

HORACE: An Automatic System for the Hierarchical Classification of Chemical Reactions

J. Royce Rose and Johann Gasteiger*

Organisch-Chemisches Institut, Technische Universität München,
Lichtenbergstrasse 4, D-85747 Garching, Germany

Received July 14, 1993*

An automatic system for hierarchically classifying chemical reactions is presented. A detailed analysis of the classification of reactions based first on topological and then on physicochemical features is performed on two data sets. Each of these studies is concluded with an analysis of the hierarchical classification produced by the HORACE algorithm which combines both approaches in a synergistic manner. Searching and updating of reaction hierarchies is then demonstrated with the hierarchies produced for the two data sets by the HORACE algorithm. The reaction hierarchies provide an efficient access to reaction information. They indicate the major reaction types for a given reaction scheme, define the scope of a reaction type, allow one to find unusual reactions, and can help in locating the reactions most relevant for a given problem.

1. INTRODUCTION

In recent years a cornucopia of information on chemical reactions has become available in computer-readable reaction databases. Databases with more than a million reactions have already been compiled.^{1,2} Others are growing by as many as 60 000 reactions annually.³⁻⁵ However, the rapidly increasing amount of information stored in large databases also causes problems. Searches may result in such a large number of hits that the user is no longer willing to scan through them. Clearly, limiting the number of reactions in a database cannot be the answer to this problem. Each additional reaction brings with it some specific information and thus adds to the knowledge and value of the database. A guiding mechanism that helps the user to find the information he requires is needed. This method should provide him with a focused answer to the question he has in mind and should produce a minimum number of hits that are peripheral to the problem at hand.

The information on individual reactions stored in a reaction database can also be used to derive knowledge about entire classes of chemical reactions. Chemists have historically learned about reaction types from series of observations on individual instances of reactions. Several different approaches have been developed to automatically derive knowledge from reaction databases that can be used for the prediction of chemical reactions or for the design of organic syntheses.⁶⁻¹³

Two approaches are briefly mentioned to illustrate the broad potential that currently lies untapped in reaction databases. Databases containing values of experimentally observed reaction rates have been used to derive quantitative mathematical functions based on physicochemical variables. These functions predict absolute rate constants for specific reaction types like acid and base catalyzed amide hydrolysis.¹⁴ On the other end of the scale, inductive and deductive machine learning techniques have been employed to extract schemes of retroreactions from reaction databases for use in computer-assisted synthesis design.⁷

Managing the information in a reaction database is not an easy task since the kinds of problems that one has to solve in planning reactions can be quite different. First, one might be interested in the different types of reaction that can achieve a desired conversion: What kind of ester cleavages are known? Next, one wants to know whether a specific reaction can be achieved in the molecule that one works with. Can I perform

an acid catalyzed ester hydrolysis although my molecule also contains a β -lactam ring? Obviously, getting all ester cleavages in the reaction database is not the desired solution to either one of these problems. In the first problem, one wants to obtain a representative of each of the various *types* of ester cleavages. In the second problem, all those acid catalyzed ester hydrolyses of compounds having a β -lactam ring are desired. The problem of obtaining both general reaction methodologies and specific instances of reactions can be solved by a hierarchical organization of reactions (Figure 1).

In this paper we are presenting data-driven machine learning techniques for finding the inherent hierarchy in a data set of reactions. These methods are combined in the HORACE system (Hierarchical Organization of Reactions through Attribute and Condition Education).

2. OBJECTIVES

A hierarchical classification allows one to answer the following questions:

What are the major reaction types contained in a reaction database (RXDB)? An answer to this question can be found by looking for those classes of reactions that contain many instances.

Which unusual reactions are contained in a RXDB? This involves searching for those classes of reactions that contain only a few members.

From a practical point of view a hierarchical classification of a set of reactions can be performed both on an entire reaction database or on a hit list of reactions, obtained from a query and having too many hits for convenient sequential processing. Classification of a complete reaction database allows one to better navigate through the flood of available information in order to quickly find the desired information on a reaction type or a specific instance of a reaction. However, this requires that the functionality of the reaction retrieval system be extended to be able to work with a hierarchically organized reaction database.

Alternatively, when only the hit list resulting from a reaction query is to be hierarchically classified, a process independent from the retrieval system can be started. This process, taking advantage of the hierarchical classification, allows the user to traverse the hierarchy going directly to the desired information without having to sequentially scan the entire list of reactions.

* Abstract published in *Advance ACS Abstracts*, January 15, 1994.

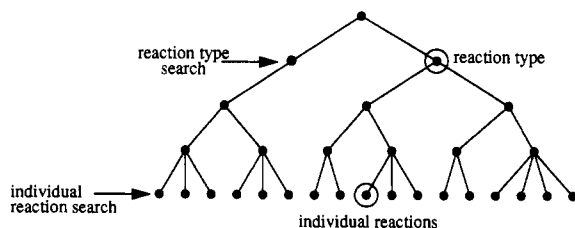


Figure 1. Reaction hierarchy, a prerequisite for searching for both reaction types and individual reactions.

We will also show how a query can be launched into a hierarchically organized data set of chemical reactions and how an *update* of such a data set can be performed. Furthermore, the hierarchical classification of reaction databases provides a foundation for the *comparison* of different RXDBs: how many different types of reactions are contained in a RXDB, how large is the overlap between two RXDBs, etc.

Finally, the classification of reactions involves a generalization phase which is implicit in the formation of classes of reactions. The classification and generalization of instances of reactions is equivalent to the *extraction of knowledge* from individual examples of reactions. Such automatic extraction of knowledge from a reaction database can provide a basis for the representation and manipulation of reaction types in systems for *reaction prediction* and *synthesis design*.

3. THE METHODOLOGY

In principle, creating hierarchical classifications is quite simple. It is only necessary to define the criteria on which the classification should be based and an efficient method for computing these criteria. In practice, however, things are much more difficult. Ideally, one might want to classify reaction instances by reaction mechanism and then refine the hierarchy by considering the relative functionality of the chemical structures involved. In fact, a machine learning system, TRISTAN,⁷ has been developed that takes advantage of a system that models reaction mechanism.¹⁵ Unfortunately, although good results have been achieved, reaction-mechanism modeling systems have not yet matured to the point that they could form part of a general-purpose robust algorithm for classification of reactions by reaction mechanism. The *model-driven* approach exemplified by TRISTAN is limited by the understanding of chemistry coded into the model. This fundamental limitation is not shared by data-driven methods. A *data-driven* approach, on the other hand, is primarily limited by the data it is processing. Seen in a different light, the vast amounts of chemical knowledge present in current reaction databases make a very powerful data-driven approach possible. This was the motivation for pursuing a data-driven approach and for finding classification criteria that are more practical but still somehow reflect the underlying mechanism. A look at the kind of general information contained in a reaction suggests practical classification criteria that are less knowledge-intensive and thus require less coding of chemistry knowledge in the program.

There are three kinds of fundamental information specified in a reaction: the reaction transformation; what is required by the reaction; what is allowed by the reaction.

The essence of the reaction transformation is embodied in the reaction center, the atoms and bonds directly involved in bond and electron reorganization. Clearly, if two reaction instances have reaction centers that differ in the number and/or kinds of bond changes, then they do not represent the same

reaction type, although they may be related. Thus, the reaction center can be used as a crude classification criterion. And, in fact, the classification algorithms that will be discussed first partition the reactions according to reaction center.

The last two points, the reaction requirements and reaction tolerance, specify the reaction context. They are directly dependent on the reaction mechanism. Unfortunately, for most reaction instances it is not obvious what part of the description belongs to which point. The second point, the reaction requirements, gives an indication of the driving force behind the reaction. It answers the question: what motivates the reaction? The third point describes what is compatible with the reaction. It hints at the kind of substructures that could survive the reaction unchanged.

Thus, a reasonable approach to classifying reactions is to characterize them according to these three points. A hierarchy can then be constructed by comparing the resulting characterizations at different levels of abstraction.

3.1. Hierarchical Descriptions. A first step in developing a system for creating a hierarchy of related reactions is the formulation of a language for describing the reaction characterizations. This language should be capable of describing specific reaction instances as well as abstractions of reaction types. Generally, molecules are represented by graphs in which atoms are the graph vertices and bonds are the connecting edges. Reactions are then represented by sets of educt graphs and sets of product graphs joined by reaction arrows. In addition, chemists have developed symbols to represent equivalence classes of atoms and fragment variables. For example, X is often used to represent halogen atoms and fragment variables like R can be used to represent generalized substructures such as arbitrarily long alkyl chains. This allows the chemist a limited amount of description generalization which is usually sufficient for describing closely related instances of reactions. However, a more extensive description language is required to support rich hierarchical classifications.

The description language used by HORACE consists of components which represent three levels of abstraction. At the lowest level of abstraction is the partial order of atom types described in the following section. This partial order gives an explicit hierarchy at the atom level. It specifies the degree of similarity between atoms. The next level of abstraction is the structural level which builds on the first level. HORACE makes use of a list of structural features (functional groups) to characterize molecules. This characterization is anchored to the reaction center. That is, the features used to characterize the molecules in a reaction are specified with respect to the atoms in the reaction center (section 3.1.2).

At the highest level of abstraction are physicochemical variables which describe aspects of the *functionality* of chemical structure (section 3.1.4). These variables can be used to recognize similarity between structurally dissimilar reaction instances. They support a much higher level of abstraction than that based on structural features alone. Although it may be possible to try to simulate this with a hierarchical classification of structural features, the results produced by any static classification of structural features would be inferior. The rich possibility of combination of structural features in organic chemistry is so great that the combinations defy any practical comprehensive explicit enumeration. Fortunately, physicochemical variables are much more flexible than any possible fixed hierarchy of structural features since they are context sensitive. Thus, they are able to reflect the richness of interactions between

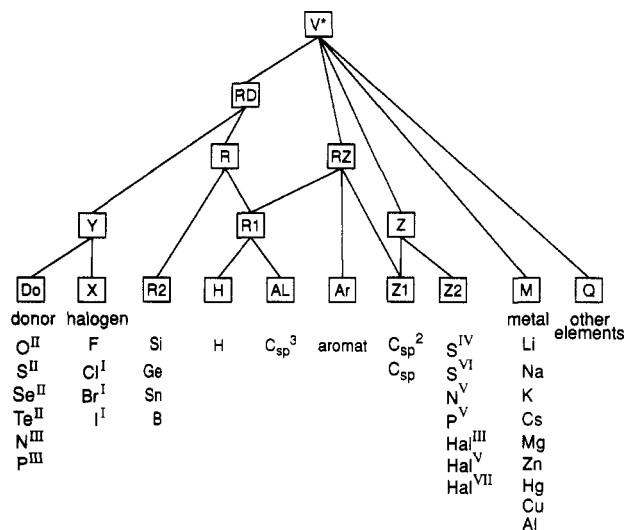


Figure 2. Atom type hierarchy for generalization.

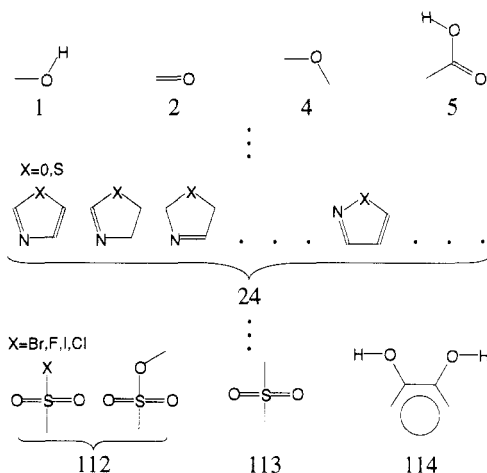


Figure 3. Some of the 114 structural classification features used by HORACE.

structures found in organic chemistry, albeit with certain accuracy limitations.

