Automatic Extraction of Chemical Knowledge from Organic Reaction Data: Addition of Carbon–Hydrogen Bonds to Carbon–Carbon Double Bonds

Lingran Chen, Johann Gasteiger,* and John R. Rose†

Computer-Chemie-Centrum, Institut für Organische Chemie, Universität Erlangen-Nürnberg, Nägelsbachstr. 25, 91052 Erlangen, Germany

Received March 16, 1995®

Historically, chemists have gained a large portion of their understanding of chemical reactions through inductive learning and analogical reasoning by considering individual reactions. The advent of reaction databases supports the development of machine-learning techniques for the automatic acquisition of knowledge on chemical reactions. Hierarchical classification and generalization can collect similar reactions into a reaction type and detect those features—functional groups and electronic effects—that are necessary for a reaction type as well as those that are dispensable. HORACE, a system based on these principles, was applied to a dataset of 120 reactions involving the addition of a C–H bond to a C=C double bond, a reaction that comprises such important reaction types as Michael additions, Friedel–Crafts alkylation, and free radical additions. The automatic classification of these reactions was able to find these and other important reaction types, indicated the scope and limitations of these reaction types, and discovered some more unusual processes.

Introduction

In recent years computer-readable reaction databases that are collections of many individual reactions have become available. Some of them contain more than a million reactions. The various search methods built into reaction retrieval systems such as STN-Messenger[1] or REACCS[2] provide a variety of avenues for retrieving the information on individual reactions. These systems output sets of individual reactions, sets that may become with the more voluminous reaction databases, to the chagrin of the user, quite large and thus laborious to digest.

It is tempting to go beyond simple retrieval methods and use the computer-readable information on organic reactions as a basis for automatic knowledge acquisition. One would like to develop computer methods that model the inductive approach to learning from information on individual reactions. This is an approach that chemists have been successful with over many decades. In fact, several research groups have already embarked on the development of systems that extract knowledge on chemical reactions from reaction databases. This knowledge can be used for the prediction of the course of chemical reactions or for the design of organic syntheses.[3–7]

We have recently reported on the HORACE system (Hierarchical Organization of Reactions by Attribute and Condition Eduction) and have shown how it can be used for the hierarchical organization of reactions.[6,7] Reactions are iteratively classified and generalized into reaction types. Such a hierarchical organization of individual reactions is of considerable importance when a search in a reaction database provides a large set of answers, providing more reactions than the user is willing to scan sequentially. A hierarchical organization of such a set of reactions allows the user to better zero in on the reaction type and thus locate the subset of individual reactions in which she/he is most interested.

Furthermore, the knowledge automatically extracted from reaction databases provides a rich source of chemical information for developing reaction prediction and synthesis design systems. We will use the potential of HORACE to generate knowledge bases for EROS (Elaboration of Reactions for Organic Synthesis),[8] our system for reaction prediction, as well as for WODCA (Workbench for the Organization of Data for Chemical Applications),[9] our system for synthesis design.

In this paper we will concentrate on the potential of HORACE to extract knowledge from reaction databases, to group reactions into various types, to show the scope of such reaction types, to indicate novel or uncommon reactions, and to give hints as to the driving forces and mechanisms of a reaction type. Thus, emphasis is put on the benefits an organic chemist might obtain by using HORACE. The methods built into HORACE are only detailed to an extent deemed necessary to understand the kind of knowledge that can be gained on organic reactions.

Objectives and Basic Methods

A clear definition of our vocabulary will facilitate the exposition of our approach and the subsequent discussion of the results. The *reaction center* is made up of the atoms and

* Phone: +49-(0)9131–856570. Fax: +49-(0)9131–856566. e-mail: Gasteiger@EROS.CCC.Uni-Erlangen.de.

† Department of Computer Science, The University of South Carolina, Columbia, SC 29208. e-mail: Rose@cs.sc.edu.

® Abstract published in *Advance ACS Abstracts,* November 1, 1995.

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

(2) REACCS and ISIS/HOST are available from MDL Information Systems, Inc., San Leandro, CA.

(3) Gelernter, H.; Rose, J. R.; Chen, C. *J. Chem. Inf. Comput. Sci.* **1990**, *235*, 163.

(4) Chen, L.; Gasteiger, J.; Rose, J. R. *Software Development in Chemistry 9*; Moll, R., Ed.; Gesellschatt Deutscher Chemiker: Frankfurt a. Main, Germany, 1995.

(5) Hendrickson, J. B.; Miller, T. M. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 403.

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

(7) Gasteiger, J.; Rose, J. R. *Software Development in Chemistry 8*; Jochum, C., Ed.; Gesellschaft Deutscher Chemiker: Frankfurt a. Main, Germany, 1994.

(8) Röse, P.; Gasteiger, J. *Anal. Chim. Acta* **1990**, *235*, 163.

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

0022-3263/95/1960-8002$09.00/0

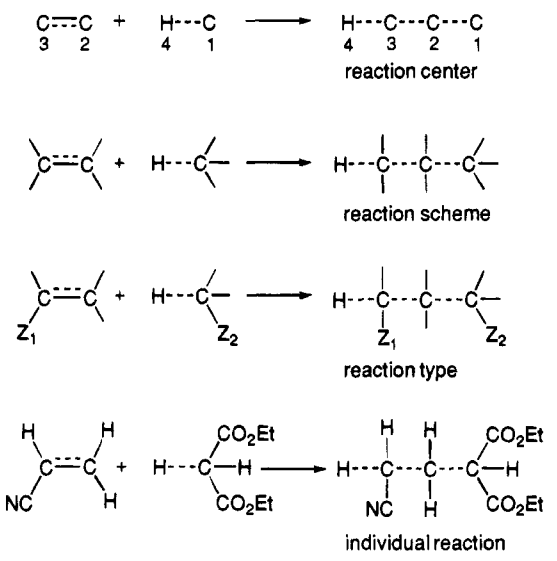


Figure 1. Relationships between the concepts of reaction center, reaction scheme, reaction type, and individual reaction. A reaction type like the Michael addition is an instantiation of a reaction scheme. The C=C and C−H bonds make up the reaction center (Z_1 and Z_2 are electron-withdrawing groups).

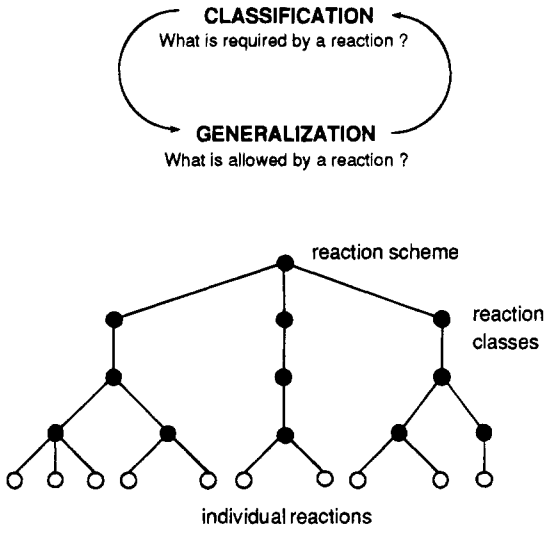


Figure 2. Basic procedure of HORACE: iterative classification and generalization. This leads to a grouping of the individual reactions into reaction classes. These reaction classes have the same reaction scheme, as HORACE processes reactions only after isolating the reaction schemes.

bonds directly involved in the bond and electron rearrangement of a reaction; a *reaction scheme* specifies how the bonds and electrons of a reaction center are shifted during a reaction. A *reaction type* is a set of reactions commonly grouped together by chemists because of similarities in the substituents at the reaction center or as a consequence of having the same reaction mechanism. Clearly, a reaction scheme can comprise several reaction types. The relationships of these concepts are indicated in Figure 1. A *reaction class* consists of reactions that have been grouped together by HORACE. What one wants to achieve is that the reaction classes of HORACE correspond to the reaction types defined by chemists.

However, HORACE can offer more information than that: It can build a reaction classification hierarchy for a given reaction dataset (see Figure 2). In order for two reactions to be grouped together into the same reaction class, they must belong to the same reaction scheme. Thus they must have the same reaction center. It is therefore natural that the analysis of reactions in HORACE starts at the reaction center. Two major questions have to be answered in the classification process. What features must be present in order for two reactions to be placed into the same class? Or, in other words, what is required for a certain reaction type? This question has to be solved by a *classification* scheme. Secondly, what features may or may not be present? Or, what is allowed by a certain reaction type? Such features should be combined in a summarization by *generalization*.

Thus, the classification of reactions into reaction classes requires both methods for classification and for generalization. In fact, HORACE consists of an iterative procedure which alternates between steps of classification and generalization. In each iteration a new level in the classification hierarchy is generated (Figure 2). The procedure terminates when no further grouping of reactions into fewer classes is found to be possible.

The decisions as to which reactions should be grouped together and which features can be generalized are made on the basis of the set of reactions shown to HORACE. Thus, HORACE is a *data-driven* system, and in this sense it is a genuine knowledge discovery system: The chemistry contained in the set of reactions determines the classifications and generalizations. This is quite different from the *model-driven* approach where a preconceived model would be imposed onto the classification such that, for example, all reactions that have an electron-withdrawing group on atom 1 of the reaction center are grouped into a single class. This is not done in HORACE!

In contrast, HORACE uses unsupervised machine-learning techniques. Specifically, it relies heavily on conceptual clustering methods.[12] The concepts, the features used to decide which reactions are to be clustered into classes, are 2-fold: physicochemical features and topological features.

