3H_7	0.11	4.11	0.21	−1.71	−3.29	0.98	−0.22
i-C_3H_7	0.11	−15.07	1.53	−2.25	−1.92	−4.57	0.89
t-C_4H_9	−2.35		0.21	1.55	−0.12	−15.80	3.62
C_6H_5	6.61	1.36	5.64	11.46	7.98	7.92	9.82
$C_6H_5CH_2$	5.53		−0.16	−0.05	−0.08	−0.16	30.94
allyl	3.31			−0.13	−0.03	2.24	
CH_3CO	1.11	−0.01	−0.69	−0.06	−0.05	−1.48	2.68
C_2H_5O	−0.30					−23.10	
CH_2=CH	−1.18	−1.27	6.30	−0.75			

A/B	CH_3	C_2H_5	i-C_3H_7	t-C_4H_9	C_6H_5	CN	NO_2
H	0.57	0.26	0.11	−2.35	6.61		
CH_3	0.26	0.11	−2.35	−8.29	5.53	−2.36	0.80
C_2H_5	0.11	−0.13	−1.28	−5.81	5.16	−0.53	−3.58
n-C_3H_7	−0.13	−0.77	−1.72	−4.95	4.46	2.49	−5.38
i-C_3H_7	−2.35	−1.28	0.98	2.56	6.40	−3.06	−5.87
t-C_4H_9	−8.29	−5.81	2.56		8.58	2.48	−4.14
C_6H_5	5.53	5.16	6.40	8.58	8.48	−1.97	−0.36
$C_6H_5CH_2$	5.16	4.46	2.65		−15.29		−31.22
allyl	0.98	0.88	−1.53	−3.32		0.53	
CH_3CO	−1.84	−1.63	3.19	5.25	−0.22		
C_2H_5O	0.15	−2.73			12.50		−15.47
CH_2=CH	3.31	0.98	−2.14	−8.82	8.47	−0.22	

[a] Experimental values from ref 20.

Table II. Difference between Experimental and Calculated Dissociation Energies for the Bonds A – B → A• + •B (Standard Deviation = 5.20 kJ/mol)

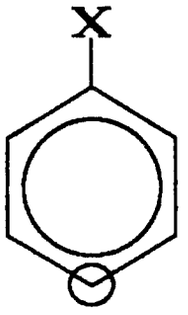
A/B	H	F	Cl	Br	I	OH	NH_2
H	0.00	0.00	−0.01	0.00	−0.01	0.67	
CH_3	0.38		1.50	2.08	−0.16	0.49	3.48
C_2H_5	−1.44		−2.10	−0.30	−1.36	−2.00	1.19
n-C_3H_7	−4.23	−0.23	−4.13	−6.05	−7.63	−3.36	−4.55
i-C_3H_7	−3.34	−18.52	−1.92	−5.70	−5.37	−8.02	−2.55
t-C_4H_9	−5.20		−2.64	−1.30	−2.97	−18.65	0.78
C_6H_5	7.07	1.82	6.10	11.92	8.44	8.38	10.29
$C_6H_5CH_2$	−1.48		−7.17	−7.06	−7.09	−7.17	23.94
allyl	−3.71			−7.15	−7.05	−4.78	
CH_3CO	1.15	0.03	−0.65	−0.02	−0.01	−1.44	2.73
C_2H_5O	0.33					−22.47	
CH_2=CH	−2.14	−2.23	5.34	−1.71			

A/B	CH_3	C_2H_5	i-C_3H_7	t-C_4H_9	C_6H_5	CN	NO_2
H	0.38	−1.44	−3.34	−5.20	7.07		
CH_3	−0.12	−1.78	−5.99	−11.33	5.80	−2.55	0.67
C_2H_5	−1.78	−3.53	−6.43	−10.36	3.92	−2.23	−5.22
n-C_3H_7	−4.66	−6.81	−9.51	−12.14	0.58	−1.85	−9.66
i-C_3H_7	−5.99	−6.43	−5.92	−3.74	3.41	−6.51	−9.26
t-C_4H_9	−11.33	−10.36	−3.74		6.19	−0.37	−6.93
C_6H_5	5.80	3.92	3.41	6.20	9.41	−1.51	0.16
$C_6H_5CH_2$	−2.04	−4.25	−7.81		−21.83		−38.17
allyl	−6.23	−7.84	−12.00	−13.19		−6.49	
CH_3CO	−1.99	−3.29	−0.22	2.44	0.28		
C_2H_5O	0.59	−3.80			13.59		−14.78
CH_2=CH	2.16	−1.68	−6.55	−12.63	7.97	−1.18	

The course of radical reactions is largely determined by the values of the bond dissociation energies. Thus, an all-important parameter for predicting radical reactions can easily be calculated.

Conclusion: Heats of reaction and bond dissociation energies can be calculated fast and with high accuracy. This allows determination of the thermochemical framework of a reaction. Furthermore, the predominant parameter governing radical reactions is available.

Problem: Many reactions are not under thermodynamic but kinetic control. How can kinetic effects be predicted?



1

Figure 1. Correlation of σ_R substituent constants with the π charge, q_π, on the p-carbon atom.

THE REACTION SITE IS KNOWN: AN LFER APPROACH

The first and widely accepted approaches to predicting data on chemical equilibria and on reaction rates are now summarized as linear free energy relationships (LFER). These include the Hammett and the Taft equation and a large series of offshoots therefrom. The central idea of the LFER approach is to dissect a system into a skeleton, the reaction site, and a substituent. The influence of the substituent on the reaction site is then expressed by a substituent constant, σ, that has been derived from a reference system.

It was found that more and more reference systems had to be defined when the approach was extended to a wider range of reactions. The reason is that the type and magnitude of interactions between the reaction under study and the reference reaction have to be quite similar. The proliferation of reference systems has led to quite a series of different tables of substituent constants (more than 50) and thus to a loss of generality of the LFER approach. In the end, it leaves the user with the problem of choosing the appropriate scale of substituent constants.

The artificial separation of a molecule into a skeleton, the reaction site, and a substituent lies at the basis of the loss of generality of the LFER approach. We therefore developed methods that calculate the influence of the various parts of a molecule on a reaction site directly by algorithms that walk through the entire molecule and do not make a distinction between a skeleton and a substituent.

One such approach is an empirical method for the calculation of *partial atomic charges* in a molecule.[4,5]

The σ bonds and the π systems have to be treated separately, but both parts are based on the idea of the dependence of orbital electronegativity on charge and on the concept of electronegativity equalization on bond formation. In the end, one obtains for each atom in a molecule uniquely defined σ- and π-charge values and σ- and π-electronegativity values that reflect the identity of this atom and its molecular environment.