3.1.1. Hierarchical Ordering of Atoms. Chemists often organize atoms into equivalence classes according to their chemical similarity. HORACE's lowest level of abstraction is based on a hierarchy of atom types (Figure 2). It consists of ten basic equivalence classes grouping the individual atoms, six intermediate equivalence classes representing abstractions of some of the base classes, and a general universal atom type. This partial order of atoms and atom types is used to support the classification and generalization of reactions (sections 3.3 and 3.4). It provides a basis for hierarchically classifying substructures at the level below that of functional groups. In the current version of HORACE, atom types are used to describe the generalization of atoms. In the case of nonterminal generalized atom types, they also represent fragment variables as well as the generalization of atoms. Thus they support a medium level of structural abstraction. The use of the hierarchy of atom types in reaction generalization is described in section 3.4.

3.1.2. Structural Classification Features. A set of 114 structural features is used by HORACE to characterize reactions having the same reaction center (Figure 3). This set was developed as part of the BRANGÄNE⁷ system and is actually a subset of a larger existing set of structures that was originally created for use in the SYNCHEM2 synthesis planning system.¹⁶ The current set of features has been arrived

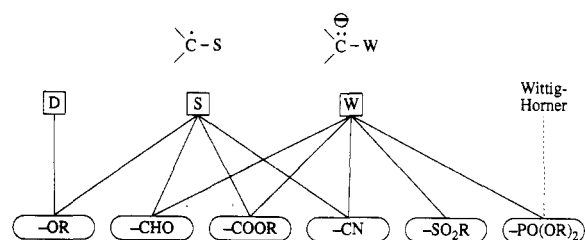


Figure 4. Different classifications of the substituents at the reaction site dependent on whether a free radical reaction or a nucleophilic reaction is being considered (D, donor group; S, radical stabilizing group; W, electron withdrawing group).

at empirically. Initially a set of 140 structures were considered, and then quasi-ubiquitous structures were discarded. Clearly, structures that occur with very high frequency tend to skew any similarity measure. Such structures are not very useful as discriminating features. They tend to give the misleading impression that most reactions are very similar by outweighing those features that indicate dissimilarity. This set of features is contained in an external ASCII file that is read in at run time. Since these features are external to the program, it is easy to modify them. In addition HORACE uses a vector to tell it which features it should use. Thus although HORACE reads in all of the features in the external file, it only uses those features indicated by this vector. This makes it even easier to select or deselect classification features.

3.1.3. Static vs Dynamic Classification. The structural classification features in the previous section are very useful for distinguishing between different types of reactions because of the medium level of abstraction they represent. However, this level of abstraction is often too low to recognize when specific reaction instances or reaction sets should be grouped together. An answer to this problem can be given by a hierarchical organization of the functional groups used as structural classification features. Thus, functional groups such as aldehyde, ester, nitrile, sulfone, phosphonate, etc., could be grouped together to form a general electron withdrawing class, W (Figure 4).

This class of functional groups could then be used to represent the influence of substituents on the reaction center of a series of reactions in order to generate an additional level of abstractions in the classification. It is true that the class of electron withdrawing groups, W, can help in the classification of Michael additions. However, the various functional groups cannot be grouped together in a generally valid manner. Two warnings have to be given that make the use of such generalized functional groups highly questionable.

First, once a substituent of this group (e.g. $-\text{PO}(\text{OR})_2$) is found in a specific reaction, one cannot automatically generalize it to its common representation (W) since this would imply that any member of this class of functional groups could just as well be present at the reaction center. In the case of the $\text{PO}(\text{OR})_2$ group, any other group of the W-class can be activating on the Michael addition. However, for the Wittig-Horner reaction, it must specifically be the $\text{PO}(\text{OR})_2$ group. None of the other members (CO_2R , CN , etc.) can replace the phosphonate group to make this reaction go.

Second, the way the various functional groups are grouped together is dependent on the reaction being considered. Thus, whereas the $-\text{CHO}$, $-\text{COOR}$, and $-\text{CN}$ groups can be classified as electron withdrawing and the $-\text{OR}$ group as electron donating, both types of substituents have to be combined into a common class, S, when it comes to reactions that proceed through free radicals (see Figure 4).

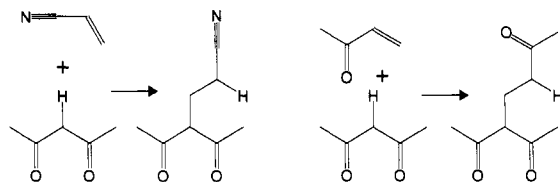


Figure 5. Two related reaction instances with different functional groups activating the reacting olefin bond.

In summary, no static classification of functional groups that is generally valid can be given. Instead, the effects exerted by a functional group are dependent on the type of reaction being considered and have therefore to be derived dynamically from available reaction information. Thus, rather than imposing a *static model* on the classification, the *data* have to *dynamically* determine the classification.

3.1.4. Physicochemical Classification Features. Although the structural classification features described in section 3.1.2 can be used to distinguish between different types of reactions, the level of abstraction is not always sufficient for recognizing when reactions should be grouped together. This is particularly true of reaction instances that are mechanistically related but happen to have different complements of functional groups around the reaction center. Figure 5 shows two such reactions. In one case, the olefin bond is activated by a ketone and in the other by a nitrile group. Both the ketone and the nitrile are strongly electron withdrawing structures and function identically in these two instances. However, the relatively low level of abstraction afforded by that part of the classification facility based purely on structural features does not recognize this similarity.

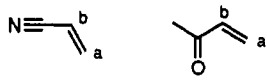
We strongly believe that any progressive scheme for the classification of chemical reactions should be able to uncover and incorporate the *physicochemical* driving forces of chemical reactions, since the electronic, energy, and steric effects determine a reaction mechanism and the reaction mechanism determines the reaction conditions. Clearly, common reaction mechanisms would form the soundest foundation for reaction classification. However, currently the mechanisms of reactions are not stored in reaction databases and are quite often not known. Access to physicochemical variables at the reaction site constitutes the closest approach to the ideal of knowing the reaction mechanism. In fact, it has been shown that these physicochemical variables can successfully be used for the derivation of a reaction mechanism.²⁴

The similarity between the two educts is obvious as soon as physicochemical parameters expressing the functional effects of the nitrile and the ketone on the olefin bond are examined (Table 1). For example the σ - and π -electronegativity values, χ_σ and χ_π , for the corresponding atoms of the olefin are fairly close. Also note similar values for the resonance stabilization of negative and positive charges (R^- and R^+ , respectively).

From the outset it was clear that large data sets of reactions would have to be processed. This places a significant constraint on the amount of computational time allowable for calculating physicochemical parameters. We therefore resorted to rapid empirical methods for the calculation of atomic charges,^{17,18} of parameters for the inductive,¹⁹ resonance,²¹ and polarizability effects,²⁰ and of bond dissociation energies.²²

The algorithms have been designed such that basic atomic or bond parameters are combined in a way that takes into account the constitution of a molecule. This ensures a rapid progression of the algorithm through the molecular structure. Thus, the calculation of the σ -atomic charges, q_σ , and

Table 1. Comparison of Physicochemical Features for Two Educts from Figure 5

| | |  | |
|--------------------|------|--|---------|
| | | Educt 1 | Educt 2 |
| parameter | atom | educt 1 | educt 2 |
| χ_σ (eV) | a | 7.99 | 7.92 |
| | b | 8.96 | 8.55 |
| χ_π (eV) | a | 5.65 | 5.74 |
| | b | 5.67 | 5.48 |
| q_σ (e) | a | -0.089 | -0.097 |
| | b | 0.016 | -0.028 |
| q_π (e) | a | 0.049 | 0.065 |
| | b | 0.0 | 0.0 |
| R^- (eV) | a | 0.0 | 0.0 |
| | b | 7.17 | 7.48 |
| R^+ (1/eV) | a | 0.0 | 0.0 |
| | b | 3.73 | 3.58 |

-electronegativities, χ_σ , basically consist of two nested do-loops. Atomic π -charges, q_π , and -electronegativities, χ_π , are obtained through generation and weighting of the various resonance structures of a molecule. The resonance parameters R^+ and R^- quantify the stabilization of a positive or negative charge obtained in the formal polar breaking of a bond. Bond polarizabilities, α_b , and bond dissociation energies, BDE, are obtained by additivity schemes.

As discussed in section 3.1.3, functional groups are often related in different ways, depending on the type of functionality one is considering, e.g. radical stabilization vs electron withdrawing ability (see Figure 4). One of the most difficult problems that occurs in classifying chemical structures on the basis of functional groups is this *nonorthogonality*, the dependence of similarity on the reaction context (reaction center, mechanism, and reaction conditions). How should the similarity between functional groups be quantified, and which similarities should be considered? To complicate matters even further, chemical structures are context sensitive, meaning that the effect of a functional group is influenced by the structures surrounding it. This makes the Sisyphean task of trying to quantify the nonorthogonality by creating equivalence classes or partial orders of structures even more hopeless. Fortunately, a solution can be provided through the use of physicochemical parameters. They make it possible to quantify the degree of similarity between chemical structures along several different dimensions, reflecting the influence of context and the manifold functional aspects of chemical structure.

3.2. Reaction Center Characterization. The reaction instances are analyzed and their reaction centers are characterized in terms of their complements of structural classification features (section 3.1.2) and physicochemical features (section 3.1.4). The search for structural classification features is restricted to the near vicinity of the reaction center. A heuristic is used whereby only those features within one bond length of the reaction center are considered. Additionally, classification features may not be wholly contained in the reaction center. This last restriction is fairly obvious. Recall that one classification criterion mentioned earlier is the reaction center (reaction transformation). All reactions having the same reaction center would then share the same classification features wholly contained in the reaction center. These features would then act as a bias, distorting the subsequent measure of similarity by overemphasizing the reaction center.

The classification features are associated with those atoms within the reaction center which are within one bond length.

In addition, this association specifies the degree of proximity. A distinction is made between the case where an atom in the reaction is contained in a classification feature or whether it is one bond length from the closest atom in the feature. Limiting the search for structures proximal to the reaction center approximates the attenuation of the effect of a functional group as a result of its distance from the reaction center. It also has the benefit of reducing the amount of computation required to characterize reactions.

In contrast, the calculation of physicochemical parameters takes into consideration the complete ensemble of molecules involved in the reaction. This has the advantage of making it possible to recognize the influence of distal functional groups that may be conducted via vinylology. As the time complexity of these calculations is significantly less than that of the subgraph embedding required to recognize functional groups, processing the entire ensemble of reacting molecules does not significantly add to the computational burden. HORACE presently characterizes reaction centers and their α -substituents in terms of their electronic effects derived from the σ - and π -electronegativity. Additionally, aromaticity and the atom hybridization state is also taken into account.

3.3. Classification of Reactions. Two classification strategies which may be combined are provided by HORACE. These classification modes are based on topological features and electronic and energy parameters, respectively. Each approach has certain advantages. Classification based on topological features results in a grouping of reactions that chemists find intuitive. However, it is possible to provide a greater degree of abstraction by classifying reactions on the basis of physicochemical attributes since the classification features in this case describe functionality of chemical structure.