Physicochemical Features. A common reaction mechanism is an important criterion shared by reactions that a chemist considers to be of the same type. Unfortunately, no direct information on the mechanism of a reaction is presently stored in commercially available reaction databases. Thus, the mechanism of a reaction has to be inferred in an indirect manner. The mechanism of a reaction is determined to a large extent by electronic factors such as charge distribution and inductive and resonance effects at the reaction center.

Presently, HORACE uses empirical methods for the quantification of the inductive and resonance effects at the atoms and bonds of the reaction center and those adjacent to it. Inductive effects are measured by residual σ-electronegativities, χ_σ, calculated by the PEOE method.[13,14]

Ground state resonance effects are derived from π-electronegativities, χ_π, calculated by the extension of the PEOE method to conjugated systems.[15] For each bond from a substituent to an atom of the reaction center, the differences, $\Delta\chi_\sigma$, and $\Delta\chi_\pi$, are calculated.

While $\Delta\chi_\sigma$ and $\Delta\chi_\pi$ measure effects in the nonreacting starting materials, additionally, estimates of the effects exerted on the molecules during reaction were considered necessary. To this effect, each bond of the reaction center is formally broken in a polar manner. This results in a positive and a negative charge, and the extent of stabilization of each of these charges through delocalization, D^+ and D^-, is estimated from weighting resonance structures.[15-17] As there are two polarities for heterolysis of a bond conceivable, for each atom of the reaction center both D^+ and D^- values are obtained (Figure 3). Table 1 gives the values of these electronic variables for

(10) Parlow A.; Weiske, C.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 400.
(11) ChemInform-RX has been produced since 1991 by Fachinformationszentrum Chemie, Berlin, Germany, from the information contained in the weekly abstracting service ChemInform, marketed by MDL Information Systems Inc., San Leandro, CA.
(12) Michaelis, R. S. *Int. J. Policy Anal. Inf. Sys.* **1980**, *4*, 219.
(13) Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219.
(14) Hutchings, M. G.; Gasteiger, J. *Tetrahedron Lett.* **1983**, *24*, 2541.
(15) (a) Gasteiger, J.; Saller, H. *Angew. Chem.* **1985**, *97*, 699. (b) *Angew. Chem., Int. Ed. Engl.* **1985**, *24*, 687.
(16) Gasteiger, J.; Saller, H.; Löw, P. *Anal. Chim. Acta* **1986**, *191*, 111.
(17) Fröhlich, A. Dissertation, Technische Universität München, Germany, **1993**.

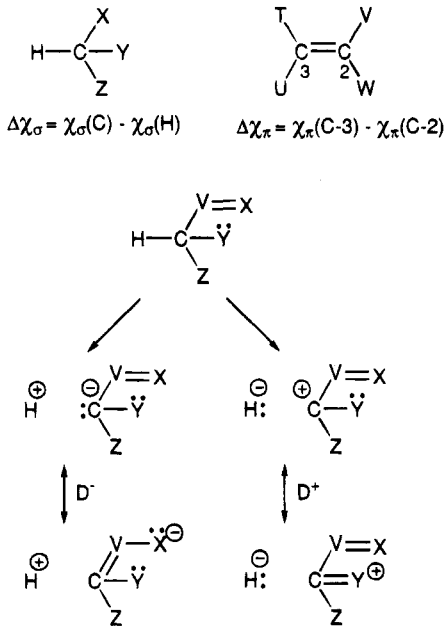
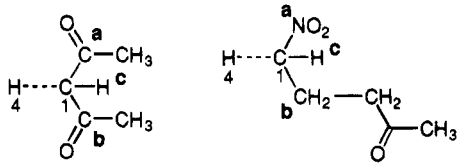


Figure 3. Physicochemical variables used in the HORACE classification.

Table 1. Comparison of Physicochemical Features for Two Educts



variable	bond/atom	educt 1	educt 2
D^+ (eV)	C-1[a]–H-4	9.12	11.72
D^- (eV)	C-1[a]–H-4	10.51	14.52
χ_π (eV)	a	6.45	9.06
	b	6.45	0.00
	c	0.00	0.00
χ_σ (eV)	a	10.10	13.12
	b	10.10	7.78
	c	7.43	7.52

[a] This is the atom that obtains the charge, indicated positive charge with D^+ and negative charge with D^-.

two molecules. As can be seen, the values quantitatively reflect the more intuitive feelings of an organic chemist.

Functional Groups (Topological features). The concept of functional groups is deeply rooted in the thinking of organic chemists. It should be realized that chemists have arrived at the notion and identification of functional groups through inductive reasoning by classification and generalization. Functional groups have been generalized, for example, into electron-donating and electron-withdrawing categories. However helpful this classification and generalization may be, it is also problematic. Clearly, most chemists will automatically classifiy a COOR group as electron withdrawing and an OR group as electron donating, but their classifications are only valid when, indeed, the COOR group is adjacent to an atom obtaining a negative charge, and the OR group is bonded to a center that obtains a positive charge (completely or partially) during the course of a reaction. In this case, this clear-cut *antagonistic* classification is valid. If, however, a radical reaction is considered, both substituents, the COOR and OR groups, exert a stabilization effect on an adjacent radical center (see Figure 4) and should therefore be grouped together into the *same* class of substituents. Thus, it is clear that the classification and generalization of such functional groups

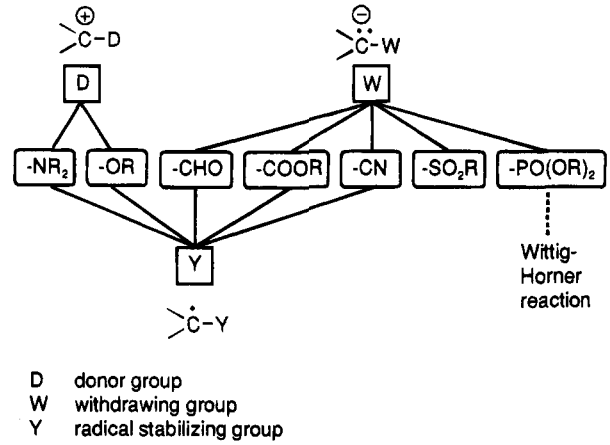


Figure 4. Classification of substituents dependent on the reaction type (see text).

should not be done without consideration of the reaction mechanism that is involved.

There is an additional problem: Not all supposedly equivalent electron-withdrawing/donating groups are interchangeable. For example, in many reaction types the phosphonate group, $PO(OR)_2$ exerts an electron-withdrawing influence in much the same way as a COOR group does. However, this should not lead to the conclusion that a $PO(OR)_2$ group can always be generalized together with a COOR substituent into an electron-withdrawing group. For example, in the Wittig–Horner reaction, a $PO(OR)_2$ group cannot be replaced by a COOR group. The conclusion that the $PO(OR)_2$ group can be generalized into an electron-withdrawing group is not allowed in the case of the Wittig–Horner reaction.

The answer to these problems can only be that there is not a fixed, static classification and generalization of functional groups. Rather, the extent of classification and generalization of functional groups has to be derived from the set of reactions under consideration. A data-driven approach will lead to a dynamic classification. Only by considering individual reactions can one come to the realization that sometimes two functional groups fall into same classes and sometimes into different classes. Likewise, sometimes a generalization of a functional group is allowed and sometimes not. This is another strength of a data-driven approach in contrast to a model-driven one which would inadvertently be constrained to work with a static classification and generalization of functional groups.

HORACE presently contains a set of 114 functional groups that are used as topological features in the classification and generalization process. Figure 5 shows an extract of this set of functional groups that are stored in an external file for easy modification and extension.

Strategies of the HORACE System. Having established the criteria that should be considered in finding the essential characteristics of a reaction type, the question is now: how, and in what sequence, does one take these criteria into account? Or, in other words, should experiments be performed by first considering the functional groups and then the physicochemical features, or should it be the other way around?[6] The best and most clear-cut results were obtained by first sorting the individual reactions into separate groups on the basis of the physicochemical features, since this causes those reaction instances which are mechanistically related but happen to have different complements of functional groups around the reaction center to be grouped together (see Table 1). Each such group is then separately submitted to the process of classification and generalization by considering the functional groups around the reaction center. As this second step involves an iterative process, a hierarchical classification into reaction types of increasing generalization is obtained.

Let us first briefly address the grouping of reactions into clusters based on *physicochemical effects*. A detailed description of the classification method will be presented later in this

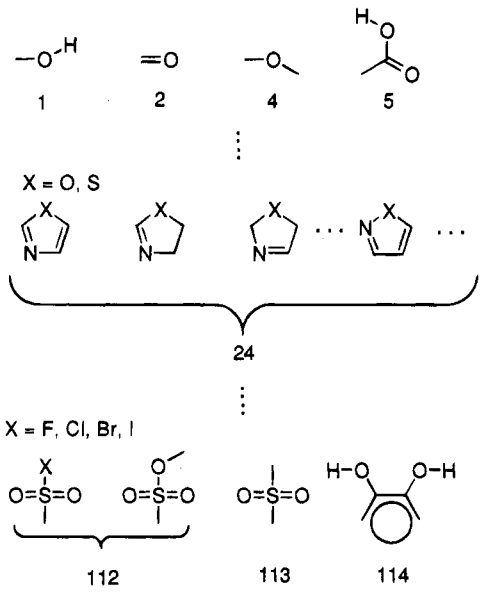


Figure 5. Some of the 114 functional groups used as topological features in HORACE.

paper using experimental results to illustrate key aspects. Summation over the differences in σ- and π-electronegativities, $\Delta\chi_\sigma$ and $\Delta\chi_\pi$, allows one to determine the overall electronic effect exerted by a substituent onto an atom of the reaction center. These values are then used to place a substituent into one of the following categories: strongly or mildly electron withdrawing (W or w), neutral (n), or mildly or strongly electron donating (d or D). These values are then used to calculate an "atom distance" for two reactions by comparing the corresponding values of the substituents of the two reactions in question. Thus, for example, if the substituent of one reaction is W and the other is w, a value of 0.5 is chosen for the "atom distance" of the influence of the two substituents on the respective reaction. Summation over all distances of the substituents to all the atoms of the reaction center gives a value for the total atom distance of two reactions. If the atom distance exceeds a selectable threshold value, T, the atom distance is set to an even higher value.