The significance of the charge values for the analysis of reactivity data was studied with σ_R substituent constants that are valid for the ionization of substituted benzoic acids and related reactions. Indeed, a correlation was found between the σ_R constants and the π charges, q_π, in the para position for a series of monosubstituted benzene derivatives (eq 1).[6] (See Figure 1.)

$$\sigma_R = 0.018q_\pi - 0.101 \quad (1)$$

In essence, this result shows that for those reactions that can be described by σ_R constants, reactivity data can be directly calculated by inherent properties of atoms and algorithms that take account of the constitution of a molecule as a whole. The effects of the atoms of a substituent and the transfer of these

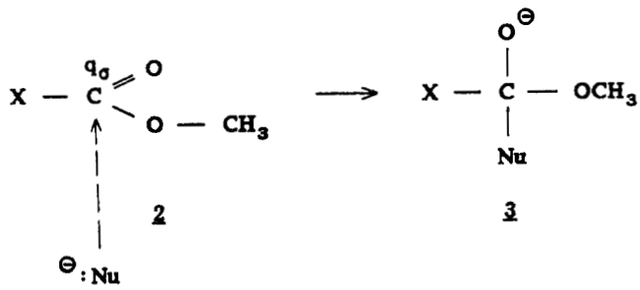


Figure 2. Reference system and equation for σ^* substituent constants; q_σ, σ charge on carbonyl carbon; χ_x, electronegativity; R_D, polarizability of substituent X.

effects through the molecule to the reaction site are automatically taken care of by the methods for charge calculation.

How general is such an approach? The disappointment immediately came when it was tried to correlate σ^* constants with the σ charge, q_σ, on the reaction site, the carbon atom of the ester group that had served as the reference system for the definition of σ^* constants. (See Figure 2.)

The direct correlation between σ^* and q_σ was rather poor. The problem was solved by a closer inspection of the reference reaction, the ester hydrolysis. The formation of the tetrahedral intermediate **3** is the rate-determining step. This reaction will not only be influenced by properties of the starting material **2** (e.g., q_σ at the carbon atom of the ester group) but also by factors stabilizing the tetrahedral intermediate **3**. The mechanistic model that we developed for this reaction step led us to eq 2.[7] This quantitative description of ester hydrolysis

$$\sigma^* = c_1 q_\sigma \chi_x + c_2 R_D + c_0 \qquad (2)$$

takes into account not only the charge distribution at the reaction site for the starting material **2** but also the inductive and polarizability effects stabilizing the reaction intermediate **3** as expressed by the electronegativity χ_x of the substituent and its polarizability R_D.

Equation 2 reflects the fact that chemical reactions are usually simultaneously under the influence of several chemical effects.

Conclusion: LFER approaches have shown their merits in quantitatively describing data on chemical equilibria and reaction rates. However, the profusion of different scales of substituent constants makes the selection of the appropriate substituent parameter a difficult task. This situation calls for a deeper understanding of the foundations of LFER. It could be shown that the magnitude and the transfer of the influence of substituents on a reaction site can be calculated directly as exemplified by the procedures for charge calculation. However, in general, chemical reactions are simultaneously influenced by various chemical effects.

Problem: How can the various chemical effects, e.g., inductive, resonance, polarizability effect, be put on a quantitative basis?

QUANTIFYING ADDITIONAL CHEMICAL EFFECTS: GAS-PHASE REACTIONS

Up until now we had developed procedures for calculating heats of formation, bond dissociation energies, and the charge distribution in molecules. The preceding studies showed that methods for the calculation of additional electronic effects had to be devised. In order to be able to study those effects in isolation, we turned our attention to gas-phase reactions.

Reactions in the gas phase allow the investigation of the reactivity of isolated molecules uncorrupted by the complicating influence of solvent effects. Furthermore, accurate experimental data had become available through ion cyclotron

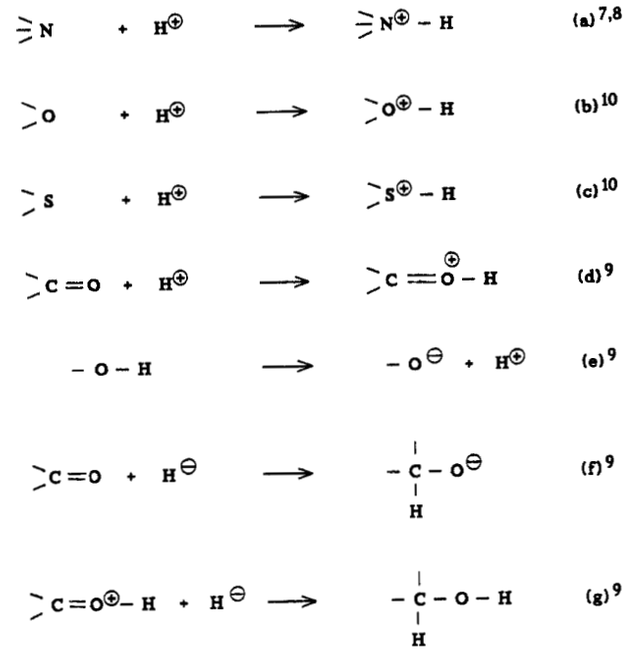


Figure 3. Gas-phase reactions for which correlations have been developed.

resonance and high-pressure mass spectrometry measurements.

It was clear from the very beginning that the methods for the calculation of the electronic effects had to be fast to be applicable to large molecules and extensive data sets. Therefore, we had to resort to empirical approaches and had to verify the significance of the methods by comparing the results with physical data and with the data on gas-phase reactions.

A simple additivity scheme incorporating a damping procedure to attenuate the influence of atoms that are further away from the reaction site was developed for the *polarizability effect*.[8]

The residual electronegativity data that are obtained concomitant with the partial charges in the procedures already mentioned above can be taken as a good measure of the *inductive effect*.[9]

The *hyperconjugation effect* can be quantified in simple systems by the number of C-H or C-C bonds.[10] In the more general case it was included in a procedure that calculates the stabilization of charges by the *resonance effect*.

Figure 3 summarizes the gas-phase reactions that were investigated. The data included values for the proton affinity of amines (a),[8,9] alcohols and ethers (b),[11] thiols and thioethers (c),[11] aldehydes and ketones (d),[9] and gas-phase acidity values of alcohols (e).[10] Furthermore, hydride ion affinity data of carbonyl compounds both in the neutral (f) and in the protonated form (g) were investigated.[10]

The data of all those chemical reactions could be reproduced by linear equations of the form of eq 3.

$$\Delta H_r = c_1 \chi_r + c_2 \alpha_d + c_3 N_{hyp} + c_0 \qquad (3)$$

In eq 3, χ_r is a residual electronegativity value of the substituents at the free sites of the reactions of Figure 3 as a measure of the inductive effect, α_d is the damped polarizability effect, and N_{hyp} gives the number of C-H and C-C bonds as a measure of the hyperconjugation effect.