3.3.1. Topological Classification. The approach taken in classifying reactions on the basis of their complement of functional groups proximal to the reaction center is a modified single-pass "leader" clustering algorithm. The standard leader algorithm is a nonhierarchical method in which each item is in turn compared with the clusters formed thus far. If an item is sufficiently close to an existing cluster, it is placed in the nearest cluster; otherwise it is used to start a new cluster. We have changed this nonhierarchical method into a hierarchical method by taking the resulting clusters and specializing them, looking for subclusters. One way of looking at this approach is as a combination of a leader algorithm followed by a polythetic divisive method.

3.3.1.1. Cluster Formation. The formation of clusters proceeds in two phases, the discovery of cluster cores and the assignment of reactions to clusters. As a first step, the reactions are sorted according to their number of proximal functional groups. The reaction having the fewest functional groups is at the head of the list. By convention this first reaction is chosen as a cluster core. With respect to the remaining reactions in the list that belong to the same as yet undiscovered class, this reaction has a minimal complement of functional groups. Thus, it may be an example of the reaction with a minimal required set of contextual functional features. This implies that other reactions in the list that are similar probably belong to the same class and that the additional functional groups they possess are not part of the required context of the reaction but rather part of the variable portion.

Additional cluster cores are found by comparing the remaining reactions with the set of current cluster cores using the closeness metric shown below. If a reaction is not sufficiently *close* (the closeness threshold is a parameter to

the program) to any existing cluster core, it is declared a new cluster core. Although it is possible to use different closeness values, a threshold of $2/3$ match using the following closeness metric is typically used. The rationale for choosing a threshold value of this approximate size will be described in the discussion on subcluster formation.

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

This closeness metric compares two feature sets, A and B . This asymmetric measure is used to determine the closeness of a reaction B to an existing cluster core A . Thus, unlike the Tanimoto index, the universe of relevant features is not the union of features in sets A and B , but simply set A , the set of features belonging to the cluster core A .

The second part of cluster formation involves assigning the remaining reactions to the cluster or clusters they most closely match, again using the asymmetric closeness metric (eq 1). In the event of a tie, a second tie-breaking metric (eq 2) is

$$\text{closeness2}(A,B) = \frac{|A \cap B|}{\max(|A|, |B|)} \quad (2)$$

used. The reaction is then placed in the winning cluster. Ties are still possible with this second metric, in which case the reaction is placed in all maximally matching clusters. As will be discussed in the section on hierarchical classification (section 3.5), placing a reaction into more than one cluster is an indication of the relatedness of the clusters and motivates the subsequent merging of those clusters at a higher level of abstraction.

3.3.1.2. Subcluster Formation. Each cluster is then analyzed separately for possible subclusters. This entails comparing the reactions within a cluster with one another and associating each reaction with the reaction it most closely matches using the second closeness metric (eq 2). Subclusters are then formed by taking the closure of all maximally matching reactions in a cluster. For example if reaction A maximally matches reaction B and reaction B maximally matches reaction C , then reactions A , B , and C will end up in the same subcluster. In the event the closure of maximally matching reactions is equivalent to the original cluster, no subcluster is formed. This divisive approach depends on the fact that the members of a cluster already have a high degree of similarity. If this were not the case, then taking the closure of maximally related reactions could result in poor subclustering. For example, if reaction A matched only half of the attributes in B and B matched only half of the attributes in C , then it might be the case that A and C have very little in common. This scenario can be avoided by requiring a high degree of similarity for the original cluster formation. It is for this reason that a similarity threshold of $2/3$ was selected in the previous section. Although a higher similarity threshold could be chosen, this would result in a greater number of small clusters. Empirically, the $2/3$ similarity threshold strikes a good balance. Experimentation has shown that the choice of what is being measured is much more important than the precise threshold value. A moderately high degree of match required for cluster formation and the tight focus on relevant attributes appears to preclude poor subclustering.

3.3.2. Functional Classification. The approach taken in classifying the reactions on the basis of electronic effects at the reaction center and the α -substituents is essentially a single linkage agglomerative clustering method. A dissimilarity matrix giving the distance between the reactions is constructed. However, rather than combining reactions in order of increasing distance as is commonly done in agglomerative

clustering, a threshold is used to merge reactions or sets of reactions in a single burst of agglomeration. Thus this method by itself is not hierarchical.

A reaction is placed in the cluster containing all other reactions that are within the specified threshold distance. Since a single linkage measure is used, the end effect is equivalent to taking the closure of all pairs of reactions within the threshold distance. The idea is to quantize the levels of abstraction represented by the merging of clusters. Each increasing threshold level represents a higher level of abstraction. At present, the main use of this clustering approach is two-fold. First, it is used to provide a classification representing a higher level of abstraction than is practical using only topological features. Second, the classification it produces can be used to place constraints on the types of clusters that may be formed in subsequent phases of topological classification.

3.4. Generalization of Reactions. In order to extract knowledge about reactions from collections of related reaction instances it is necessary to be able to generalize reactions. Thus, generalization is an abstraction process. A set of related reactions is generalized to reveal the underlying chemistry. The generalization component used in HORACE is essentially a modified version of ISOLDE.⁸ ISOLDE is able to induce generalizations of reactions as well as already generalized reactions. Before induction is performed on reaction instances, this module deduces contextual features that are not used for classification but are used to provide a more detailed description of the reaction. The deduced attributes include acid, base, water, and temperature sensitivity.

In the context of classification, the purpose of generalization is to produce for each cluster a generalized reaction that serves as a description of that cluster. The reactions in a cluster are processed sequentially in four phases. First, a mapping is found between the reaction center of the cluster member being processed and the current hypothesized generalization. The first member of a cluster is the initial hypothesis. Next, this mapping is extended to atoms in the sphere. During this phase the characterization of α -substituents in terms of electronic effects is used to constrain the possible mappings. Thus α -atoms at strongly electron withdrawing or strongly electron donating substituents must be mapped to the most compatible counterparts before the remaining atoms are mapped. This reduces the possible mappings as early as possible. It also results in generalizations that are chemically more meaningful than those that often result by simply searching for the greatest common subgraph containing the reaction center. Thereafter, the mappings are extended as far as possible. Finally, the mapping is extended one additional sphere outward if possible by using variable nodes (fragments variables) to account for nonmatching structures. At the same time, the physicochemical parameters describing the reaction center and α -substituents are generalized.

3.5. Hierarchical Classification through Iterative Classification and Generalization. HORACE's predecessor, BRANGÄNE, used an algorithm involving alternating phases of classification and generalization to produce hierarchical classifications of reactions. In HORACE we have explored the development of hierarchical classification methods through the combination of structural and functional descriptions of reactions. Our initial approach was to extend the basic BRANGÄNE algorithm⁹ which is described in the following section.

3.5.1. Hierarchical Classification on the Basis of Topological Features. A hierarchical method can be produced from a nonhierarchical approach by taking a given classification

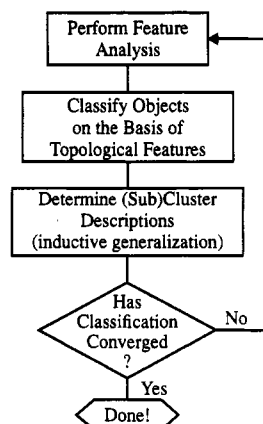


Figure 6. BRANGÄNE hierarchical clustering algorithm for reactions having the same reaction center.

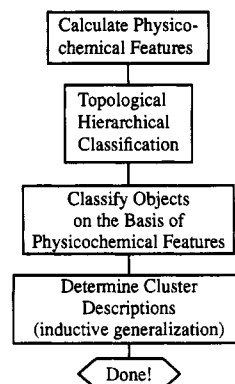


Figure 7. First extension of the topological hierarchical classification algorithm for reactions with the same reaction center.

and generalizing each class and then repeating this process of classification/generalization on the resulting objects until a stable taxonomy is produced. During all subsequent phases of classification and generalization, the objects being processed are generalized class descriptions and not the original objects. In the case of reactions, these are generalized reaction descriptions. This algorithm is shown schematically in Figure 6.

The first step in this algorithm involves analyzing reactions or generalized reactions from previous iterations for their complement of proximal functional groups (section 3.2). These features are then used in the next step for clustering (section 3.3.1). Each of the resulting clusters is then generalized to produce a generalized reaction describing the cluster (section 3.4). Finally, beginning with the second iteration, the current classification is compared to the classification produced in the previous iteration. If they are identical, then a stable taxonomy has been produced and the hierarchical classification is complete. Otherwise, the analysis, classification, and generalization processes are repeated on the generalized reactions. If no objects are combined, then a stable taxonomy has been achieved; combining n objects in a cluster results in there being $n - 1$ fewer objects in the next iteration. This algorithm is guaranteed to converge since objects are combined but never separated.

3.5.2. Hierarchical Classification Based on Topological and Physicochemical Features. The algorithm in the previous section was initially extended by adding a final additional level of clustering on the basis of physicochemical features (section 3.3.2), producing the algorithm shown in Figure 7. This extension was a logical consequence of viewing physicochemical parameters as a more abstract level of description than topological features.

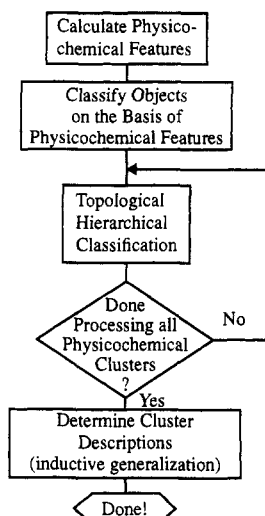


Figure 8. HORACE hierarchical classification algorithm.

3.5.3. The HORACE Algorithm. After some experimentation, it became clear that the topological classification should also be somehow constrained by physicochemical features in order to avoid incorrectly classifying reactions that might be superficially similar in terms of their complement of topological classification features (section 8.2). It was observed that on rare occasion reactions that had been classified together on the basis of topological similarity were later separated on the basis of physicochemical dissimilarity. A second algorithm combining topological and physicochemical classification was developed to address this problem (Figure 8). This is the HORACE algorithm. As will be discussed in sections 5 and 8, this algorithm combines the advantages of the superior abstract classification afforded by physicochemical classification and the more intuitive classification provided by an extension of the hierarchical classification on the basis of topological features.

The key feature of the HORACE algorithm is the manner in which it combines structural and semantic classification approaches. HORACE does not simply compose the two classification methods. Rather, it propagates constraints from the physicochemical phase of classification into that of the topological phase. This is done by first computing the physicochemical classification and then creating a topologically-based hierarchy on each of the resulting clusters (Figure 9). Since the topological algorithm is processing only reactions from one physicochemical cluster at a time, it cannot mistakenly combine reactions from separate physicochemical clusters that might appear to be topologically similar.

Also of great interest is the fact that first classifying reactions on the basis of physicochemical features and then later performing a topological hierarchical classification on the individual clusters makes *lazy evaluation* of the hierarchy possible. Rather than being forced to compute the entire hierarchy, after producing the physicochemical classification it is possible to simply compute topological classifications on those individual clusters that the user asks for. This also has the advantage of potentially reducing the amount of computational effort involved.