Similar to this "atom distance", a "bond distance" is also calculated for each reaction. The value of the bond distance is obtained by consideration of each bond of the reaction center. For all such bonds, the amount of potential stabilization by delocalization of the charges resulting from heterolysis is calculated. It is possible to influence the value of the bond distance by varying a scaling factor, F.

The distance between any two reactions is then calculated by summing atom and bond distances. If this overall distance is smaller than the threshold value T, then the two reactions compared are put into the same group. Changing the values of T and F allows one to influence the number of classes that are obtained from the grouping by the physicochemical features.

Each such group is then further processed by classification and generalization based on the *topological features*. As with a physicochemically based classification, a detailed explanation of the topological classification approach will be presented along with illustrative results later in this paper. In a nutshell, the reactions of each class obtained from consideration of the physicochemical features are sorted according to the number of functional groups proximal to the reaction center. Only those functional goups that are at most one bond length from the reaction center are considered. The reaction with the fewest number of functional groups is taken as the representative, A, of this group of reactions. Any other reaction, B, is then compared with this representative A, and a closeness between these two reactions is calculated according to eq 1:

$$\text{closeness(A,B)} = \frac{|f_A \cap f_B|}{|f_A|} \quad (1)$$

where f_A and f_B are the sets of the functional groups at the reaction sites of reactions A and B, respectively, $|f_A|$ is the cardinality of set f_A, $f_A \cap f_B$ is the intersection of sets f_A and f_B. If the closeness is larger than a certain threshold value, the two reactions, A and B, are put into the same class.

In order to specify which kind of functional groups may be present at a certain position, a notation that makes it possible to represent a collection or set of substituents is necessary. HORACE bases this generalized representation of functional groups adjacent to the reaction center on the identity of the α-atoms of these groups. A hierarchy of atom types (Figure 6) has been built based on common chemical reactions and is used for the specification of the generalized α-atoms of the substituents. This hierarchy and identification of atom types is kept in a separate file for easy modification. More details and insights into the methods used by HORACE for reaction classification will be given in the section covering the results obtained on Michael additions.

The Dataset. The construction of the carbon skeleton is an essential task in the synthesis of complex organic molecules. In choosing a dataset for extracting knowledge on organic reactions, we have elected a reaction scheme that incorporates the formation of a carbon−carbon bond. In order to develop and verify the classifications obtained by HORACE, a reaction scheme was chosen that encompasses a wide variety of reaction types. Consequently, we wanted to make sure that the different reaction types are indeed perceived and the individual reactions correctly assigned to the various reaction types.

The reaction scheme chosen is shown in Figure 1. It involves the addition of a C−H bond to a C=C double bond. Clearly, such a scheme lies at the foundation of a Michael addition. However, Friedel−Crafts alkylations of aromatic compounds by alkenes also fall into such a scheme, as do additions to alkenes of carbon radicals formed by breaking a C−H bond. Thus, we have a wide range of reaction types proceeding under catalysis by bases, Brønsted or Lewis acids, and radical initiators.

The set of reactions used in this study was taken from the ChemInform RX reaction database.[10,11] A reaction center search with the scheme shown in Figure 1 was initiated with REACCS[2] in volume 1991 of the ChemInform RX database; 120 reactions resulted from this search and formed the basis of the present study.

Results and Discussion

The results of the classification of 120 reactions, based only on physicochemical features, are summarized in Table 2. This table shows the number of reactions contained in each class. The semantic description of the reaction types was not obtained by HORACE but assigned by human inspection. It can be seen that the reactions belonging to different reaction types were clearly separated. On the other hand, in some cases, reactions belonging to the same reaction type were assigned to more than one class, particularly with the Michael additions. This is the case with reaction types covering a broad range of structural variety encompassing special reactions.

Although a dataset of 120 reactions is fairly small as compared to the number of known reactions, it nevertheless comprises a rich source of information on chemical reactions. A variety of different types of knowledge can be extracted from such a source. Clearly, it is beyond the scope of this publication to discuss every aspect and all details of chemical knowledge acquired in a HORACE run. Rather, we will concentrate on the major types of knowledge. In particular, emphasis will be placed on those kinds of questions that can be answered by the presently available reaction retrieval systems only indi-

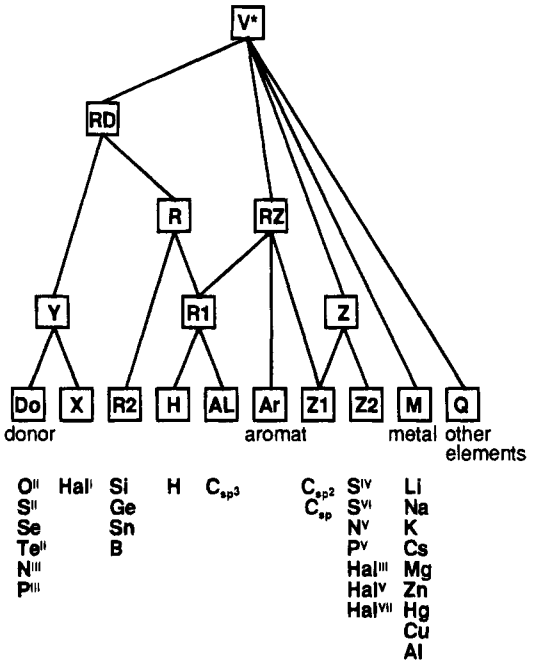


Figure 6. Hierarchy of the generalization of atom types for the α-atoms of the substituents at the reaction center.

Table 2. Results of Classification of 120 Reactions Based on Physicochemical Features[a]

class	no. of reactions	reaction type	class	no. of reactions	reaction type
1	3	P	16	1	$
2	32	M	17	4	FC
3	1	@	18	2	FC
4	1	P	19	1	R
5	2	P	20	1	M
6	1	#	21	2	M
7	3	H	22	2	M
8	1	C	23	1	P
9	1	$	24	3	FC
10	1	$	25	3	M
11	9	FC	26	2	R
12	1	M	27	1	$
13	32	M	28	1	M
14	1	R	29	1	M
15	5	N	30	1	R

[a] Reaction type: C, condensation reaction; FC, Friedel–Crafts alkylation by olefins; H, hydride abstraction reaction; M, Michael addition; N, Nazarov reaction; P, photochemical reaction; R, free radical reaction; @, reaction with wrongly assigned reaction center, #, reaction with special mechanism; $, "reaction" consisting of two reactions with different reaction conditions.

rectly or with difficulty. Specifically, the following questions on chemical reactions seem to be important and involve in one way or another a certain level of abstraction, a characteristic of inductive learning. Which reaction types have a broad scope and utility in chemical synthesis? Which functional groups are essential for a reaction type? Which functional groups, though appearing proximal to the reaction site, are not essential? Which reactions are novel or unusual?

Depending on which of the above questions is of major interest to the chemist, different strategies in analyzing the output of a HORACE run can be recommended. When one searches for reactions that have a broad scope, those reaction classes with many members or with highly generalized atom types at the reaction center should be investigated. If, however, rather unusual reactions are to be located, those reaction classes with only one or a few instances should be investigated.

The following discussion will begin with the major reaction types of the given reaction scheme, the addition of a C–H bond to a C=C double bond, such as Michael addition and Friedel–Crafts alkylation. Next, some more unusual or specialized reaction types will be discussed.

Clearly, the compression of 120 reactions into 30 reaction classes is not a very compact one. However, we have found this level of compression appropriate both for discovering the major reaction types as well as for easily detecting special reactions.

Michael Addition. The two largest classes, classes 2 and 13, from among the 30 classes obtained from the automatic classification of 120 reactions contain 32 individual reactions each. Both classes are classified by a chemist as Michael additions; the difference between the two classes is that in class 13 all reactions have an aromatic ring in conjugation to the reacting double bond, whereas the reactions in class 2 do not have this feature. Thus, the division of Michael additions into these two classes immediately provides us with an important knowledge about the Michael addition: It can proceed with or without an aromatic ring in conjugation to the C=C double bond.