These effects influence the various reactions of Figure 3 to different extents. Table III indicates which coefficients of eq 3 are needed and are therefore different from zero for the various reactions of Figure 3.

Thus, whereas the proton affinity of alkylamines can be calculated with a simple one-parameter equation from the

Table III. Coefficients Needed in Equation 3 for the Reactions of Figure 3[a]

reaction (see Figure 3)	c_1	c_2	c_3
(3a) PA of alkylamines only		X	
(3a) PA of substituted amines	X	X	
(3b) PA of ethers and alcohols	X	X	
(3c) PA of thiols and thioethers	X	X	
(3d) PA of carbonyls	X	X	X
(3e) acidity of alcohols	X	X	
(3f) HIA of carbonyls	X	X	X
(3g) HIA of protonated carbonyls	X	X	X

[a] PA = proton affinity; HIA = hydride ion affinity.

value of the polarizability effect, a measure of the inductive effect has to be included for amines also containing substituents with heteroatoms. These two parameters, inductive effect and polarizability effect, also suffice to calculate the proton affinities of alcohols and ethers, as well as of thiols and thioethers, and the gas-phase acidity of alcohols.

The hyperconjugation effect has to be included in all those reactions that involve carbenium ions, the protonation of carbonyl compounds, and the hydride ion addition to carbonyl compounds in the neutral and the protonated form.

The success of eq 3 in quantitatively reproducing data of such a variety of reactions as assembled in Figure 3 is remarkable. They encompass all different charge types of polar processes: the reaction of a neutral molecule with a positive or negatively charged species, the dissociation of a neutral molecule into two ions of opposite charge, and the combination of two ions to a neutral species. This broad range of reactions underscores the quality of the parameters calculated for the inductive, polarizability, and hyperconjugation effect and the global validity of separating reactivity data into additive contributions coming from these effects.

In passing, it should be mentioned that these quantitative analyses of gas-phase reactions do form a secure foundation for investigating data in solution. One such study has been performed with pK_a values of alcohols comparing the data from aqueous solution with those of the gas-phase acidity study.[12]

Conclusion: Parameters can easily and rapidly be calculated for the inductive, polarizability, and resonance effect by newly developed methods. These parameters have shown their merit in quantitatively reproducing reactivity data of fundamental polar reactions in the gas phase. Thus, a firm basis for trying to understand data of reactions in solution has been established.

Problem: The reaction site has been known in the reactions studied in this chapter. Can the values on the various electronic effects be used to search for the reaction site?

LOCATING THE REACTION SITE: FUNCTIONS FOR BREAKING AND MAKING BONDS

It has always been a basic intellectual exercise in our group to treat chemical reactivity without resorting to a search for functional groups. There must always be some reason why certain bonds contained in functional groups are reactive. Our work has tried to decipher these reasons and to put them on a quantitative basis. This should, in addition, enable one to evaluate the relative reactivity of various functional groups in a molecule and thus make predictions of the course of reactions when competing groups are present.

Reactivity Space. In the previous sections, methods for calculating heats of formation, bond dissociation energies, charge distributions, and inductive, polarizability, and resonance effects were introduced. These effects are decisive factors in determining the reactivity of bonds. The situation is complicated by the fact that these effects operate simultaneously and they do so to various extents. To analyze this

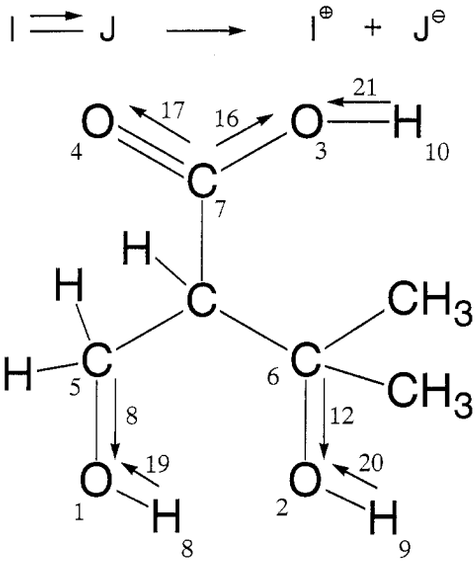


Figure 4. Numbering of atoms and polar bond breakings in 2-hydroxymethyl-3-hydroxy-3-methylbutanoic acid, **4**, with indications of the shift of the electron pair.

dependence of chemical reactivity on several parameters, we have defined reactivity spaces that have these energy and electronic effects as coordinates. A bond in a molecule is represented by a point in such a reactivity space with the values calculated by the above procedures for the various chemical effects as coordinates. More exactly this one-to-one correspondence between a bond and a point in a reactivity space only applies to the homolytic breaking of a bond. On the other hand, there are two possibilities for shifting the electron pair in the case of a polar breaking of a bond, leading to two alternatives of creating a positive and a negative fragment. Some of the above effects are different for these two alternatives of heterolysis of a bond, e.g., the polar breaking can occur in line or against the inherent polarity, or the electronegativity difference of this bond. Furthermore, resonance stabilization of charges will depend on which atoms the charges appear. Thus, the two choices for the polar breaking of a bond will be characterized by two points in a reactivity space.

An example will serve to illustrate a reactivity space and its merits for studying chemical reactivity. Structure **4**, 2-hydroxymethyl-3-hydroxy-3-methylbutanoic acid, was chosen to exemplify the different acidities of several OH groups and their dissimilar leaving group potentials. The heterolyses of the bonds to be considered are indicated in Figure 4.

The molecule **4** contains 14 constitutionally nonequivalent bonds, corresponding to 28 different ionic bond breakings. These heterolyses are indicated in Figure 5 as points in a reactivity space spanned by the resonance effect (R), the difference in σ electronegativity ($\Delta\chi_\sigma$), and the bond polarity (Q_σ) as coordinates.

It can be seen that the reactive bonds clearly separate from the nonreactive bonds. Thus, the property to study, chemical reactivity, is well represented in that space; the parameters chosen as coordinates attain values that can quantify reactivity.

Closer inspection of the reactivity space allows the extraction of more detailed information on the relative reactivity of the individual bonds. This can be achieved by looking at different projections of the reactivity space. Interactive computer graphics is a powerful tool for this purpose. However, it cannot be well reproduced in a static manner as required by a printed medium. Instead, a projection of the points in the reactivity space onto the three planes defined by combinations of two parameters each is given here in Figure 6.