4. SELECTION OF A REACTION TYPE

The approaches outlined in the previous section are illustrated with an example. Only reactions with the same reaction center, i.e., with the same set of atoms and bonds involved in the rearrangement of bonds during the reaction,

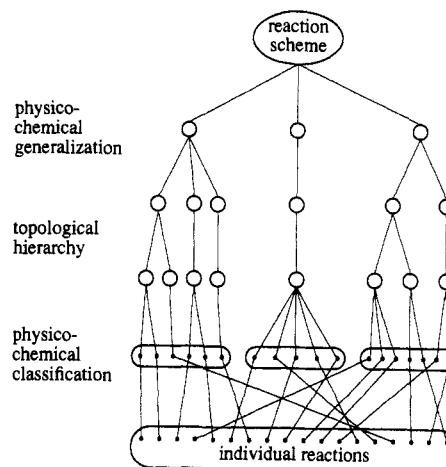


Figure 9. HORACE propagation of physicochemical constraints into the topological hierarchy by first performing the physicochemical classification. The highest level of abstraction is obtained by generalizing the reactions in each physicochemical class after the topological hierarchy has been produced.



Figure 10. Reaction center common to all atoms investigated in the examples.

have been chosen. Reaction centers provide an obvious and important vehicle for classification. The theme of this work is the hierarchical classification of reactions in a reaction database. This hierarchy is in the form of a forest. Each tree in this forest consists of a hierarchical classification of all reactions that share the same reaction center. In other words, all reactions in a database that have the same reaction center will be placed in the same tree.

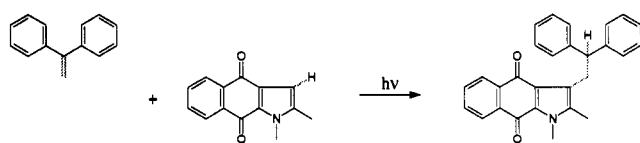
All reactions studied in this example have a reaction center that consists of a CH bond and a CC double bond that react to give a sequence of two CC single bonds and one new CH bond (Figure 10).

This reaction center was selected since it contains several types of reactions proceeding under a variety of reaction conditions and with different reaction mechanisms such as Michael addition, Friedel-Crafts alkylation by olefins, and free radical additions to olefins. First, we will investigate a small data set of 20 reactions with HORACE to show the steps in the classification explicitly. We will then present the results on a slightly larger set of 56 reactions. The first data set was taken from the ChemInform-RX reaction database,^{3-5,23} whereas the second data set was extracted from the Theilheimer reaction database.

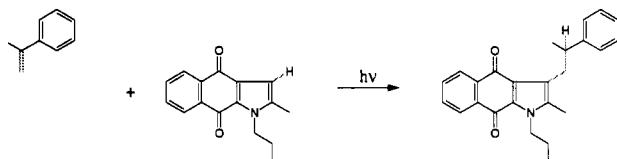
The hierarchical classification will be presented in three studies on two data sets. The first study considers only the identity of atom types and of substructures and functional groups proximal to the reaction center. These may be loosely called topological features as they build on the atoms and bonds contained in the connection table of the reaction partners.

In the second study an assortment of physicochemical variables for the atoms and bonds of the reaction center is also taken into account. The parameters used comprise a variety of electronic and energy effects. These include partial atomic charges,^{17,18} hybridization states, atomic polarizabilities,¹⁹ resonance stabilization²¹ of charges on bond heterolysis, and bond dissociation energies.²² The values of these effects are calculated by empirical methods that are rapid enough to process sizeable data sets of organic reactions with acceptable computation times. The results of the physicochemical classification will be compared with those of the topological classification produced in the first study.

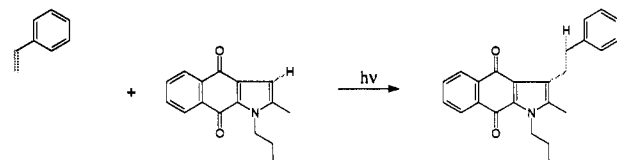
Scheme 1. 20 Reactions of Dataset I



1: ChemInform Refno 9127157 (REACCS Rireg 39828)



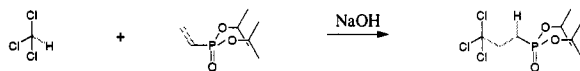
2: ChemInform Refno 9127157 (REACCS Rireg 39829)



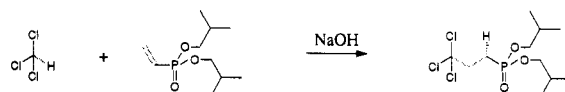
3: ChemInform Refno 9127157 (REACCS Rireg 39829)



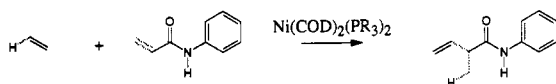
4: ChemInform Refno 9129221 (REACCS Rireg 43686)



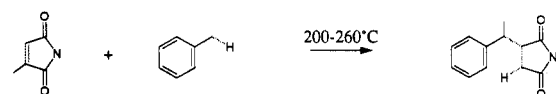
5: ChemInform Refno 9129221 (REACCS Rireg 43687)



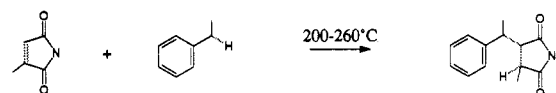
6: ChemInform Refno 9129221 (REACCS Rireg 43688)



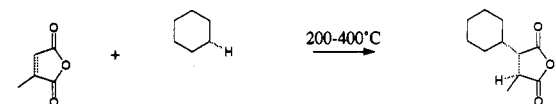
7: ChemInform Refno 9129053 (REACCS Rireg 43829)



8: ChemInform Refno 9129053 (REACCS Rireg 42126)



9: ChemInform Refno 9129053 (REACCS Rireg 42127)

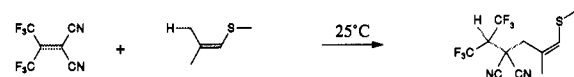


10: ChemInform Refno 9129053 (REACCS Rireg 42125)

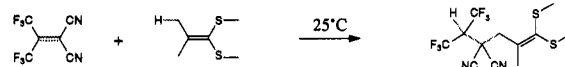
The third study shows the hierarchical classification produced by the HORACE algorithm. We have intentionally shown the results from the first two studies in order to demonstrate how the HORACE algorithm is able to combine the benefits of the two approaches to classification.

5. DATASET 1: 20 REACTIONS

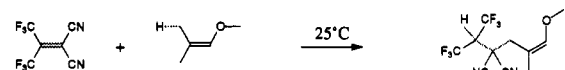
5.1. Study I: Classification by Topological Features. That part of the ChemInform-RX reaction database corresponding



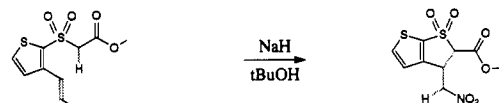
11: ChemInform Refno 9130081 (REACCS Rireg 44090)



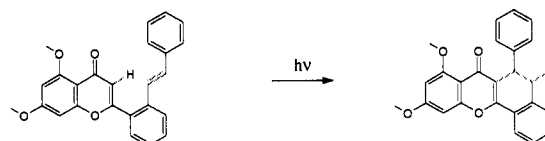
12: ChemInform Refno 9130081 (REACCS Rireg 44093)



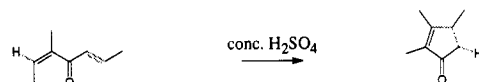
13: ChemInform Refno 9130081 (REACCS Rireg 44089)



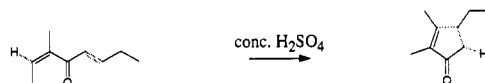
14: ChemInform Refno 9129137 (REACCS Rireg 42761)



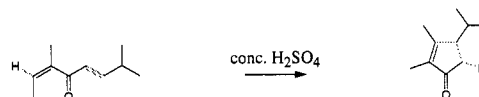
15: ChemInform Refno 9129176 (REACCS Rireg 43209)



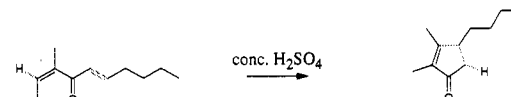
16: ChemInform Refno 9131125 (REACCS Rireg 46166)



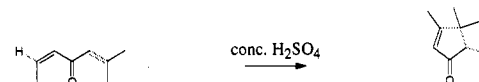
17: ChemInform Refno 9131125 (REACCS Rireg 46167)



18: ChemInform Refno 9131125 (REACCS Rireg 46168)



19: ChemInform Refno 9131125 (REACCS Rireg 46169)



20: ChemInform Refno 9131125 (REACCS Rireg 46174)

to volumes 27–31, 1991, of the printed ChemInform provided 20 reactions with a reaction center as shown in Figure 10. Scheme 1 gives the reaction equations of these 20 reactions. The bonds that are broken or made in the reaction are indicated by textured lines. The topological classification of these 20 reactions results in a hierarchy consisting of two levels (Figure 11). Most of the classification for this data set is achieved in the first level. Only those reactions that end up in class 3 in the final classification are still held in two subclasses.

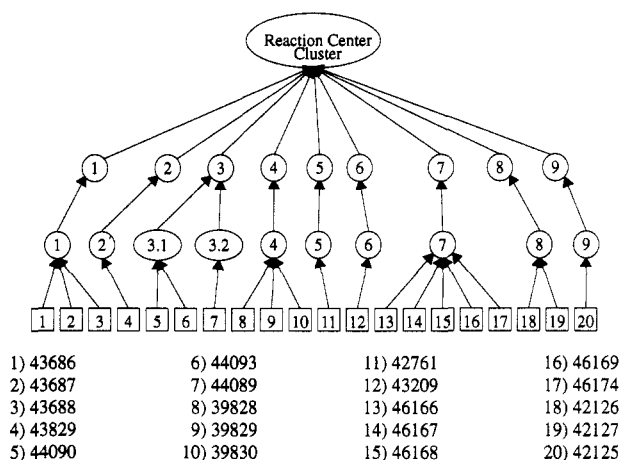
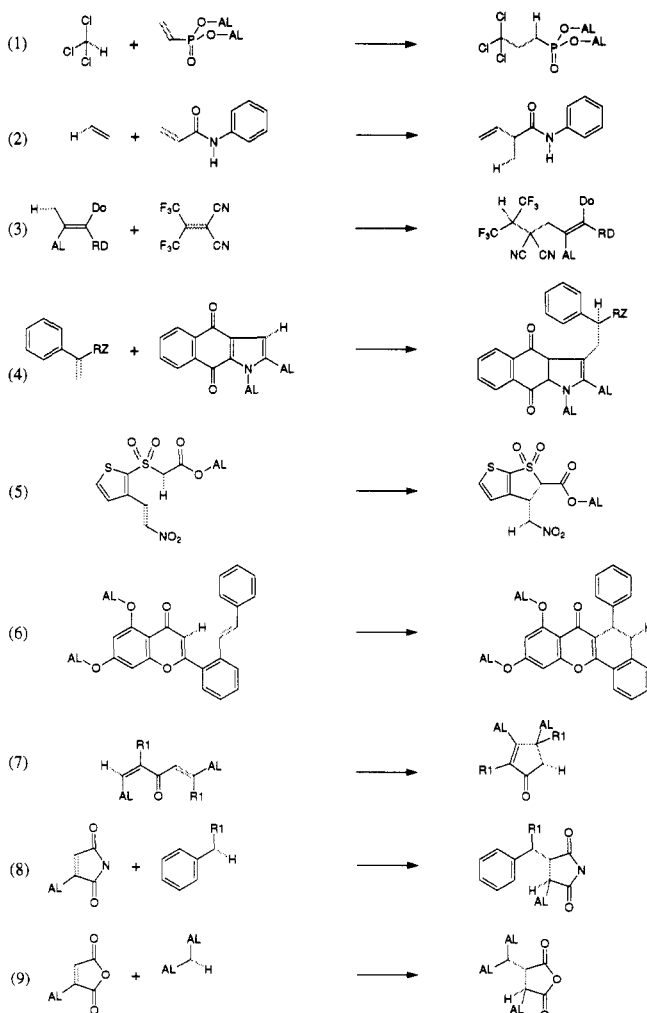


Figure 11. Hierarchical classification of the 20 reactions of data set I.