Let us now look at one of these classes, class 2, in more detail. Figure 7 shows the reaction classification hierarchy produced for this class based on the functional groups around the reaction center. In order to make HORACE's approach to classification more concrete to the reader, we will delve into the details of the creation of subclass 2.3 of Figure 7. We will simplify the exposition by examining the classification of the 12 reactions (Figure 8) that comprise subcluster 2.3 in isolation. As already mentioned, HORACE begins by classifying reactions on the basis of *physicochemical features*. In particular, the clustering is based on parameters describing the electronic effects operative at the reaction center. Only parameters from the reactants are used in the classification. The current set of parameters that are considered are χ_σ and χ_π at atoms in the reaction center and those atoms α to the reaction center and the delocalization parameters (D^+ and D^-) for bonds in the reaction center. In the following discussion reference will be made to individual atoms in the reaction center which are numbered according to the scheme shown in Figure 1. Table 3 lists the delocalization values for the 12 reactions under consideration. Note that although D^+ and D^- are bond parameters, the values are listed with respect to educt atoms in the reaction center in order to indicate direction within the bond. Thus the D^+ value of 9.12 eV listed for atom 1 in reaction 23 005 indicates that the carbon–hydrogen in the first reactant has this positive delocalization value in the direction of the carbon atom. Not surprisingly, the values for atom 4, the hydrogen atom, are all zero.

In addition to delocalization parameters, the influence of substituents on the reaction center is described in terms of the σ- and π-electronegativities of the atoms α to the reaction center. This is shown in Table 4. We see that atom 1 of each reaction has three substituents. The σ- and π-electronegativity values, χ_σ and χ_π, for each substituent are combined and quantized to derive a value that ranges from strongly electron withdrawing (W) to strongly electron donating (D) with respect to the corresponding atom in the reaction center. The possible values are W, w, n, d, and D which are strongly withdrawing, withdrawing, neutral, donating, and strongly donating, respectively. Since atom 4, the hydrogen atom

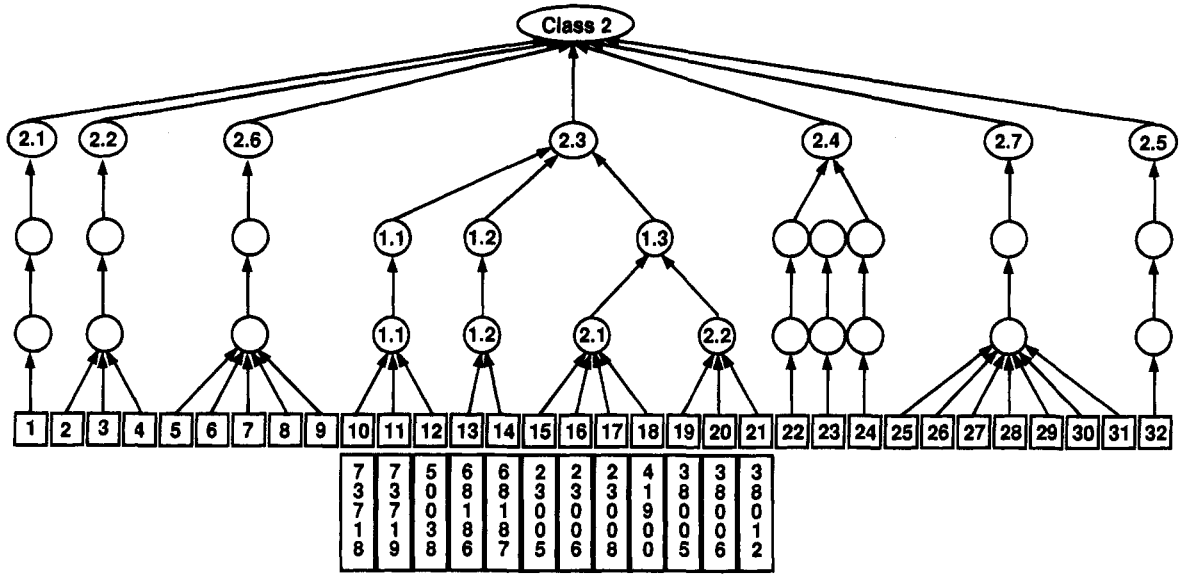


Figure 7. Hierarchical classification scheme of class 2, comprising those Michael additions that do not have an aromatic ring in conjugation to the reacting double bond. Note that the subsubclasses 1.1, 1.2, etc., are generated for each subclass separately.

Table 3. Delocalization Parameters D^+ and D^- (eV)

C=C + H–C → H–C–C–C

3 2 4 1 4 3 2 1

	C-1		C-2		C-3		H-4	
reaction no.	D^+	D^-	D^+	D^-	D^+	D^-	D^+	D^-
23 005	9.12	10.5	0.0	0.0	8.18	5.90	0.0	0.0
23 006	9.12	10.5	0.0	0.0	8.19	5.90	0.0	0.0
23 008	14.1	15.4	0.0	0.0	8.18	5.90	0.0	0.0
38 005	13.8	12.3	0.0	0.0	7.68	6.25	0.0	0.0
38 006	13.8	12.3	0.0	0.0	8.19	5.90	0.0	0.0
38 012	13.8	12.3	9.55	0.0	7.70	6.42	0.0	0.0
41 900	11.4	12.2	9.40	0.0	9.40	0.0	0.0	0.0
50 038	6.68	9.14	0.0	0.0	12.7	9.67	0.0	0.0
68 186	16.6	12.2	0.0	0.0	7.69	6.25	0.0	0.0
68 187	16.6	12.2	9.58	0.0	8.20	5.90	0.0	0.0
73 718	11.7	14.5	0.0	0.0	7.69	6.25	0.0	0.0
71 719	11.7	14.5	0.0	0.0	7.69	6.25	0.0	0.0

in the reaction center, has no substituents, it is not included in this table. A quick glance at the quantized values in this table reveals that these reactions are quite similar with respect to these parameters.

HORACE uses the values in Tables 3 and 4 to derive a distance measure between all pairs of reactions. This is stored in the distance matrix shown in Table 5. The computation of distance between reactions is accomplished by first calculating the reaction center atom distance. The distance between two corresponding atoms in the reaction center is the sum of the differences between the sets of quantized substituent electronic effects shown in Table 4.

A heuristic function is used to derive the quantized distance between corresponding substituents. A distance of 0 is returned if the substituents possess the same descriptor. If one of the pair is strongly withdrawing and the other is withdrawing, a value of 0.5 is returned. If one of the pair is strongly donating and the other is donating, a value of 0.5 is returned. Otherwise, a value of 1 is returned indicating that they are dissimilar. This value is referred to as the reaction center atom distance. The distances for all corresponding pairs of substituents for all of the atoms in the reaction center are summed, and this value is normalized by dividing by the number of substituents. The resulting quantity is then squared producing the *reaction center atom distance* between the two reactions. If this value exceeds a user supplied threshold, the value is changed to a "huge distance", a large constant that will preclude the possibility of these reactions being grouped in the same class. This is to ensure that there is not a great amount of variance in similar mean distances. If two corresponding atoms in the reaction center have different numbers of substituents, then a distance value of "huge distance" will be returned.

Next, the *reaction center bond distance* is computed. A second heuristic function is employed to derive this value. The heuristic calculates the difference between delocalization parameters, D^+ and D^-, for corresponding bonds in the two reaction centers. The maximum distance found is then scaled by a delocalization coefficient, F, to produce the *reaction center bond distance*. The scaling is done to modify the influence exerted when this value is combined with the *reaction center atom distance* to form the *reaction center distance*. The resulting *reaction center distance* values computed by HORACE are shown in Table 5. A delocalization coefficient of $F = 1.0$ was used.

After the distance matrix $d(i,j)$ (see Table 5) has been constructed, each distance is compared with the threshold T. If $d(i,j) \leq T$, then reactions i and j are placed in the same cluster. Finally, the closure of all clusters is taken, i.e., clusters with nonempty intersections are combined so that all remaining clusters have empty intersections. For example, using a threshold $T = 0.4$, it is clear that all 12 reactions will be placed in the same cluster since all reactions are within distance 0.4 of reaction 23 005 as shown in the column labeled 23 005 in Table 5.

The classification of the 12 reactions based on physicochemical features is then refined by considering *topological features*. The reactions are examined for their complements of functional groups proximal to the reaction center. Initially, all functional groups are identified with respect to the reaction center atom they are associated with as shown for reaction 23 005 in Table 6.

In this table, pfg (proximal functional group) indicates that the functional group is α to the associated reaction