These projections permit the simultaneous study of the influence of two chemical effects on reactivity. Clearly this is

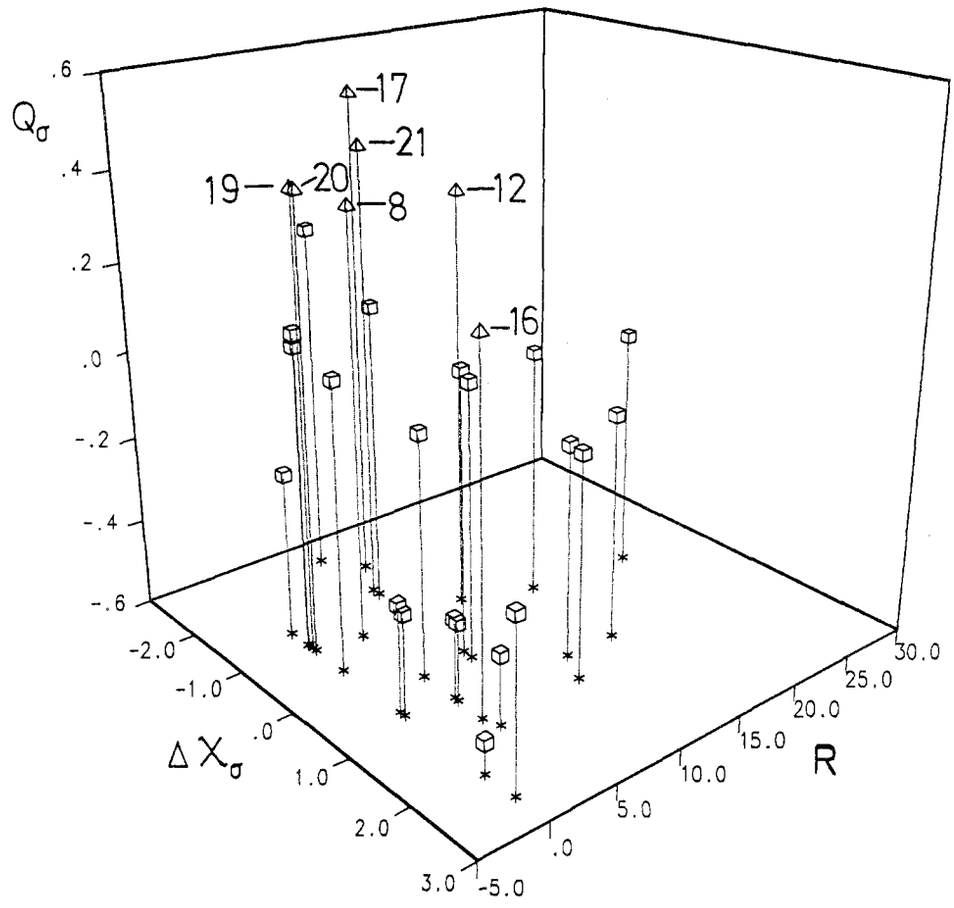


Figure 5. Heterolyses of the bonds in molecule **4**, in a reactivity space with the resonance effect, R, the σ electronegativity difference, $\Delta\chi_\sigma$, in [eV], and the bond polarity, Q_σ, in [e], as coordinates. Nonreactive bonds are indicated by small cubes, reactive ones by square pyramids. The numbering of the reactive bonds corresponds to that in Figure 4.

a drawback against an investigation of the full three-dimensional space where the influence of all three factors can be seen in concert. Nevertheless, one can arrive in steps at the same conclusions.

Electronegativity difference ($\Delta\chi_\sigma$) and resonance effect (R) together do not suffice to separate reactive and nonreactive bonds. On the other hand, there is a clearly visible separating line between reactive and nonreactive bonds in the plot of electronegativity difference against σ-bond polarity ($\Delta\chi_\sigma$ vs Q_σ). However, this two-dimensional plot does not give the full information as it cannot make a marked difference between the generation of a tertiary (heterolysis 12) and a primary carbocation (heterolysis 8). This difference only comes to full light in the plot of the resonance effect against the bond polarity (R vs Q_σ). Heterolysis 12 is distinguished by a higher value of the resonance effect, showing that the farther the bonds are to the right-hand side in this plot, the more reactive they are.

This is also indicated for the losses of a proton from the three different OH groups. There is hardly any reactivity difference for the deprotonation of a primary or a tertiary alcohol (points 19 and 20). However, the loss of a proton from the carboxylic acid is distinguished by a clear separation of point 21 to the right-hand side of the plot of the resonance effect against the σ-bond polarity, the direction that has been perceived above as indicating higher reactivity.

To summarize, the more reactive bonds in the reactivity space of Figures 5 and 6 are to be found in the upper part of the space and into the corner farthest away from the observer. Thus, it can be concluded that the heterolyses corresponding to the points 17, 21, and 12 are the most reactive ones. These correspond to nucleophilic attack at the carbonyl group, loss of a proton from the carboxylic acid, and loss of hydroxide to give the more stable tertiary carbocation, respectively.

Reactivity spaces constitute powerful tools for studying the influence of various electronic and energy parameters on chemical reactivity. By taking different sets of three factors, the simultaneous influence of these three effects can be visualized and analyzed with graphical projections such as those of Figures 5 and 6.

More detailed investigations of the aldol condensation, the Grob fragmentation, and the haloform reaction have been published elsewhere.[13-15]

Statistical Analysis. It is even possible to investigate the coincident influence of more than three chemical effects when one is prepared to sacrifice direct visualization by computer graphics. We have extensively studied reactivity spaces of six dimensions with a host of statistical and pattern recognition methods. This allows us to unravel the relative importance of the various factors on the reactivity of specific bonds and reactions.

Multilinear regression analysis can be used to derive reactivity functions when quantitative data on chemical reactivity are available. Unfortunately, this favorable situation is quite often not given. However, reactivity functions can be derived even in cases where one can only give a classification whether a bond is reactive or not. Methods such as linear discriminant analysis or logistic regression analysis can be used to develop functions that reproduce this classification and thus give a measure of chemical reactivity. Obviously, a large and carefully selected data set has to be used when such a coarse measure of reactivity—reactive or not—is administered.

In a typical study, a data set of 29 aliphatic molecules containing bonds that encompass 770 different polar bond