Scheme 2. Generalized Classes Produced by Topological Classification



However, it is already perceived that these are subclasses; i.e., the reactions contained in these subclasses are somehow similar.

The generalization of reactions in a class isolates those parts of the reactions common to all members in the same class. Scheme 2 shows this generalized notation for the nine classes of the second and final level of classification in the first study. The notation for the generalized atom types is that given in the atom type hierarchy of Figure 2.

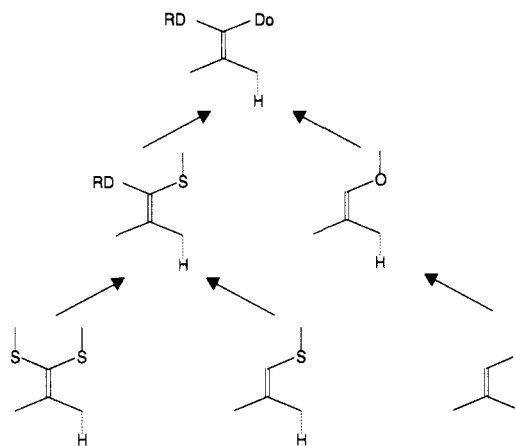


Figure 12. Hierarchical generalization of the reaction partner containing the CH bond of the reaction center in the reactions of class 3 (for an explanation of the symbols RD and Do, see Figure 2).

Class 1 contains three instances of the addition of trichloromethane to phosphono enolates. The only difference in these three reactions is found in the two alkyl groups of the phosphono enolates, that can be either ethyl, isopropyl, or isobutyl. This is reflected by the representation of the alkyl group as the generalized atom type, *AL*.

Class 2 consists of only one reaction, and thus there is nothing to generalize.

The three reactions that finally end up in class 3 show some interesting features of the classification and generalization. All three involve the addition of an allylic CH bond to bis(trifluoromethyl)dicyanoethene. In the first level of classification, the two reactions containing a methylthio group in the allylic system are perceived as having this common feature. In addition, it is found that the carbon atom bearing the methylthio group can have either a hydrogen atom or another methylthio group. The two atoms, H and S, of these two groups bonded to this olefinic carbon are generalized to the *RD* atom type (*RD*: hydrogen, alkyl, or electron donating atoms) (cf. Figure 2). The third reaction having a methoxy group on the allylic system is found to be sufficiently different to warrant keeping it in a separate subclass in the first level of classification.

In the second level of classification the methylthio and the methoxy group are generalized away and replaced by the generalized atom type *Do*, representing donor atoms and encompassing both an O and an S(II) atom (cf. Figure 2). This generalization across atom types in this reaction partner is shown in Figure 12.

Class 4 groups together the three instances of a Friedel-Crafts type reaction of styrene derivatives with the pyrrole ring of a tricyclic system.

The one instance of a CH bond activated by two electron withdrawing groups adding to an electron deficient double bond (Michael addition) forms a class (no. 5) of its own.

Class 6 also contains a single member, the addition to a conjugated double bond of a CH bond of a ring that is aromatic if it is considered in a zwitterionic form.

All five reactions leading to a cyclopentenone are obviously quite similar to each other and form class 7. The reactions differ only in the alkyl substituents proximal to the reaction site. HORACE generalizes them and indicates this by the *R1* and *AL* groups.

There are two reactions involving the addition of a benzylic CH bond to a substituted succinimide, and they are both put into class 8.

The addition of cyclohexane to the substituted succinic anhydride is taken as sufficiently different from the reactions in class 8 to put it in a class of its own (class 9).

It should be noted that HORACE in classifying the reactions solely on the basis of topological features did not produce an extended hierarchy of generalization and classification. Only two levels of classification were produced in this phase, and the generalized reactions are very similar to the individual reactions. This is not surprising, considering the information stored in the ChemInform-RX reaction database and the selection of the data set. The ChemInform reports on new reactions from the literature and then gives a series of examples obtained by structural variations of such a reaction, i.e., by changes in the substituents. The similarity in these structural variations of a reaction type contained in one abstract are already perceived in the first level of classification, and these variants of a reaction are grouped together. However, as these structural changes are usually not very extensive, e.g., varying an alkyl group in the series methyl, ethyl, propyl, etc., there is not much to generalize. Secondly, by selecting reactions from such a narrow time range as five consecutive volumes of the ChemInform covering 5 weeks of reporting from the literature, the likelihood of finding similar reactions in different abstracts is rather low. Thus, beyond the first level of classification and generalization no additional clustering of reactions, except for combining the two subclasses in the third class, can be performed. The reactions in the nine classes are structurally too different to be grouped together. This is where classification based on physicochemical features can be used to great advantage.

5.2. Exercise: Investigation of Physicochemical Variables.

Descriptions that are more general than constitutional aspects of the substituents and functional groups at the reaction center have to be used in order to allow a further clustering of reactions. A selection of physicochemical variables for both bonds directly taking part in the reaction is required. Therefore, a variable was selected for each of the two bonds to illustrate the importance and effects of physicochemical variables.

Among the main features influencing the reactivity of the addition of a CH bond to a CC double bond (see Figure 10) are, at the C_1H bond, bond polarity, the stabilization of a carbanion on C_1 by inductive and resonance effects, and bond dissociation energy; and at the C_2C_3 bond, electron density in the π -part, bond polarity (particularly in the π -part), the electrophilicity, and stabilization of a negative charge on C_2 after addition of a nucleophile at C_3 . These reactions, like most reactions, are influenced by a variety of physicochemical effects. However, to keep the discussion simple, one variable deemed predominant was chosen for each bond. For the CH bond the difference in σ -electronegativity, $\Delta\chi_\sigma(C_1-H)$, was selected as a measure of bond polarity and of the inherent potential to stabilize a carbanion at carbon atom C_1 by inductive effects.¹⁹

For the C_2C_3 double bond, the difference in π -electronegativity, $\Delta\chi_\pi(C_2=C_3)$, was chosen as a measure of the polarity in the π -bond and of the potential to stabilize a negative charge generated by the attack of the nucleophile on this bond.¹⁹

These two variables, $\Delta\chi_\sigma(C_1-H)$ and $\Delta\chi_\pi(C_2=C_3)$, were calculated by empirical methods¹⁷⁻¹⁹ for the reaction centers of each of the 20 reactions of the data set and used for identifying these reactions in a plot of these two variables (Figure 13).

The reactions of each class found in section 5.1 based on a topological classification end up in separate clusters in the

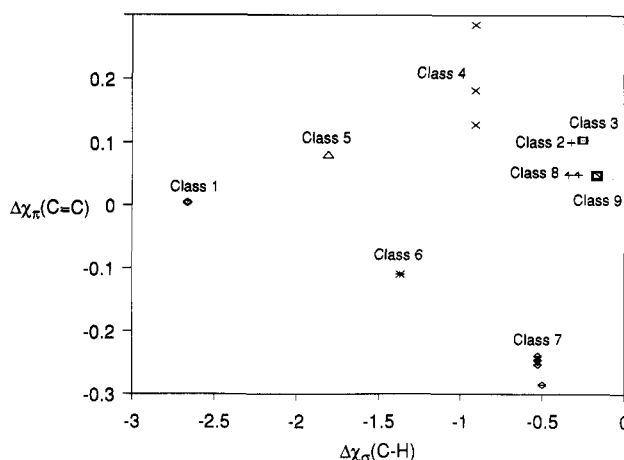


Figure 13. Identification of the reactions of Scheme 1 in a plot of the difference in σ -electronegativity in the CH bond, $\Delta\chi_\sigma(C_1-H)$, and the difference of the π -electronegativities of the C_2C_3 double bond, $\Delta\chi_\pi(C_2=C_3)$.

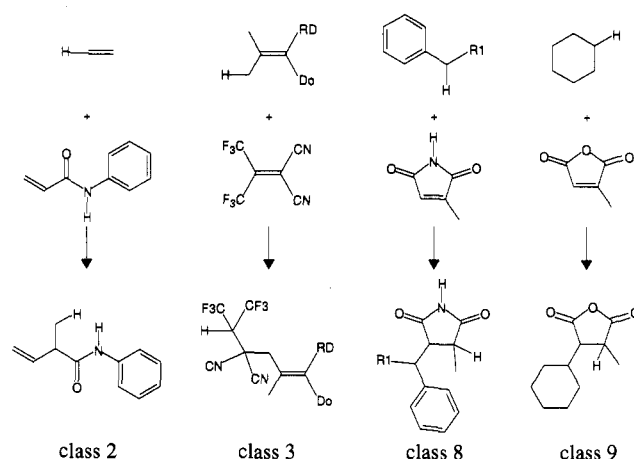


Figure 14. Four classes from the topological classification grouped together by the physicochemical classification.

plot of Figure 13. Thus, a classification based on physicochemical features that works independently of the previous topological classification, taking the individual reactions, produces confirming results. This indicates that the topological features used by HORACE grasp some fundamental essence of these reactions that is also reflected in the electronic variables $\Delta\chi_\sigma(C_1-H)$ and $\Delta\chi_\pi(C_2=C_3)$. However, the electronic variables offer the possibility for greater levels of abstraction by combining classes found to be similar with respect to these variables. This is illustrated with classes 2, 3, 8, and 9. In Figure 13 these four classes are rather close together and can be perceived as forming one single cluster. Indeed, this is supported by chemical evidence showing that all these reactions have several features in common. The reactions in classes 2, 3, 8, and 9 involve the addition of a $C_{sp^2}-H$ or $C_{sp^3}-H$ bond to a double bond bearing electron withdrawing groups at both ends of the olefinic system or have, with the amide group, a substituent that is only slightly electron withdrawing (Figure 14).