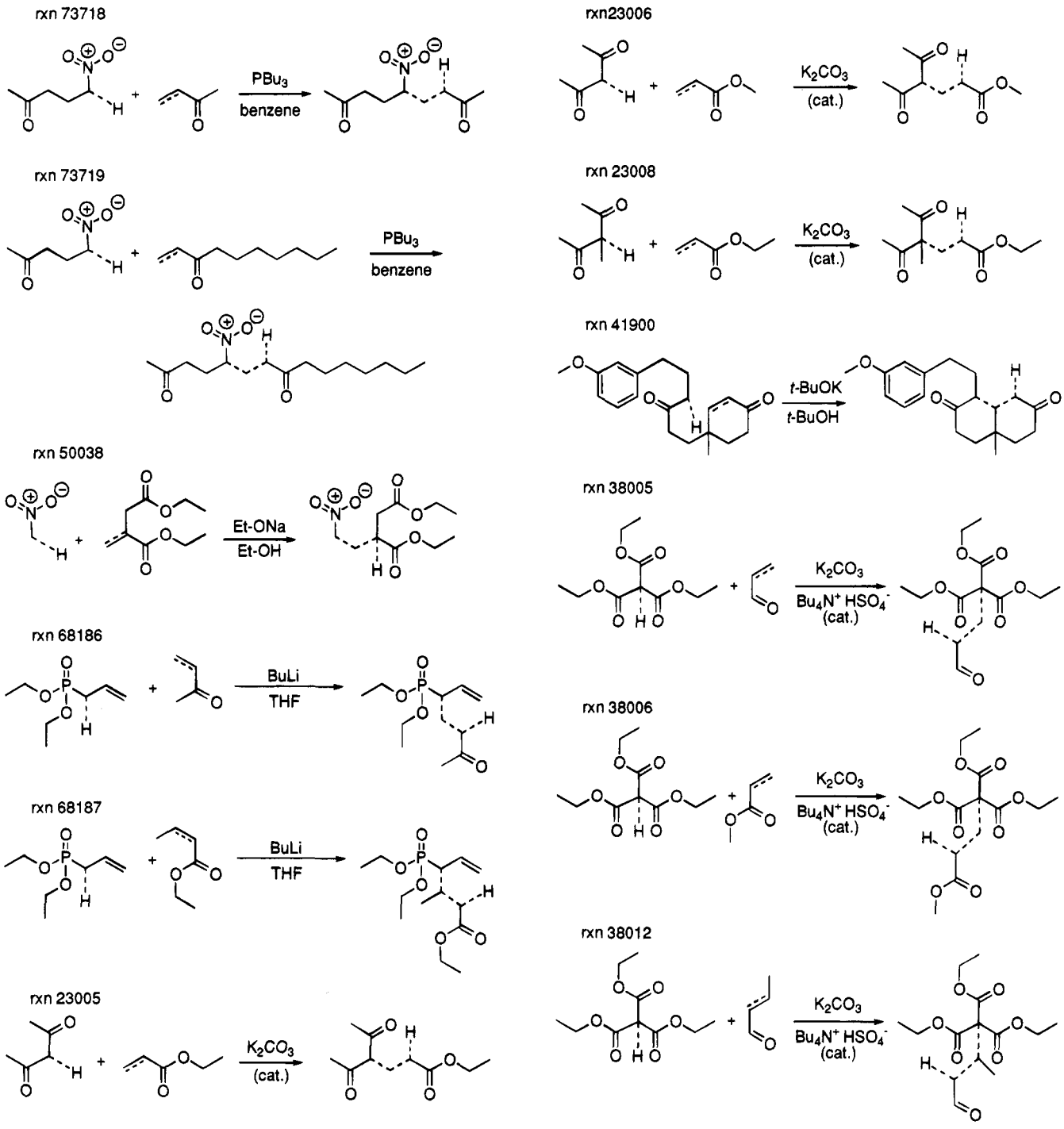


Figure 8. Twelve individual reactions of subclass 2.3.

center atom. For example, the second line of Table 6 denotes a carbonyl α to atom 1 of the reaction center. If the functional group actually contains the corresponding reaction center atom, then this is denoted by the label ppfg. Thus the ketone mentioned in line 6 of Table 6 contains atom 1 of the reaction center.

After perceiving a reaction's complement of proximal functional groups, those functional groups that are wholly contained in some other functional group are weeded out. For example, the ether listed in Table 6 is discarded since it is contained in the carboxylic ester. Likewise, the allylic ether is discarded. The remaining 10 functional groups are used by HORACE in its topological classification.

In order to perform a topological classification, it is necessary to determine the number of classes. HORACE does this by first identifying reactions that will form cluster cores. The 12 reactions are sorted according to the number of functional groups that were perceived. The reactions are listed in sorted order in Table 7. The first column lists the reaction names, and the second column indicates the number of perceived functional groups. By convention, the reaction with the fewest number of functional groups is taken as the first cluster core. In this case, reaction 73 718 is the first cluster core since it has only six perceived functional groups. The remaining reactions are in turn compared with the set of cluster cores. If a reaction does not match any cluster core, it in turn becomes a cluster core. The degree of required match is selectable. A match of 67% or better is the

Table 4. Influence of Substituents onto the Reaction Center: σ- and π-Electronegativities (χ_σ and χ_π, in eV)

	substituents on C-1			substituents on C-2			substituents on C-3		
reaction no.	χ_σ	χ_π	descriptor	χ_σ	χ_π	descriptor	χ_σ	χ_π	descriptor
23 005	10.1	6.4	W	8.8	5.6	w	11.6	7.1	W
	10.1	6.4	W	7.5	0.0	n	7.9	5.7	w
	7.4	0.0	n	7.5	0.0	n	7.6	0.0	n
23 006	10.1	6.4	W	8.8	5.6	w	11.6	7.1	W
	10.1	6.4	W	7.5	0.0	n	7.9	5.7	w
	7.4	0.0	n	7.5	0.0	n	7.6	0.0	n
23 008	10.1	6.5	W	8.8	5.6	w	11.6	7.1	W
	10.1	6.5	W	7.5	0.0	n	7.9	5.7	w
	7.5	0.0	n	7.5	0.0	n	7.6	0.0	n
38 005	11.6	7.1	W	8.5	5.5	w	10.2	6.4	W
	11.6	7.1	W	7.5	0.0	n	7.9	5.7	w
	11.6	7.1	W	7.5	0.0	n	7.6	0.0	n
38 006	11.6	7.1	W	8.8	5.6	w	11.6	7.1	W
	11.6	7.1	W	7.5	0.0	n	7.9	5.7	w
	11.6	7.1	W	7.5	0.0	n	7.6	0.0	n
38 012	11.6	7.1	W	8.6	5.5	w	10.2	6.4	W
	11.6	7.1	W	7.6	0.0	n	8.0	5.8	w
	11.6	7.1	W	7.5	0.0	d	7.6	0.0	n
41 900	10.1	6.4	W	8.6	5.5	w	10.3	6.5	W
	7.6	0.0	n	7.9	0.0	n	8.1	5.8	w
	7.4	0.0	n	7.5	0.0	d	7.6	0.0	n
50 038	13.1	9.0	W	9.0	5.7	W	11.6	7.1	W
	7.5	0.0	n	7.5	0.0	n	8.0	5.7	W
	7.5	0.0	n	7.5	0.0	n	8.6	0.0	w
68 187	11.7	7.5	W	8.6	5.5	w	10.3	6.5	W
	8.1	5.2	w	7.5	0.0	n	7.9	5.7	w
	7.4	0.0	n	7.5	0.0	n	7.6	0.0	n
68 187	11.7	7.5	W	8.6	5.6	w	11.6	7.1	W
	8.1	5.2	w	7.6	0.0	n	8.0	5.8	w
	7.4	0.0	n	7.5	0.0	d	7.6	0.0	n
73 718	13.1	9.1	W	8.6	5.5	w	10.2	6.5	W
	7.8	0.0	w	7.5	0.0	n	7.9	5.7	w
	7.5	0.0	n	7.5	0.0	n	7.6	0.0	n
71 719	13.1	9.1	W	8.6	5.5	w	10.3	6.5	W
	7.8	0.0	w	7.5	0.0	n	7.9	5.7	w
	7.5	0.0	n	7.5	0.0	n	7.6	0.0	n

Table 5. Distance Matrix

reaction no.	23 005	23 006	23 008	38 005	38 006	38 012	41 900	50 038	68 186	68 187	73 718	73 719
23 005	0.0											
23 006	0.0	0.0										
23 008	0.1	0.1	0.0									
38 005	0.2	0.2	0.1	0.0								
38 006	0.2	0.2	0.1	0.0	0.0							
38 012	0.4	0.4	0.4	0.3	0.3	0.0						
41 900	0.4	0.4	0.5	3.6	3.6	3.3	0.0					
50 038	0.4	0.4	0.5	3.7	3.7	7.1	3.8	0.0				
68 186	0.1	0.1	0.1	0.2	0.2	0.5	0.5	0.5	0.0			
68 187	0.4	0.4	0.3	0.5	0.5	0.2	0.2	4.0	0.3	0.0		
73 718	0.1	0.1	0.1	0.2	0.2	0.5	0.4	0.5	0.1	0.4	0.0	
73 719	0.1	0.1	0.1	0.2	0.2	0.5	0.4	0.5	0.1	0.4	0.0	0.0

Table 6. Functional Groups Found at the Atoms of the Reaction Center

atom	label	functional group
C-1	pfg	carbonyl
	pfg	carbonyl with α-proton
	ppfg	ketone
	ppfg	carbonyl with α-proton
	ppfg	β-dicarbonyl with α-proton
C-3	pfg	carbonyl
	pfg	ether
	pfg	carboxylic ester
	ppfg	carbonyl with α-proton
	ppfg	α,β-unsaturated carbonyl
	ppfg	allylic ether
	ppfg	simple carboxylic ester

Table 7. Topological Classification Level 1

reaction no.	no. of functional groups	matching functional groups		classification	
		cluster 1	cluster 2	cluster	subcluster
73 718	6	6	4	1	1.1
73 719	6	6	4	1	1.1
68 186	7	5	4	1	1.2
68 187	7	5	4	1	1.2
50 038	8	5	3	1	1.1
23 005	10	5	8	2	2.1
23 006	10	5	8	2	2.1
23 008	10	5	8	2	2.1
38 005	11	4	11	2	2.2
38 012	11	4	11	2	2.2
38 006	12	5	11	2	2.2
41 900	12	5	7	2	2.1

standard threshold value. Thus, we see in Table 7 that reaction 38 005 is the first reaction to fail to meet the required degree of match with reaction 73 718 because the match is $^4/_6 = 66.67\%$ (<67%). Consequently, reaction 38 005 is designated a new cluster core. All subsequent reactions adequately match either reaction 73 718 or 38 005, so no new cluster cores are added.