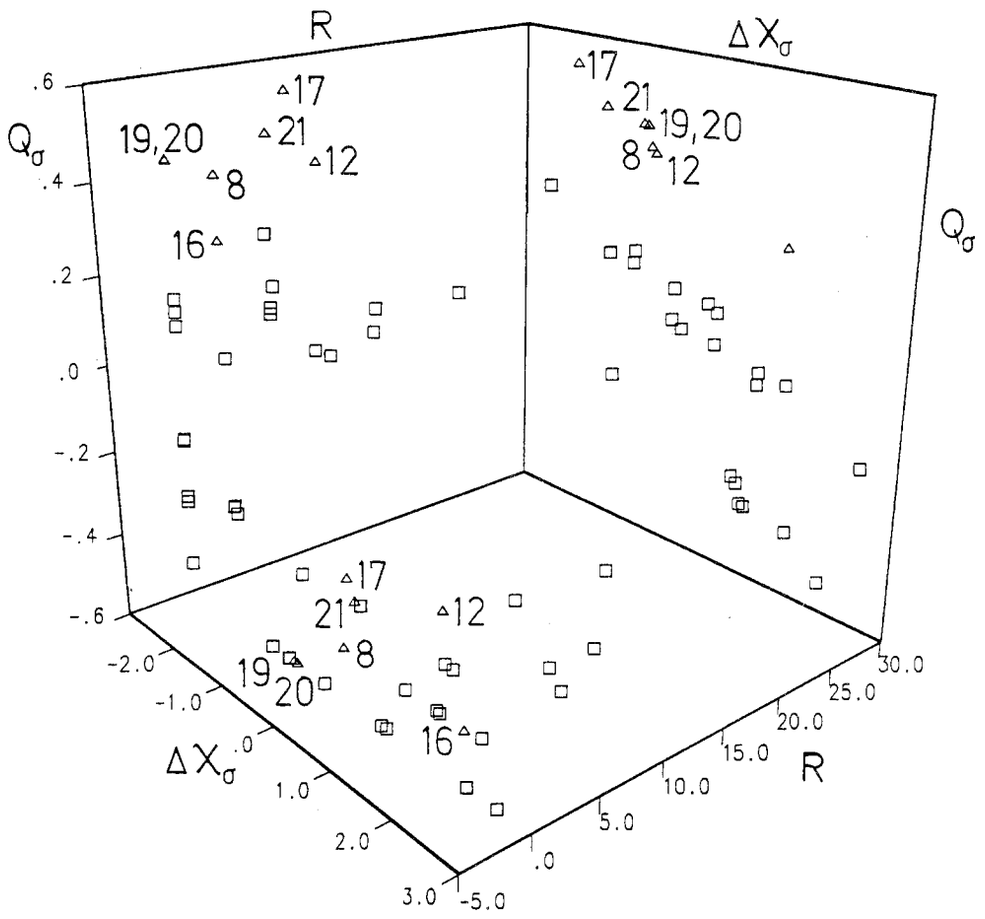


Figure 6. Projection of the points in the reactivity space of Figure 5 onto three planes composed of two parameters only. (R, resonance effect; $\Delta\chi_\sigma$, in [eV], σ electronegativity difference; Q_σ, in [e], σ bond polarity).

breakings was used to develop a function for quantifying the reactivity of aliphatic single bonds.[16] From this data set, 116 bonds were selected and classified as either reactive (42 bonds) or nonreactive (74 bonds).

Logistic regression analysis (LoRA) uses the combination of the two functions (eqs 4 and 5) to reproduce the initial

$$P = \frac{1}{1 + e^{-f}} \tag{4}$$

$$f = c_0 + c_1x_1 + c_2x_2 + \ldots = \mathrm{RI} \tag{5}$$

classification (reactive = 1, nonreactive = 0) as a probability, P.

The various chemical effects calculated by our methods were taken as independent variables, x_i, in eq 5. The large number of bonds and molecules for which the values of charge distribution, bond dissociation energy, and inductive, resonance, and polarizability effect have to be evaluated again underscores that these methods have to be rapid to handle such large data sets.

Those parameters selected as important by LoRA were also found as primarily relevant by principal component analysis and linear discriminant analysis. The function, f, obtained by LoRA was taken as a measure of reactivity, the reactivity index, RI.

In a similar fashion, reactivity functions have been developed for single bonds in charged species, for double bonds in neutral as well as for double bonds in charged species, and for aromatic bonds.

The reactivity index, RI, can be taken as a measure of the prospensity of a bond for heterolysis. Thus, those bonds in a molecule with the highest values of the reactivity index locate the reaction site.

In addition to the functions that determine the ease of breaking a bond, functions for making a bond have also been developed: one for neutral molecules and one for charged species.

The automatic search for the reaction site has achieved success. It starts with an evaluation of the magnitude of the various chemical effects for each bond of a molecule. These values are entered into the appropriate reactivity function. The highest values of the reactivity index define the reaction site. Functional groups are thus found based on the first principles. However, the approach goes beyond the concept of functional groups. It also locates the most reactive bonds in those cases where functional groups would give the wrong answer as exemplified by the Grob fragmentation and the haloform reaction.[14,15] C-C single bonds, not normally considered as functional groups, become reactive in these reactions. The reasons can only be found by working in steps through the reaction mechanism. Then one can see that by breaking a C-C bond in an intermediate, ions are generated that have extra factors for stabilization. This is one strength of our approach.

EROS 5.2. The procedures for calculating the chemical effects and the reactivity functions that determine which bonds to make and break have been incorporated into the EROS system and define the version 5.2 that can be applied to the prediction of the course of chemical reactions and thus lead to the reaction products. This is achieved by modeling the mechanism of a reaction as a sequence of polar bond breakings and makings.

It is clear that the condensation of the plurality and complexity of chemical reactions into five functions for breaking a bond and two for making a bond constitute a drastic oversimplification of the problem. Nevertheless, the mechanisms and products of a series of complex organic reactions have been

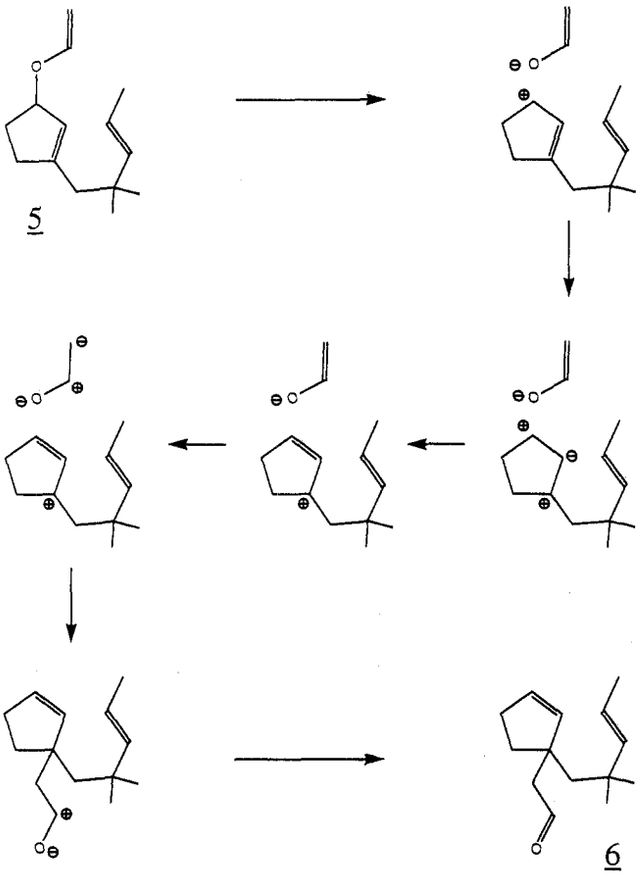


Figure 7. Reaction mechanism obtained by EROS 5.2. Charges indicate the direction of heterolysis and are not fully developed.