Clearly, the reaction in class 9 involving an addition to a cyclic anhydride is closely related to the two reactions in class 8 that involve an addition to the corresponding imide. In fact, the similarity of these reactions is also supported by bibliographic evidence as all three reactions are reported in the same publication.²⁵

The reactions in classes 8 and 9 share a greater degree of similarity with each other than they do with the reactions in

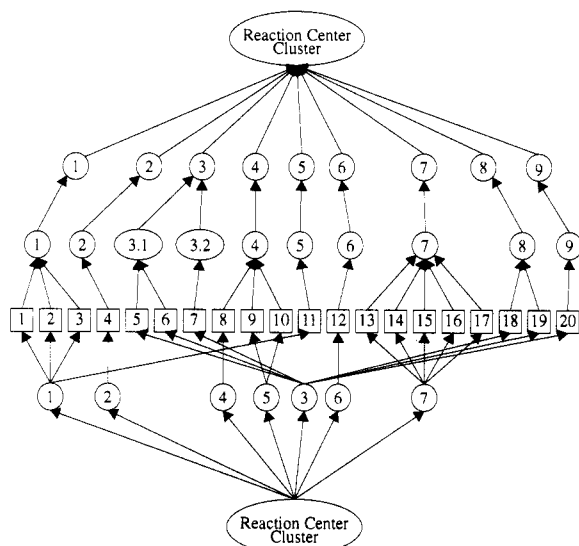
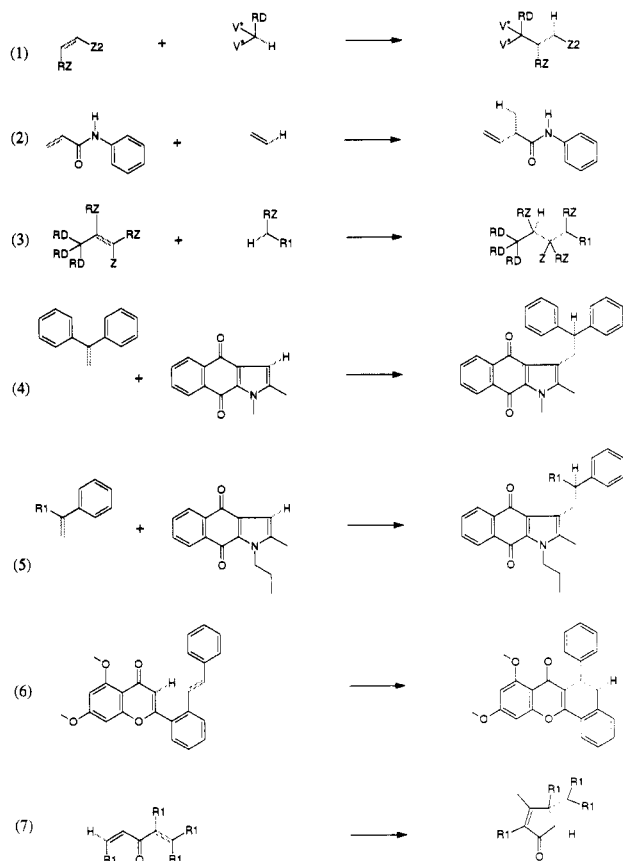


Figure 15. Classification of data set I (20 reactions) by topological features (top) and physicochemical variables (bottom).

Scheme 3. Generalized Classes Produced by Physicochemical Classification



class 3. However, they do show distinct similarities with the reactions in class 3. All three reaction classes have strongly electron withdrawing groups at both ends of the double bond. Furthermore, the reactions of class 3 involve the addition of an allylic CH bond, whereas those of class 8 show the addition of a benzylic CH bond, closely related to allylic CH bonds.

The reaction of class 2 involves the breaking of a C_{sp^2} -H bond, whereas those in classes 3, 8, and 9 all are a result of the breaking of a C_{sp^3} -H bond. This difference is sufficient for separating the reaction in class 2 from those in classes 3, 8, and 9. In other words, it is not possible to adequately describe the differences in this set of reactions through the use of only

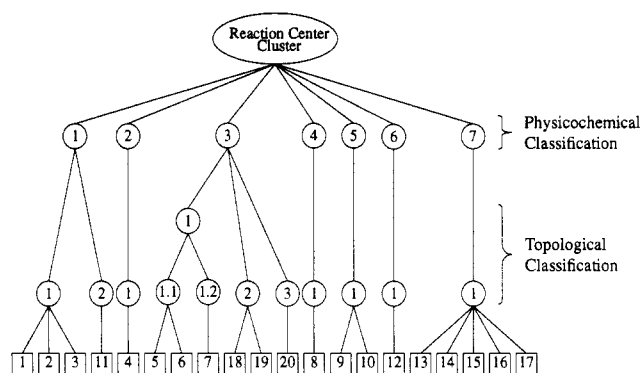


Figure 16. Classification of the 20 reactions by sequential topological and physicochemical classification.

one electronic parameter for each of the two reacting bonds. This underlines the necessity of using a more complex scheme considering more electronic parameters in the classification based on physicochemical variables. In fact, this is done in the full classification as detailed in section 3.2 giving the result presented in the next section.

5.3. Study II: Classification by Physicochemical Variables.

The results of the classification of the data set of 20 reactions based on physicochemical variables is shown in the bottom part of Figure 15 and compared with the previously obtained classification based on topological features which is shown in the top part of Figure 15 (cf. Figure 11). The two classifications agree to a large extent. This indicates that both methods, the classification based on topological features and the one employing physicochemical parameters, are able to extract the major features common to the individual instances of the reactions of a given class. The major difference can be seen in the potential of the physicochemical parameters to provide a higher level of abstraction and thus produce seven classes (see Scheme 3) instead of the nine that were produced on the basis of topological features. In so doing, classes that were originally distinguished in the topological classification are combined to form a common class. This is true for classes 1 and 5 which are merged to form class 1. In fact, all of these reactions are Michael additions, and combining them in a single class is thus warranted. Classes 3, 8, and 9 are also combined to form a single class. This set of reactions has been discussed in the previous section in conjunction with the plot shown in Figure 13. It is gratifying to see that these closely related reactions are combined in a single class and that the reaction of class 2 is excluded from this common class. This shows that the full scheme of weighting by physicochemical variables is able to distinguish the breaking of a C_{sp^2} -H bond from that of a C_{sp^3} -H bond.

Only the former class 4 is split into two new classes. All three reactions involve the addition of a phenyl-substituted olefin to a pyrrole ring. However, the 1,1-diphenylethene is considered to be sufficiently different from styrene and α -methylstyrene in terms of electronic effects to warrant a class of its own. In this case, too much importance is attributed to the difference in the electronic nature of the substituted ethenes, a problem that has to be remedied by further improvement in the weighting of the physicochemical effects.

5.4. Study III: The HORACE Algorithm. The results of the HORACE algorithm which combine the classifications based on physicochemical and topological features are shown in Figure 16. As can be seen, it exhibits the advantages of the two approaches: the intuitive topological similarity and the higher physicochemical abstraction. At the highest level of abstraction is the classification based on physicochemical

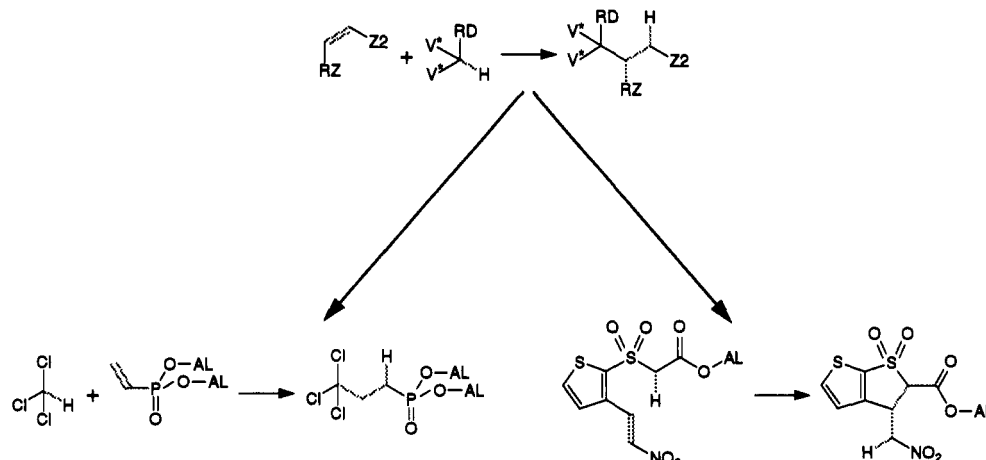


Figure 17. Next lower level of classification of class 1 of Scheme 2 obtained in a topological classification.

variables. Below this level are the two levels of topological classification. By performing the physicochemical classification first and then topologically classifying each of the resulting clusters independently, HORACE is able to propagate physicochemical (dis)similarity constraints into the topological classification. In this way, it is possible to avoid mistakes that could occur in the classifications based on topological features alone (section 8.2).

The fewer classes there are, the more generalization has occurred. This is concurrent with a loss of specific details of the individual reactions. However, the hierarchy offers the user the opportunity to decide which level of abstraction he wants and thus tailor a query to his specific needs. The generalizations of the final 7 classes produced by the HORACE algorithm are the same as those shown in the previous section in Scheme 3. This is a consequence of the fact that HORACE first creates the highest level of classification, the physicochemical level, before creating lower levels.

6. SEARCHING IN A HIERARCHY: 20 REACTIONS

The set of 20 reactions will now be used to illustrate how a search in a hierarchically organized reaction database or hit list resulting from a reaction query can be performed. Class 1 of Scheme 3, resulting from the physicochemical classification of the full HORACE classification, is represented by a reaction equation having highly generalized atom types. In particular, two atoms are of type V^* which is the highest level of atom type generalization (see Figure 2). This is an indication that this reaction class encompasses quite a variety of reaction instances with many different atom types and functional groups around the reaction center.

The potential variety embodied in this class tempts one to further explore the scope of this reaction class. This can be achieved by a topological classification of the four reaction instances grouped into this class by the physicochemical classification. Figure 17 shows the generalized reaction equation of this class on the top level and the two subclasses obtained by the topological classification at the next lower level.

We now see that this reaction class has been split into two subclasses. One consists of three examples for the base catalyzed addition of chloroform to an alkene phosphonate. The other class is comprised of a single reaction, the intramolecular addition of a CH bond activated by two electron withdrawing groups to an electron-poor double bond. Thus, all of the reaction instances of this class are Michael additions, but these examples are from extreme ends of the spectrum of

Michael additions. The first three instances show the reaction of a CH bond activated by three groups (Cl), exerting an electron withdrawing influence by an *inductive* mechanism, a feature that is not very common in Michael additions. The other subclass with a single representative is also quite a remarkable reaction. First, it is an example of an intramolecular Michael addition. Second, whereas the functional groups at the CH bond (CO_2R and SO_2) show common activating effects, it is interesting to note that the double bond is under the influence of *two* electron withdrawing groups: one, the nitro group, is directly bonded, whereas a second one, the SO_2 group, can exert an influence through conjugation of an additional double bond in a thiophene ring.

The analysis of the two subclasses also allows one to understand the generalized atom types. At the double bond, the symbol Z2 represents both the $\text{PO}(\text{OR})_2$ and the NO_2 group, whereas RZ results as a generalized representation for both an H and a C_{sp^2} atom (see Figure 2). At the CH bond, the symbol RD is the common denominator for an H and a Cl atom, whereas of the two V^* atom types one is the common representation of a Cl and an S^{VI} atom (SO_2 group), and the other of a Cl and a C_{sp^2} atom (COOR group).

In summary, an analysis of this highly generalized representation of this reaction class has revealed the broad scope of the Michael addition to us.

7. SEARCHING IN AND UPDATING OF A HIERARCHY: 20 + 1 REACTIONS

One of our aims in developing machine learning methods for the hierarchical classification of reactions is to expand the capabilities for searching in reaction databases. In the introduction we have mentioned two main questions for searching reaction databases: searching for a reaction type corresponding to a query reaction and searching for reactions that are most similar to a query reaction.