The reactions are then compared with the set of cluster cores and classified according to which cluster core they

Table 8. Topological Classification Level 2

cluster from previous classification	functional groups	matching functional groups	classification	
			cluster	subcluster
subcluster 1.1	7	7	1	1.1
1.2	8	6	1	1.2
2.1	10	6	1	1.3
2.2	13	6	1	1.3

Table 9. Topological Classification Level 3

cluster from previous classification	functional groups	matching functional groups	classification	
			cluster	subcluster
subcluster 1.1	7	7	1	1.0
1.2	8	6	1	1.0
1.3	11	6	1	1.0

most closely match. For example, reactions 23 005, 23 006, and 23 008 have more functional groups in common with the second cluster core (reaction 38 005) than with the first one (reaction 73 718) and are therefore put into cluster 2. Although it is possible for a reaction to match two or more cluster cores with an identical degree of closeness, that did not occur in the case of the reactions under discussion. As shown in Table 7, each reaction is placed uniquely in one cluster.

Each cluster is then analyzed separately for possible subclusters. The basic idea is that each reaction is associated with the reaction in its cluster that it most closely matches. Subclusters are then formed by taking the transitive closure of all maximally matching pairs. The subclusters among the 12 reactions under consideration are shown in the last column of Table 7. Thus, we find that the initial cluster core, reaction 73 718, has been placed in subcluster 1 of cluster 1, denoted by 1.1, while the second cluster core, reaction 38 005, has been placed in subcluster 2 of cluster 2, denoted by 2.2.

During the next iteration of classification, subcluster 1.1 is selected as the initial cluster core since it has the smallest complement of functional groups (seven), as shown in Table 8. Since the remaining subclusters match the new cluster core, no additional cluster cores are added. Consequently, all subclusters from the previous iteration are grouped in the new cluster 1. The subcluster formation phase results in the creation of three subclusters. Subclusters 2.1 and 2.2 from the previous classification iteration are grouped together.

During the final level of classification, shown in Table 9, the three subclusters formed in the previous iteration are combined to produce a single cluster. This cluster is not partitioned into subclusters. This is denoted by "1.0" in the subcluster column of Table 9. By convention, HORACE does not produce subclusters during the final topological classification iteration.

In summary, the physicochemical classification method grouped 12 reactions into one class. This is consistent with the fact that all 12 reactions are Michael additions. The topological classification of this small dataset led to the building of a hierarchy which is in agreement with the subhierarchy generated for subclass 2.3 in Figure 7, allowing easy analysis of the structural varieties of these reaction instances.

We stress once again that HORACE not only generates reaction classification hierarchies but produces generalized representations for reaction (sub)classes as well. As an example, Figure 9 shows the generalized representations of subclasses 2.1, 2.2, and 2.3.

In order to assist the reader in further developing an understanding of the generalized atom types, the symbols

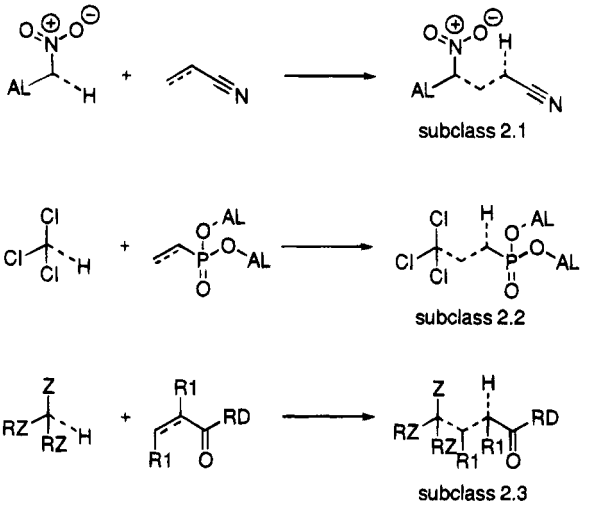


Figure 9. Generalized representations of subclasses 2.1, 2.2, and 2.3. The bonds in the reaction center are indicated by dotted lines.

in the generalized representation of subclass 2.3 are explained in detail. The reaction instances in this subclass all have a carbonyl group in conjugation with the reacting double bond, but this carbonyl group can either be a ketone, bearing a C_{sp^3} atom at the carbonyl group, or an ester, bearing an oxygen atom at the carbonyl group. The generalized representation of these two atoms, C_{sp^3} and O^{II}, is the symbol RD (alkyl or donor group) (see Figure 6). At either end of the double bond, either a hydrogen atom or an alkyl group might be bonded, the generalized symbol for these two atom types is R1.

Now let us turn our attention to the situation at the reacting C−H bond. First, at least one electron-withdrawing group, Z, has to be present; in the examples in this dataset, an NO_2, CO_2R, COR, or $PO(OR)_2$ was found. The second α-atom can either be an H or a C_{sp^3} atom or the C_{sp^2} atom of an alkene or a second carbonyl group. These different atoms meet in the atom hierarchy of Figure 6 at the atom type RZ. Also, there are three reactions that carry three ester groups in α-position to the C−H bond. Thus, a third C_{sp^2} atom has to be generalized together with an H or C_{sp^2} atom to form the second RZ symbol.

The generalized representation of a (sub)class reveals the common features of the reaction instances in that (sub)class. For example, from Figure 9 we immediately know that a C−H bond can be activated to undergo a Michael addition not only by substituents that can stabilize a carbanion through delocalization (subclasses 2.1 and 2.3) but also by the combined inductive effect of three chlorine atoms (subclass 2.2). This is the merit of a data-driven approach that can discover such latent knowledge. In the present case it results from the use of a combination of σ- and π-electronegativities as well as delocalization variables in phase I, the physicochemical classification, of a HORACE run.

To summarize, a wide variety of Michael additions was correctly perceived to show common characteristics that resulted in their placement in the same reaction class (class 2). That these reactions have much in common is underlined by the fact that 30 of the 32 reactions proceed under base catalysis and only two reactions (in subclass

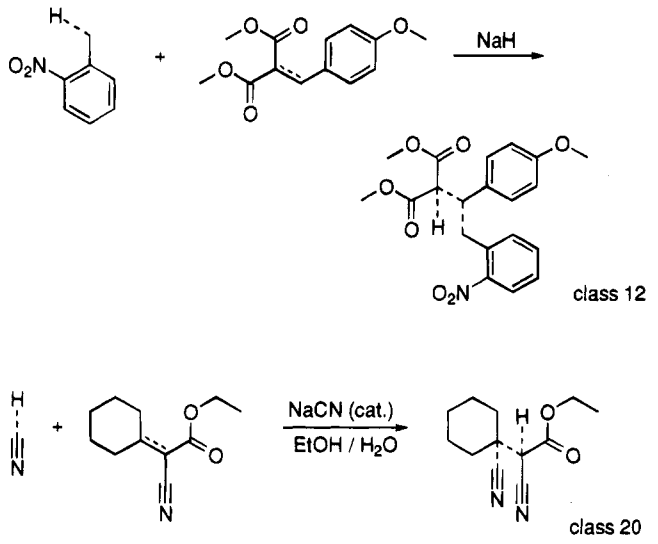


Figure 10. Two special types of Michael additions found in classes with one entry only.

2.3) are catalyzed by Bu_3P.[18] It should be emphasized, however, that no information on the catalysts was used in the classification. The common characteristics—including the implicit perception that these reactions are to be catalyzed in the same manner—were only derived from an analysis of the physicochemical effects and functional groups around the reaction site. Analysis of the seven subclasses of this group of reactions showed the variety of structural features that can occur in a Michael addition.

A similar analysis was made with class 13, comprising 32 reactions that are all Michael additions, differing from those of class 2 by having an aromatic ring in conjugation to the reacting double bond. This aromatic ring changes the electronic effects at the double bond to an extent that warranted the grouping of these reactions into a separate class. The knowledge that can be derived from an analysis of the subclasses of class 13 shows many similar traits with those mentioned in the preceding discussion. We therefore refrain from giving details on class 13.

The discussion has heretofore focused on the large bulk of instances of Michael additions. Now, we address the question of whether we can also find reactions that are outside the usual scope of Michael additions but might still be included in this reaction type. As was stated in the introductory remarks of Results and Discussion, novel or unusual types of reactions can be found in those classes having only one or a few reaction examples.