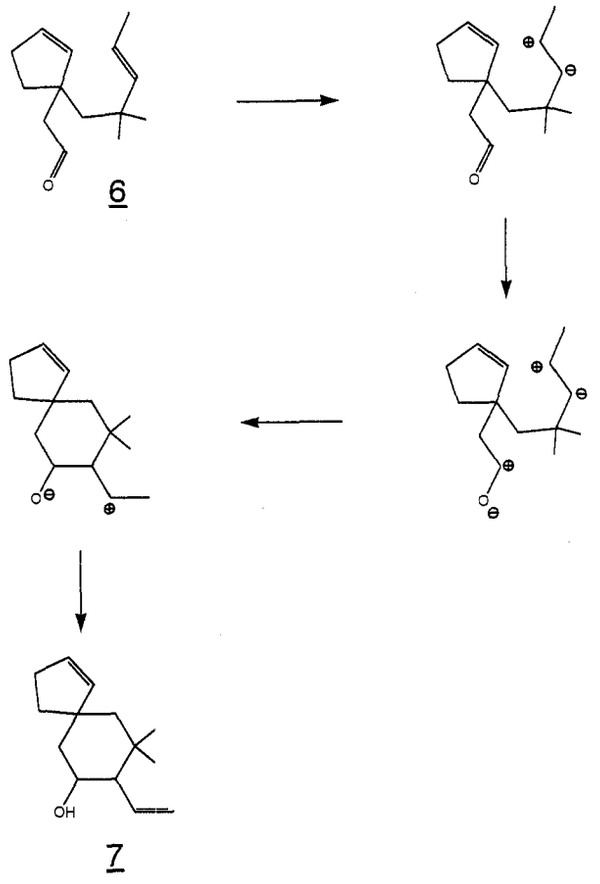


Figure 8. Reaction mechanism modeled by EROS 5.2.

correctly predicted.[1] In particular, the course of capricious rearrangement reactions could be elucidated.

Figures 7 and 8 give a sequence of rearrangements from **5** to **6**, and from **6** further to **7**. It should be noted that these sequences are the result of the evaluation of about 50 intermediates. Not all of those are given so as not to overburden the reaction schemes. Only those intermediates lying on the direct path from the starting materials to the product have been included. The predicted reaction products **6** and **7** also are the ones observed experimentally.[17] Clearly, the reaction of **5** to **6** is a concerted Claisen rearrangement which is modeled here in a stepwise manner. Some other intermediates lead to products that do seem quite likely candidates. In fact, rearranging **6** at higher temperature gives additional products that have precursors found in the tree of reaction intermediates.

Conclusion: Methods have been developed for the calculation of fundamental electronic and energy effects. These provide parameters to be included in reactivity functions to determine which bonds will be broken or made in a reaction. Thus, the reaction site in a molecule can be determined. The course and products of complex reactions can be derived from the modeling of reaction mechanisms as a sequence of bond-breaking and bond-making steps. The methods have been included in a program system for the prediction of the course of chemical reactions.

Problem: The dissection of a reaction mechanism into a sequence of polar bond breakings and bond makings is an oversimplification of the problem. Many reactions occur with the simultaneous breaking and making of bonds. In addition, it must be clear that five reactivity functions for determing the breakability and two for calculating the ease of making a bond cannot suffice to embrace the whole range of organic reactions. How can progress be made in both areas?

FUNDAMENTAL PROCESSES: USING REACTION DATABASES

Any chemical reaction can be obtained through sequences of individual (heterolytic or homolytic) bond-breaking and bond-making steps. However, such a breakdown of a reaction into single steps of bond breaking and bond making might not correspond to the actual mechanism of a reaction. For example, such an approach will not make a distinction between an S_N1 and an S_N2 process—the S_N2 process would also be modeled as a two-step process of polar bond breaking and bond making.

The truth is that many reactions proceed by the simultaneous breaking and making of bonds, by the shifting of more than one electron pair. Any process that involves the concerted shift of electrons is considered an elementary step.

Thus, elementary reaction steps are the division of the two electrons of a single bond into two radicals (homolysis), heterolysis of a single bond (e.g., the initial step in an S_N1 reaction), the simultaneous polar bond breaking and bond making (S_N2 process), or a concerted shift of three electron pairs (e.g., Diels–Alder reaction) (Figure 9).

Any endeavor to develop a deeper understanding of reactivity in organic chemistry must be able to predict the onset and the causes for such elementary steps.

EROS 6.0. The task is threefold. (1) All the essential elementary processes controlling organic reactions have to be collected. (2) The causes of one elementary reaction in preference to another, in a given situation, have to be found and generalized. (3) And, finally, reactivity functions for the various elementary processes have to be developed.

In order to prepare the EROS system for solving these problems, the system was redesigned. Up to version 5.2, the reaction generators to break and make bonds and the reactivity

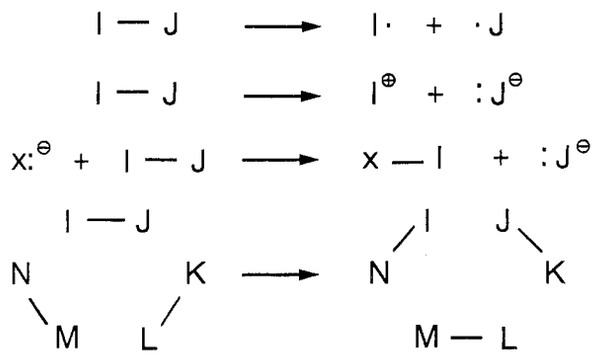


Figure 9. Elementary processes, proceeding with a concerted shift of electrons.

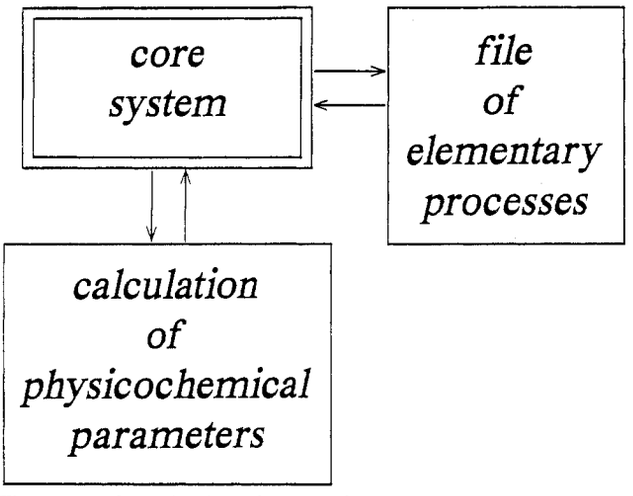


Figure 10. Organization of the EROS 6.0 version.

functions for evaluating reactions were contained in subroutines of the system. In the new version 6.0, a stronger separation of the various parts of the system was made (Figure 10).