The approach taken in searching is simply that of a forest search. It is first necessary to find the appropriate classification tree. This is done by finding the reaction hierarchy with the same reaction center as that of the query reaction. Next a tree search starting from the root and searching in the direction of the leaves is performed. In the case of a hierarchy of reactions, this involves starting at the reaction center level and going down the tree to the desired level of abstraction. The first problem can be addressed by comparing the query reaction to the classes in the top levels of a hierarchical classification. The second problem can be addressed by classifying the query together with the reactions in the

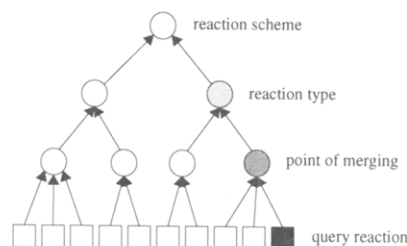


Figure 18. Merging a query reaction into a hierarchical classification of reactions.

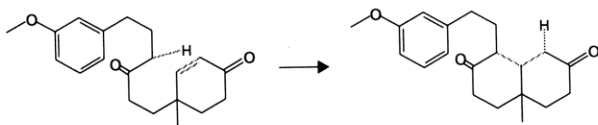


Figure 19. Query reaction of first experiment with the hierarchy of 20 reactions (REACCS Rireg 41900).

database. That point in the hierarchy where the query reaction is merged into a reaction class gives those reactions that are most similar to the query reaction (Figure 18). In fact, the reaction type of the query reaction can be found by going from this point of merging to the top level of the hierarchy (Figure 18).

To illustrate these points and show the performance of a reaction hierarchy with a query reaction, we have used the hierarchical classification produced by the HORACE algorithm from the database of the 20 reactions (section 5.4) and added an additional reaction as a prototype of a query or update reaction. Two different query reactions were used.

Query 1: Figure 19 shows the first reaction used as a query.

Processing of the query begins first by finding the appropriate hierarchy of reactions, that is, by finding the hierarchy with the same reaction center as that of the query reaction. In this case, the reaction center of the query reaction is identical to that of the hierarchy of the 20 reactions produced by the HORACE algorithm (section 5.4). Next, the query reaction is in turn compared with the lower levels of the hierarchy in order to determine its point of merging in the hierarchy.

Starting with the highest level of classification in the hierarchy of 20 reactions, the physicochemical level, the query reaction merges into class 7 of Scheme 3. This is a correct result as the reaction in class 7 and the query reaction all involve the addition of a C–H bond to an α,β -unsaturated ketone. There are noteworthy structural differences in the C–H bond of the query reaction and the reactions in class 7. However, the electronic variable, $\Delta\chi_{\sigma}(\text{C}_1\text{--H})$, which is one of the variables used in the classification, is a measure of the inductive stabilization of a carbanion at carbon atom C_1 and thus brings to light the electronic similarity of those two CH bonds.

Had the query reaction not merged into an existing class then the search would be finished, since that would indicate that the query reaction is unlike any of the other reactions in the hierarchy. Since the query reaction merges into class 7 in the physicochemical classification, it is only necessary to inspect the branch of the hierarchy extending from class 7. Topological classification and generalization of the five reactions of class 7 and the query reaction produce two classes. These consist of the previous existing class, as shown in Scheme 2 (generalized class 7) and an additional class consisting only of the query reaction (Figure 19). Apparently, the reaction in Figure 19 is considered to be substantially different from the other five reactions. Thus, a classification based solely on topological features will not place it into any cluster

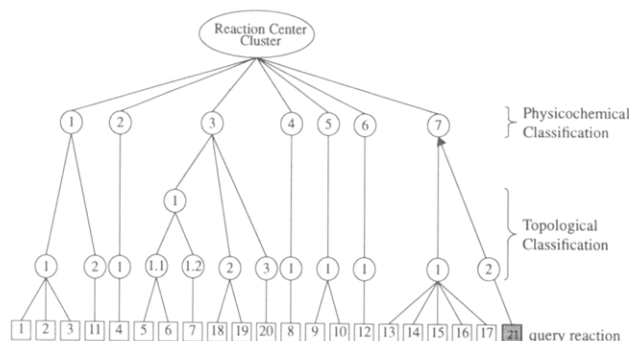


Figure 20. Query reaction merging into the hierarchy during the physicochemical classification.

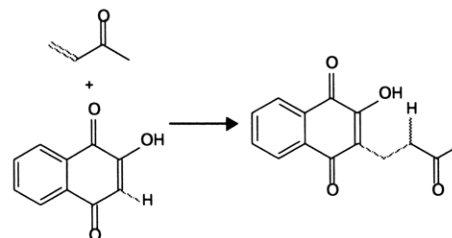


Figure 21. Query reaction of the second experiment with the hierarchy of 20 reactions (REACCS Rireg 24900).

containing other reactions from class 7. It is gratifying that the addition of an extra reaction does not disturb the classification previously obtained with 20 reactions.

Figure 20 shows how the query reaction is merged into the reaction hierarchy already given in Figure 16. Using only topological variables, no similarity is perceived between the query reaction and the other reactions in class 7. In fact, no similarity is perceived to any of the reaction classes of study I. Only at the higher level of abstraction obtained with physicochemical parameters is the similarity of the query reaction to one of the reaction types (class 7) perceived.

It is important to note that it is not necessary to redo the entire classification and generalization with the 21 reactions. Rather, the results produced previously with the 20 reactions can be used, and then it is only necessary to compare the one extra reaction, the query reaction, with this hierarchy. This is an essential feature for the application of HORACE to large data sets of chemical reactions. It would be extremely inefficient to have to recompute the entire hierarchical classification already obtained for the reaction database. Rather, as is done here, it is possible to simply compare the query reaction with the existing hierarchical structure of the previous classification.

In Summary: The query reaction of Figure 19 is perceived as substantially different from the 20 reactions in the database and thus is not merged into any of the classes obtained in the topological classification and generalization. However, similarity is perceived at a higher level of abstraction provided by the electronic parameters chosen for the full HORACE classification.

Query 2: The reaction used in the second experiment as a query is shown in Figure 21.

The search begins by finding the appropriate classification tree which in this case is the same as that of the previous query, the hierarchy produced in section 5.4. The search then proceeds from the physicochemical classification level to the topological levels in order to find the point where the query reaction merges into the hierarchy. Like the first query reaction, this reaction (Figure 21) also merges into class 7 of the physicochemical classification. It is then necessary to

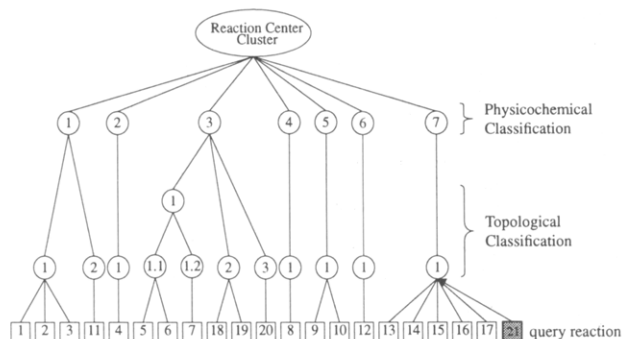


Figure 22. Merging of the query reaction of Figure 21 into the reaction hierarchy of Figure 16.

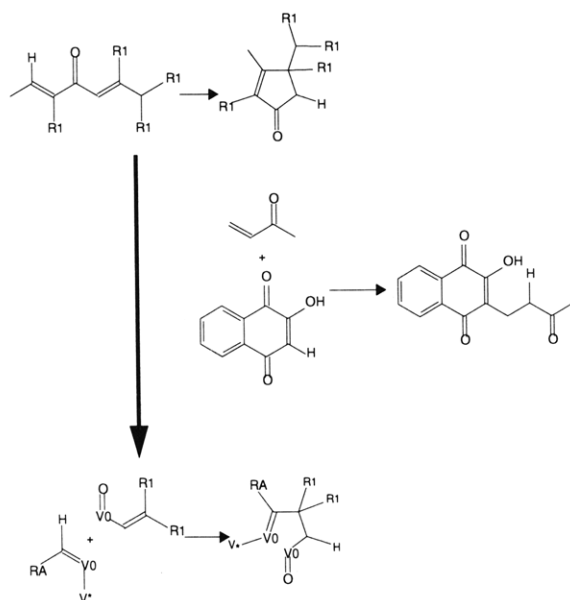


Figure 23. Extension of the scope of a reaction by the addition of a new reaction. This leads to a more generalized representation of the reaction.

examine the rest of this branch so that the degree of similarity of the query reaction to the other reactions in class 7 can be determined.

Topological classification and generalization of the reaction contained in cluster 7 together with the query reaction results in a single cluster (Figure 22). Unlike the previous query, the reaction is found to be similar to the reactions of class 7 in the first level of classification and is therefore merged into this class.

Over and above this perception of similarity in reactions, this experiment also shows that the range of validity of the reactions in class 7 can be extended if the query reaction is added. The previous three reactions in class 7 are intramolecular reactions producing a five-membered ring system. The added reaction (Figure 21) is an example of an intermolecular addition. Thus, while previously the double bond containing the reacting CH bond was directly bonded in the starting material, there is no such bond in the starting materials of the query reaction. This is indicated by the symbol $V0$ in the generalized representation of the reactions in class 7 instead of a carbon atom at the carbonyl group and at the olefinic bond (Figure 23). The two instances of the $V0$ atom type indicate that these are two carbon atoms that are bonded to each other by a carbon chain of varying length. This chain can have either the length of 0 or 1 in this example, indicating intra- or intermolecular reactions. This notation can also be used to describe rings of different sizes.

Additional extensions of the scope of this reaction that are perceived include the following: the carbon atom of the reacting CH bond can either bear a methyl group or a carbonyl group. Figure 2 shows that in the hierarchy of atom types, C_{sp^3} and C_{sp^2} atoms merge at the level of the RA atom type, and therefore this atom type is indicated in the extended generalized form (see Figure 23). It is also found that the carbon atom adjacent to the CH bond can bear either a hydrogen, a C_{sp^2} , or an oxygen atom. The most specific atom type combining both the carbon atom adjacent to the CH bond and a C_{sp^2} is the RA type (cf. Figure 2). However the most specific atom type combining both this carbon atom and an oxygen atom is the RD atom type. In order to accommodate both RA and RD atom types, the atom type is generalized to V^* (see Figure 2 and Figure 23).

In summary: The query reaction of Figure 21 is found to be similar to the reactions in class 7 at level 1 of the hierarchy. Adding this reaction to the hierarchy results in its joining this class at level 1. This indicates a much greater reaction similarity than that found with the query reaction of Figure 19 which when added to the hierarchy, only merges at level 3. Furthermore, adding this reaction to the database increases our knowledge about the range of validity and applicability of the reactions of class 7.

8. DATASET II: 56 REACTIONS

The set of reactions dealt with in the previous section was taken from recent publications abstracted to present results deemed to be novel and interesting. One might expect to have reactions that are somehow biased toward rather *unusual* or at least *novel* reactions.

An additional data set of reactions having the same reaction center as contained in all previous reactions and shown in Figure 10 was investigated for the following reasons:

1. The data set was taken from the Theilheimer database that has been compiled so as to comprise a *representative* selection of reactions for a given reaction type.
2. A slightly more extensive data set was selected to investigate the dependency of the number of classes in the hierarchy on the overall number of reaction instances.
3. It should be investigated whether similar kinds of reaction classes are obtained from reactions extracted from different databases.

8.1. Study I: Topological Classification. The topological classification of these 56 reactions produces a two level hierarchy. The first level consists of 13 classes of which 6 contain subclasses so that there are a total of 29 (sub)classes. The second and final level of the topological classification results in 15 classes (see Figure 24). It may be noted that two reactions (in subcluster 6.3 and subcluster 8.2) are assigned to two classes each.