Consider classes 12 and 20, each comprising only one reaction as shown in Figure 10. The single reaction of class 12 involves the addition of a benzylic C–H bond without any other α-substituents.[19] Such bonds are usually not prone to be part of a Michael addition. However, in the specific case shown here, this bond is activated by a nitro group in the ortho-position that is able to stabilize the incipient carbanion. This potential is perceived by the routine used for calculating the stabilization of charges through delocalization, leading to delocalization variables, D^-, for the benzyl anion of 6.92 eV and for the *o*-nitrobenzyl anion of 14.57 eV. Although the nitro group is three bonds away from the

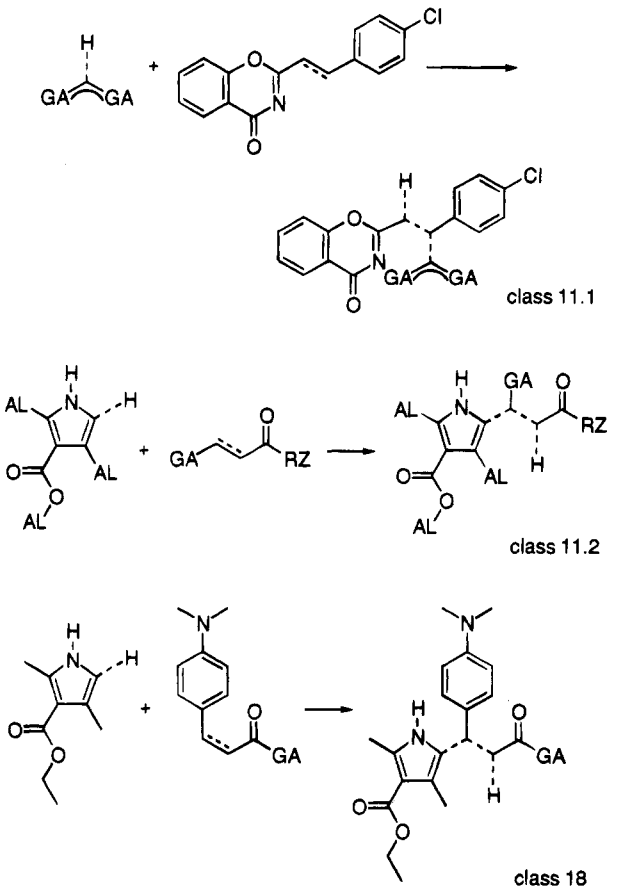


Figure 11. Generalized representations of subclasses 11.1 and 11.2 and class 18 of Friedel–Crafts alkylations found by HORACE (GA = generalized aromatic ring: aromatic ring with different substituents).

reaction center, its influence is perceived. Once again, this emphasizes the fact that a search limited to the functional groups directly bonded to the reaction center is not sufficient. Rather, as is done in HORACE, the physicochemical features of the reacting bonds must also be considered in the classification process.

In the reaction of class 20, the carbanion is formed at a C_{sp} center.[20] This is quite different from all the other Michael additions met so far that had the carbanion formed at a C_{sp^3} center. The fact that a cyanide ion can also undergo a Michael addition thus clearly extends the knowledge so far derived for Michael additions. It is therefore quite appropriate that this reaction was put in a class of its own, facilitating the discovery of this extra knowledge.

Both reactions of Figure 10 that had been put in classes of their own extended our knowledge on Michael additions. This confirms the strategy of searching for reactions that extend the scope of a reaction type in classes with one or a few entries only.

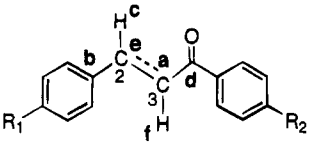
Friedel–Crafts Alkylation. The next largest group of reactions that fall into the scheme treated in this paper are Friedel–Crafts alkylations of aromatic compounds by alkenes. In fact, four classes of Friedel–Crafts alkylations, comprising 18 reactions altogether, were perceived and separated from other reactions by

(18) Janowitz, A.; Vavrecka, M.; Hesse, M. *Helv. Chim. Acta* **1991**, *74*, 1352.

(19) Floyd, D. M.; Moquin, R. V.; Atwal, K. S.; Ahmed, S. Z.; Spergel, S. H.; Gougoutas, J. Z.; Malley, M. F. *J. Org. Chem.* **1990**, *55*, 5572.

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

Table 10. Comparison of Physicochemical Features for Four Reaction Educts



	variable	bond/atom	reaction 23 484	23 487	23 485	23 488
R_1			H	H	NMe_2	NMe_2
R_2			H	NO_2	H	NO_2
substituents on C-2	D^+ (eV)		17.09	17.08	45.82	45.79
	D^- (eV)		10.55	10.57	29.30	29.34
		a	5.54	5.54	5.29	5.29
	χ_π (eV)	b	5.49	5.49	5.27	5.27
		c	0.00	0.00	0.00	0.00
		a	8.67	8.67	8.67	8.67
	χ_σ (eV)	b	8.57	8.57	8.58	8.58
		c	7.56	7.56	7.56	7.56
substituents on C-3	D^+ (eV)		8.21	8.12	8.51	8.42
	D^- (eV)		5.80	5.89	5.49	5.59
		d	6.59	6.59	6.57	6.58
	χ_π (eV)	e	5.83	5.84	5.80	5.81
		f	0.00	0.00	0.00	0.00
		d	10.59	10.59	10.59	10.59
	χ_σ (eV)	e	8.33	8.33	8.33	8.33
		f	7.58	7.58	7.58	7.58

HORACE. Two of these classes (classes 11 and 18) are discussed in more detail. Figure 11 shows the generalized representations of subclasses 11.1 and 11.2 and class 18.

The alkylations of aromatic rings by a double bond exhibit quite a structural variety, both at the aromatic system (benzenoid, heterocycles) and at the double bond. Because of this wide structural variety, these reactions have not been collected into a single class—at least not with the setting of the parameters of the threshold, T, and the scaling factor, F, chosen here. The largest class, class 11, consists of nine reactions. However, because of the structural differences of both educts in these reactions, they were further divided into two subclasses. Subclass 11.1 comprises two reactions that involve alkylations of substituted benzenes by alkenes, while the seven reactions in subclass 11.2 are the alkylations of substituted pyrroles (see the generalized representations of these two subclasses in Figure 11). Perhaps it would have been desirable to combine the two reactions of class 18 with the seven reactions of subclass 11.2 because of their close structural similarity. The reason why the two reactions of class 18 were kept apart from those of subclass 11.2 lies in the *p*-dimethylamino substituent of the phenyl ring. This substituent in the reactions of class 18 is in strong conjugation with the carbonyl group. This gives a large weight to a zwitterionic resonance structure, in stark contrast to the situation of the reactions in subclass 11.2 where there is only an unsubstituted phenyl ring or a *p*-nitrophenyl group. This fact is perceived by the procedures for the evaluation of electronic effects, leading to an unusually high value of D^+ on C-2 for heterolysis of the C=C double bond (Table 10). This high value of D^+ is found to be sufficient reason for establishing a unique class.

All reactions of the four classes (classes 11, 17, 18, and 24) involve the reaction of a C=C double bond that is in conjugation to at least one carbonyl group (or its aza analog). Thus, one might term these reactions also as Michael additions, and in fact, one of the research groups having performed the original experiments[22] preferred to name its reactions (in subclass 11.2 and class 18) Michael additions. However, as it is very difficult to generate a carbanion from an aromatic C–H bond, we have preferred to classify all of these reactions as Friedel–Crafts alkylations. This terminology is supported by the fact that nearly all reactions proceed under catalysis by acids or Lewis acids.[21-23] Clearly, HORACE avoids this semantic problem but realizes that these reactions are definitely different from those put into classes 2 and 13 (that we have named Michael additions).

Free Radical Addition. Michael additions and Friedel–Crafts alkylations comprise the large bulk of reactions that fall into the scheme chosen for this investigation. The addition of radicals, formed by homolysis of the C–H bond to alkenes, was expected as another type of reactions for this scheme. Only five reactions in this dataset of 120 reactions were identified by the experimental investigators as radical reactions. HORACE placed these reactions into four classes shown in Figure 12.

The reaction of class 14 is quite unique;[24] there are no other instances similar to it in the entire dataset. The radical is generated at a C_{sp^2} center. This warrants it being classified as a separate group. Class 19 comprises a single reaction. The rather unusual addition to a simple alkene, cyclohexene, without any functionality is initiated by an electron transfer reaction that generates a radical.[25] It is quite important to note that one of the starting materials of the reaction of class 19, ethyl cyanoacetate, is a reagent typically involved in Michael additions. However, HORACE does not fall into the temptation of putting this reaction into the large bulk of Michael additions (class 2). Rather, HORACE perceives

(21) Soliman, A. Y.; Sayed, M. A.; El-Bassiouny, F. A.; El-Hashash, M. A. *Indian J. Chem., Sect. B* **1991**, *30*, 754.

(22) Lüönd, R.; Neier, R. *Helv. Chim. Acta* **1991**, *74*, 91.

(23) Balsamini, C.; Duranti, E.; Salvatori, A.; Spadoni, G.; Barone, D. *Farmaco Ed. Sci.* **1990**, *45*, 1111.

(24) Macias, F. A.; Molinillo, J. M. G.; Collado, I. G.; Massanet, G. M.; Rodriguez-Luis, F. *Tetrahedron Lett.* **1990**, *31*, 3063.

(25) Shundo, R.; Nishiguchi, I.; Matsubara, Y.; Hirashima, T. *Tetrahedron* **1991**, *47*, 831.

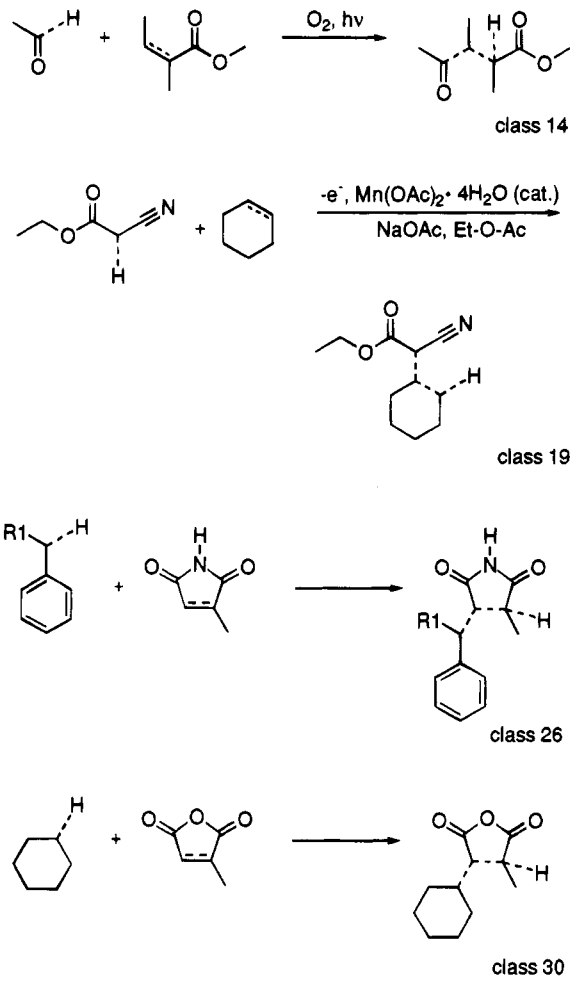


Figure 12. Four classes of free radical additions.

that the other reaction partner is an alkene devoid of functionality and thus has distinct characteristics of its own that warrant its categorization into a class of its own.