The core system performs the analysis and manipulation of structures and reactions, as well as the evaluations and decision-making processes. These are based on information obtained for the various electronic and energy effects through the procedures mentioned in the previous sections. Knowledge of the various elementary processes is kept in a separate datafile, the rule file. Each elementary process is a chapter of its own in this rule file, containing restrictions for the atoms, bonds, and molecules in order that this rule can be applied. In addition, each rule contains a reactivity function for evaluating this elementary process. The rule file is an ASCII file which makes changing of the rules by editing an easy process. Thus, the knowledge base of the system contained in this rule file can be inspected and changed without difficulty.

However, the question is, how can the knowledge be acquired that is to be stored in the rule file to begin with? A variety of methods can be used. Firstly, the knowledge contained in EROS 5.2 in subroutines can be transfered into EROS 6.0 by storing the various reactivity functions as separate rules in the rule file.

Furthermore, all the model-building methods mentioned in the previous chapters can be employed. As an example, logistic regression analysis (LoRA) was used to develop a function that distinguishes between an S_N1 or S_N2 process. A data set of structures was built, and the entries were classified as either reacting according to an S_N1 process or not. This binary information was modeled by a probability distribution through LoRA by the combined application of eqs 4 and 5 using parameters on electronegativity difference, on charge stabilization, and on the polarizability effect. The resulting function reproduced the initial classification S_N1/S_N2 well and also gave good performance in predicting cases not contained in the training set. And finally, the information contained in reaction databases can be put to use.

Use of Reaction Databases. The increasing availability of computer-readable reaction databases led us to search for methods for the automatic extraction of knowledge on chemical reactions from a reaction database. We have developed an approach that condenses essential information in a reaction database into a reaction rule, a chapter of the rule file (Figure 11).[18]

The information of a reaction database is standardized and checked for errors before being entered into the rule-generation phase. At the center of rule generation is a structure–reactivity analysis that uses parameters on electronic and energy effects calculated by the methods mentioned above. The only human intervention is made in this structure–reactivity analysis by the selection of various parameter combinations, different mathematical equations, and different statistical methods. Apart from this possibility for the researcher to interfere in

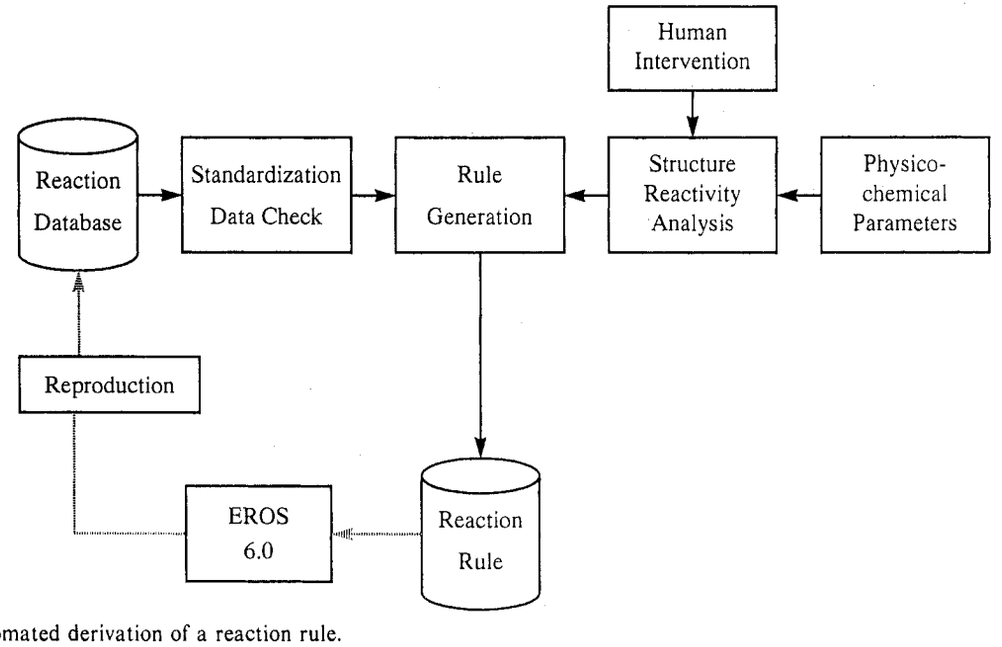


Figure 11. Automated derivation of a reaction rule.

the building of the chemical and statistical model, the rest of the rule generation is performed completely automatically.

The reaction rule obtained by such a process is entered into the rule file of the EROS 6.0 system. Then, an immediate check of the performance of the rule in reproducing the original data in the reaction database is made.

Diels–Alder Reaction. The Diels–Alder reaction serves as an illustration for this process. Two important aspects of the Diels–Alder reaction have been investigated: the *regioselectivity* of this reaction governs the substitution pattern in the products to give one or a ratio of two isomers. The *reactivity* determines the rate of forming the products.

A database of 148 Diels–Alder reactions containing information on isomer distributions was extracted from the literature. From the yield of isomers, [x] and [y], a value for the selectivity Sel was calculated:[19]

$$\mathrm{Sel} = \frac{[x] - [y]}{[x] + [y]} \times 100 \qquad (6)$$

This value leads to a difference in the free energy of activation $\Delta\Delta G^{\ddagger}$ under the assumption that $T\Delta\Delta S^{\ddagger} \ll \Delta\Delta H^{\ddagger}$:

$$\Delta\Delta G^{\ddagger} = -RT \ln \frac{100 + \mathrm{Sel}}{100 - \mathrm{Sel}} \qquad (7)$$

The distribution of isomers in Diels–Alder reactions is determined by the directing power of the diene and the dienophile. These were calculated from electronic effects in the two reactants. For the diene the difference in the +M effect in positions 1 and 4, the difference in the +M effect in positions 2 and 3, and the electronegativity difference of atoms 1 and 4 were taken as quantitative measures of the directing power, $D_{4\pi}^{(i)}$.

For the dienophile, the difference in electronegativity and the difference in the –M effect of the two atoms participating in the reaction were considered as measures of the directing power, $D_{2\pi}^{(j)}$.

Three products, V_{ij}, were taken from these electronic effects; the product of the inductive effect (electronegativity difference) in the diene with the one in the dienophile, the product of the combined resonance effects in the diene with the one in the dienophile, and the product of the combined resonance effect in the diene with the inductive effect in the dienophile:

$$V_{ij} = D_{4\pi}^{(i)} D_{2\pi}^{(j)} \qquad (8)$$

Finally, the difference in the free enthalpy of activation was fitted to these three products by linear regression analysis:

$$\Delta\Delta G^{\ddagger} = c_1 V_1 + c_2 V_2 + c_3 V_3 \qquad (9)$$

The correlation was only moderate ($r = 0.85$). However, it allowed the correct prediction of the major product for 139 of the 148 reactions when the quality of the function in reproducing the primary data was tested (Figure 11). Deviations occurred for highly polar dienophiles, or when steric effects came in.