The two reactions of class 26 and that of class 30 show similarities and come from the same laboratory.[26] However, in the reaction of class 30, a molecule, cyclohexane, is reacting that is devoid of any functionality, quite in contrast to the compounds met in class 26 that can form quite stabilized benzyl radicals. The example of class 30 clearly extends our knowledge on which compounds can undergo an addition as radicals to double bonds. Thus, in addition to being justified, HORACE's classification of two sets of reactions into two separate classes also provides didactic benefit.

Photochemical Reactions. Several photochemical reactions were found in the classification process and put into classes 1, 4, 5, and 23, respectively. For reasons of brevity, Figure 13 shows only the generalized representation of class 1 consisting of three reactions.[27] From this class representation, one can immediately see that the three corresponding individual reactions are extremely similar. It should be noted that the generalized representations of this type have widely been used by organic chemists to represent a set of similar reactions under investigation. This fact again indicates the usefulness of HORACE for organic chemists.

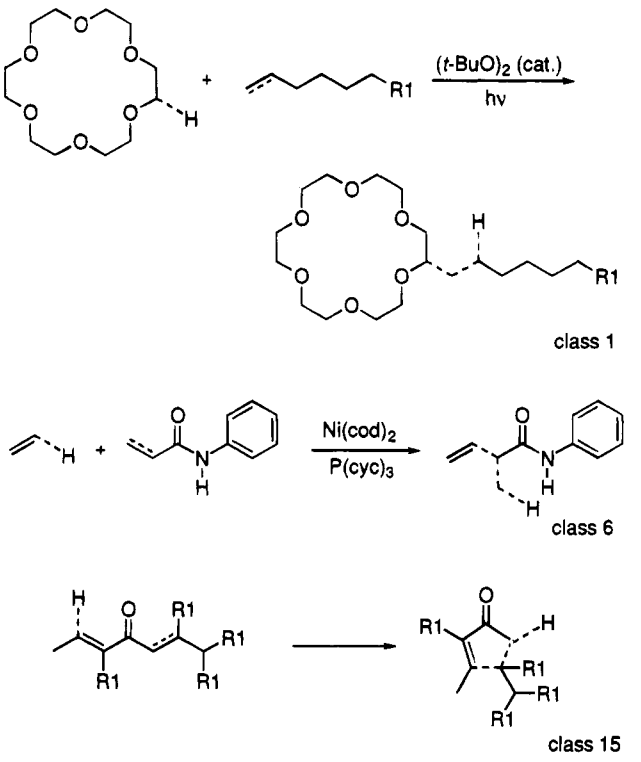


Figure 13. Generalized reactions for class 1 (photochemical reaction) and classes 6 and 15 (diverse reaction types).

Diverse Reaction Types. The classes obtained by HORACE consisting of only one or a few reactions show additional diverse reaction types underlining the variety of reactions falling into the scheme chosen here. From the four classes only two are shown here in Figure 13.

Class 6 consists of one instance, showing the rather unusual addition of a C–H bond of ethene to the double bond of a *N*-phenylacrylamide.[27] The pecularity of this reaction is underscored by the rather special organometallic catalyst.

The generalized representation of class 15 (see Figure 13) keeps the all important structural features of the reaction center, and thus, from this representation, we can see that the five individual reactions exhibit a high degree of similarity; they are, in fact, all Nazarov reactions initiated by concentrated sulfuric acid.[28]

Errors in the Reaction Database. Classification of reactions by HORACE can also aid in the detection of certain types of coding errors in a reaction database. Reactions with coding errors can be found in those classes of reactions that have been singled out, having been put into special classes that show unusual reaction behavior which, in fact, is not borne out in reality.

Figure 14 shows two examples. The first case is a reaction[29] that had been assigned an incorrect reaction center during the building of the reaction database. HORACE found such a bond rearrangement scheme unusual, in the context of the other reactions in this study, and therefore put this reaction into a class of its own (class 3). Closer inspection of this reaction by inquiry of the primary literature[29] showed that this reaction does not fall into the reaction scheme studied

(26) Giese, B.; Farshchi, H.; Hartmanns, J.; Metzger, J. O. *Angew. Chem.* **1991**, *103*, 619.
(27) Hoberg, H.; Ballesteros, A.; Sigan, A.; Jegat, C.; Bärhausen, D.; Milchereit, A. *J. Organomet. Chem.* **1991**, *407*, C23.
(28) Motoyoshiya, J.; Yazaki, T.; Hayashi, S. *J. Org. Chem.* **1991**, *56*, 735.
(29) Cheskis, B. A.; Isakov, Ya. I.; Novikov, A. V.; Moiseenkov, A. M.; Minachev, Kh. M. *Izv. Akad. Nauk SSSR Ser. Khim.* **1990**, *4*, 902.
(30) Nelson, P. H.; Nelson, J. T. *Synthesis* **1991**, *3*, 192.

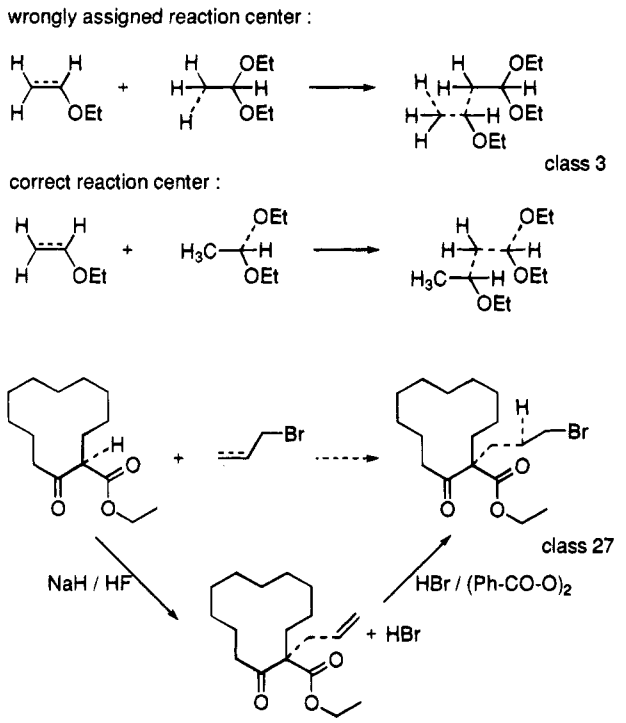


Figure 14. Reactions with an error in the reaction center assignment. In the first reaction, a C–O bond is broken and not a C–H bond. The second reaction involves a two-step sequence of an alkylation and then an HBr addition.

here but involved a bond rearrangement process that is different from the one chosen here.

Another rather unusual reaction[30] was also put into a reaction class of its own (class 27). Again, the reaction scheme coded in the database is the one chosen as the basis of this study. However, it is a two-step reaction and follows a different mechanism as indicated in Figure 14.

Thus, the analysis of those classes of reactions with only a few entries and unusual distributions of a functional group can point out errors in the coding or assignment of reaction centers.

Summary

The automatic classification of organic reactions can greatly extend our knowledge about chemical reactions. It involves a process of learning about the scope and limitations of a certain reaction type from a series of individual reactions. This approach is based on a learning method, inductive learning, which has been used successfully by chemists from the very beginning.

Through classification and generalization of reaction instances, HORACE is able to find the major reaction types, point out the more unusual reactions, and identify anomalies that may turn out to be coding errors in a reaction database. The secret to the success of a study with HORACE is its unique consideration of both electronic effects and functional groups around the reaction center.

The processing of a dataset of 120 reactions involving the addition of a C–H bond to a C=C double bond has underlined the overall importance and broad scope of the Michael addition. It also showed examples of some of the more unusual Michael additions.

The next important reaction type is the Friedel–Crafts alkylation of aromatic compounds by alkenes. A set of reactions was correctly classified as consisting of members of this reaction type. However, this classification also showed that there are reactions where a clear-cut classification into either a Michael addition or a Friedel–Crafts alkylation becomes questionable. Several reactions were found that proceed through the addition of a carbon radical to the C=C double bond, and these reactions were faithfully compiled into a separate reaction class.

Thus, the processing of a fairly small reaction dataset has already shown that a reaction database is a rich source of information and has underscored its importance as the basis for the automatic extraction of knowledge of chemical reactions from individual reaction instances. Furthermore, this study underlines the advantages of a data-driven approach.

Acknowledgment. We are grateful to the Alexander von Humboldt Foundation for the support provided to J. R. Rose and L. Chen in the form of research fellowships. We thank Fachinformationszentrum (FIZ) Chemie, Berlin, Germany, and MDL Information Systems Inc., San Leandro, CA, for providing us with ISIS Host and the ChemInform reaction database. We thank the editor and one referee for their suggestions for improvements in the presentation.

JO950508Y