Thus, we had managed to condense the information on the distribution of isomers of about 150 Diels–Alder reactions into a function that allows the prediction of the product ratio, the *regioselectivity* of Diels–Alder reactions to a reasonable accuracy and with high reliability. This process of concentration of information, of knowledge acquisition, runs to a large extent automatically. The only human intervention necessary—and wanted—is the building of a physicochemical and mathematical model as embodied in eqs 6–9.

In a similar manner, a database of 170 Diels–Alder reactions was built, containing information on kinetic data for the reaction of symmetrical dienes and dienophiles. The free energy of activation of these reactions was regressed against values of the inductive and resonance effects in the diene and the dienophile. This gave a function for quantifying the *reactivity* in Diels–Alder reactions.

Conclusion: Methods have been developed to extract knowledge of chemical reactions from reaction databases. The essential characteristics of a series of chemical reactions can be stored in a reaction rule. The EROS 6.0 system has such a rule file for predicting the products of chemical reactions. The clear-cut separation of knowledge base and inference methods facilitates future development.

Problem: The depth of knowledge to be extracted from a reaction database is heavily dependent on the quality of information stored in the database. All the information known for a given reaction should be stockpiled. Important features to be given for a reaction are

- reaction conditions
- product ratios and yields
- mechanistic information
- kinetic data

Unfortunately, none of the available reaction databases lives up to this standard. Presently there is no commercial reaction database that gives information on the individual steps of a reaction mechanism or contains kinetic data. Some databases even give only one product for each reaction, thus blurring any information on product ratios. It is clear that the knowledge to be gained on chemical reactions has to be rather rudimentary when only such crude information is available.

We have to strive for better reaction databases and, before that, more thorough experimentation including full kinetic analysis.

SUMMARY

The products and their rates of formation for a wide range of organic reactions can be derived from simple equations. These mathematical models are based on the values of electronic and energy effects. The applications range from the calculation of parameters of linear free energy relationships to the correlation of data on fundamental gas-phase reactions, and to the automatic determination of reactive bonds in a molecule. Methods have been developed for the direct deduction of knowledge of chemical reactions from the information contained in reaction databases. The equations and methods have been incorporated into the EROS system which can predict the course and products of organic reactions by explicit modeling of their mechanisms.

ACKNOWLEDGMENT

We express our gratitude to other members of our group for their contributions to the overall objectives. Financial support of our work came from Deutsche Forschungsgemeinschaft, ICI plc, U.K.; Sumitomo Chemical Co., Japan; Tecnofarmaci, SpA, Italy; Stiftung Volkswagenwerk; Verband der Chemischen Industrie; and the Bundesminister für Forschung und Technologie.

REFERENCES AND NOTES

(1) Gasteiger, J.; Hutchings, M. G.; Christoph, B.; Gann, L.; Hiller, C.; Löw, P.; Marsili, M.; Saller, H.; Yuki, K. A New Treatment of Chemical Reactivity: Development of EROS, an Expert System for Reaction Prediction and Synthesis Design. *Top. Curr. Chem.* **1987**, *137*, 19–73.

(2) Benson, S. W. *Thermochemical Kinetics*, 2nd ed.; Wiley: New York, 1976.

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

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

(5) Gasteiger, J.; Saller, H. Berechnung der Ladungsverteilung in konjugierten Systemen durch eine Quantifizierung des Mesomerie-konzeptes. *Angew. Chem.* **1985**, *97*, 699–701; *Angew. Chem., Int. Ed. Engl.* **1985**, *24*, 687–689.
(6) Marsili, M.; Gasteiger, J. Pi-Charge Distributions from Molecular Topology and Pi-Orbital Electronegativity. *Croat. Chem. Acta* **1980**, *53*, 601–614.
(7) Gasteiger, J.; Marsili, M.; Paulus, B. In *Data Processing in Chemistry*; Hippe, Z., Ed.; Elsevier: Amsterdam: 1981, pp 229–246.
(8) Gasteiger, J.; Hutchings, M. G. Quantification of Effective Polarisability. Applications to Studies of X-ray Photoelectron Spectroscopy and Alkylamine Protonation. *J. Chem. Soc., Perkin Trans. 2* **1984**, 559–564.
(9) 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.
(10) 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., Perkin Trans. 2* **1986**, 447–454.
(11) Gasteiger, J.; Hutchings, M. G. Quantitative Models of Gas-Phase Proton Transfer Reactions Involving Alcohols, Ethers, and their Thio Analogues. Correlation Analyses Based on Residual Electronegativity and Effective Polarizability. *J. Am. Chem. Soc.* **1984**, *106*, 6489–6495.
(12) Hutchings, M. G.; Gasteiger, J. Correlation Analyses of the Aqueous Phase Acidities of Alcohols and Gem-Diols, and of Carbonyl Hydration Equilibria, using Electronic and Structural Parameters. *J. Chem. Soc., Perkin Trans. 2* **1986**, 455–462.
(13) Gasteiger, J.; Hutchings, M. G.; Saller, H.; Löw, P. In *Chemical Structures*; Warr, W. A., Ed.; Springer: Heidelberg, 1988; pp 343–359.
(14) Gasteiger, J.; Röse, P.; Saller, H. Multidimensional Explorations into Chemical Reactivity: The Reactivity Space. *J. Mol. Graphics* **1988**, *6*, 87–97.
(15) Gasteiger, J.; Ihlenfeldt, W. D.; Röse, P.; Wanke, R. Computer-Assisted Reaction Prediction and Synthesis Design. *Anal. Chim. Acta* **1990**, *235*, 65–75.
(16) Gasteiger, J.; Saller, H.; Löw, P. Elucidating Chemical Reactivity by Pattern Recognition Methods. *Anal. Chim. Acta* **1986**, *191*, 111–123.
(17) Ziegler, F. E.; Mencel, J. J. *Tetrahedron Lett.* **1984**, 123–126.
(18) Röse, P.; Gasteiger, J. Automated Derivation of Reaction Rules for the EROS 6.0 System for Reaction Prediction. *Anal. Chim. Acta* **1990**, *235*, 163–168.
(19) Röse, P.; Gasteiger, J. In *Software Development in Chemistry IV*; Gasteiger, J., Ed.; Springer: Heidelberg, 1990; pp 275–287.
(20) Pedley, J. B.; Naylor, R. D.; Kirby, S. P. *Thermochemical Data of Organic Compounds*, 2nd ed.; Chapman and Hall: London, 1986.