Thus, although the number of individual reactions has increased from 20 to 56 with respect to the first data set, the number of classes has only increased from 9 to 15, supporting the idea that only a limited number of reaction classes exists for this reaction type (reaction center).

8.2. Study II: Classification on the Basis of Physicochemical Variables. The use of the physicochemical variables in classifying the 56 reactions allows one to reduce the number of classes even further to 10. Thus, in comparison to the data set of 20 reactions, the number of classes has only increased from 7 to 10, supporting the conjecture that the number of reactions for this reaction center might soon reach a limit.

As was done in Figure 15, the classification obtained from the physicochemical parameters is merged in Figure 25 with

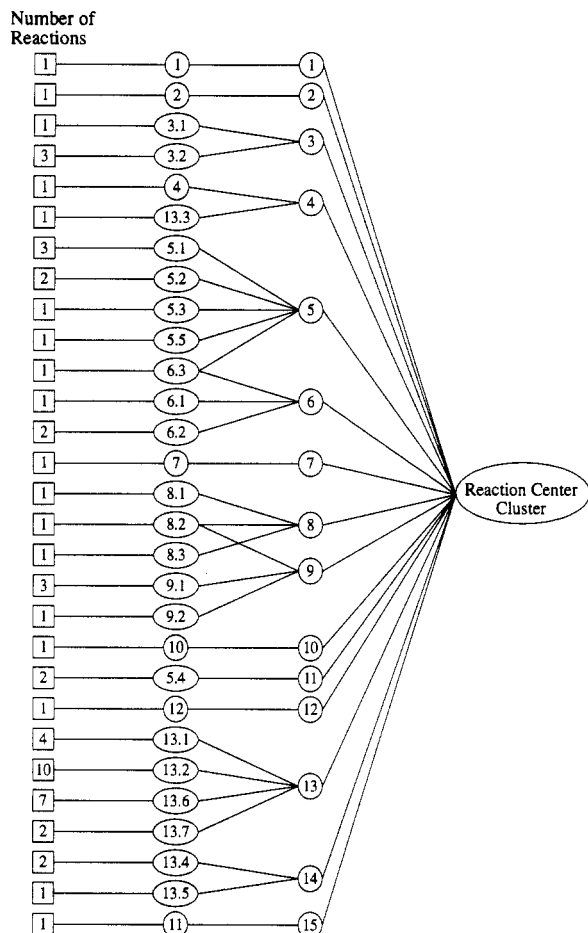


Figure 24. Results of the topological classification of data set II. The numbers in the leaves (boxes) indicate the number of reactions that are combined.

the topological classification. However, in this case the physicochemical classification is added as an extension to the topological classification in order to facilitate comparison.

The two types of classification agree to a large extent. The major difference between the two classifications are two-fold. First, several classes in the topological classification are merged in the physicochemical classification. Second, a few classes of the topological classification are split in the physicochemical classification. The differences in the two classification schemes are quite revealing and are therefore discussed in detail:

1. The classes 3, 8–10 and 13–15 are combined into a single class. This merging is substantiated on the basis of chemical evidence as all of the reactions in the new class 3 are Michael additions.

2. In this merging process four reactions from the 23 reactions of the previous class 13 are separated and put into classes other than class 3; one reaction forms a class of its own (class 9); three reactions are put into class 4. It is gratifying that the physicochemical classification has separated these four reactions from all the others of class 13, as these four reactions are free radical reactions, whereas the rest are Michael additions.

3. The reaction of class 1 is combined with most of the reactions in class 5 to form a new class 1. This is also a correct finding, as all reactions are Friedel–Crafts alkylations of aromatic compounds by olefins. The single reaction of the previous class 1 is a reaction of thiophene, whereas those in the previous class 5 are alkylations of substituted benzenes. Thus, the physicochemical parameters correctly perceive the

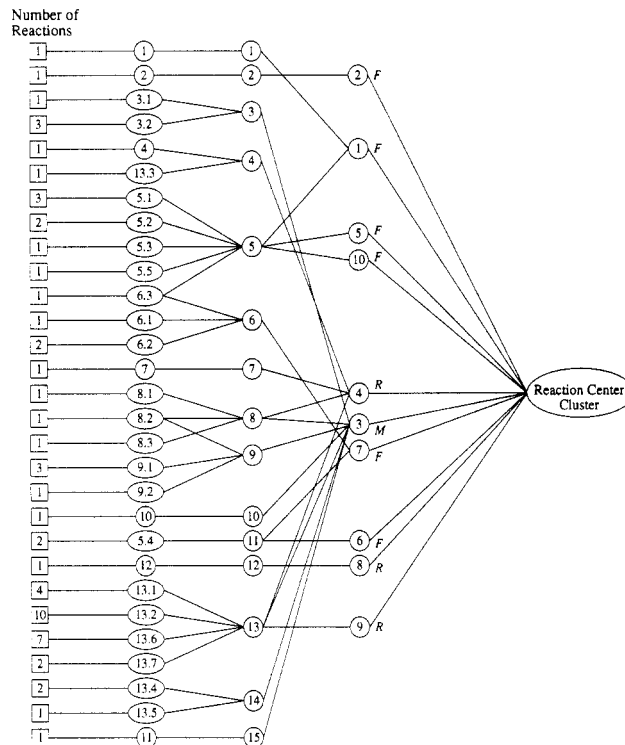


Figure 25. Sequential topological and physicochemical classification of data set II. F indicates Friedel–Crafts reactions, M, Michael additions, and R, free radical reactions.

similarity between thiophene and benzene derivatives in the context of Friedel–Crafts alkylation by olefins.

4. Two reactions of the previous class 5 are each put into a separate class, the new singleton classes 5 and 10. Although these two reactions are also Friedel–Crafts alkylations by olefins and their grouping with the other Friedel–Crafts reactions had been correct, they nevertheless are somehow different from the others. The one reaction forming the new class 5 shows an unusual regioselectivity in the addition reaction. The reaction forming class 10 involves the addition of an α,β -unsaturated ketone and thus extends the scope of the other reactions which all consist of unfunctionalized olefins.

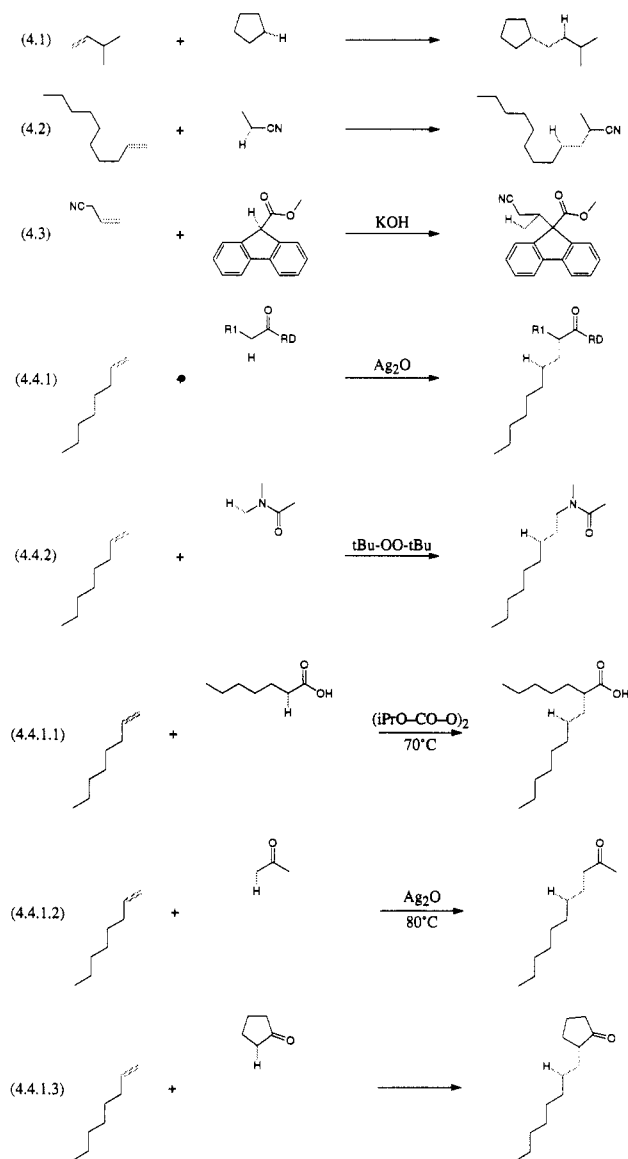
5. The previous classes 4 and 7 as well as one reaction from class 8 and three from class 13 are combined to form the new class 4. This merging of classes is supported by chemical evidence as all reactions in class 4 are free radical reactions.

6. The three reactions in class 8 of the topological classification are separated in the physicochemical classification; two end up in class 3 (Michael additions) and one in class 4 (radical reaction). This is justified on the basis of the types of reactions involved.

In summary, the topological and physicochemical classifications produced by HORACE largely correspond to each other. This is evidence that the selection of classification features and the manner in which they are compared is justified. This is further supported by the fact that the resulting classifications are largely correct with respect to the reaction classes extracted. However, the physicochemical classification leads to a higher level of abstraction and is able to find a classification that even better corresponds to chemical evidence, remedying some deficiencies of the topological classification.

8.3. Study III: Classification by the HORACE Algorithm. The way in which HORACE combines the topological and physicochemical classification methods produces a synergistic effect. By propagating physicochemical constraints into the topological portion of the classification, the full HORACE

Scheme 4. Reaction Instances of Class 4



CE, an automatic system for hierarchically classifying reactions, combines topological and physicochemical classification methods to produce hierarchies expressing a broad range of abstraction of similarity. The HORACE algorithm is conducive to *lazy evaluation*, thus potentially providing a significant reduction in the amount of computation involved for hierarchy searches and updates.

ACKNOWLEDGMENT

We appreciate the discussions on reaction classification we had in a task force organized under the auspices of the German Federal Ministry for Research and Technology. These meetings confirmed our views on hierarchical classification and forced us to define various features of our approach more clearly. We would like to thank Dr. C. Weiske and Dr. Axel Parlow of Fachinformationszentrum Chemie, Berlin, for providing us with the 20 reactions of data set I. We are also indebted to Prof. Herbert Gelernter for having made the 56 reactions in data set II available to us. Finally, we would like

to thank the Alexander von Humboldt Foundation for having made this international and interdisciplinary collaboration possible.

REFERENCES AND NOTES

- (1) Blake, J. E.; Dana, R. C. CASREACT: More than a Million Reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 394–399.
- (2) ChemReact, available from InfoChem, Munich.
- (3) Parlow, A.; Weiske, C.; Gasteiger, J. ChemInform—An Integrated Information System on Chemical Reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 400–402.
- (4) Gasteiger, J.; Weiske, C. ChemInform—Vom gedruckten Referatedienst zur Reaktionsdatenbank (ChemInform—From a printed abstracting service to a reaction database). *Nachr. Chem. Tech. Lab.* **1992**, *40*, 1114–1120.
- (5) ChemInform-RX: Produced since 1991 by Fachinformationszentrum Chemie, Berlin from the information contained in the weekly abstracting service ChemInform, marketed by Molecular Design Ltd. (MDL).
- (6) Röse, P.; Gasteiger, J. Automated Derivation of Reaction Rules for the EROS 6.0 System for Reaction Prediction. *Anal. Chem. Acta* **1990**, *235*, 163–168.
- (7) Gelernter, H.; Rose, J. R.; Chen, C. Building and Refining a Knowledge Base for Synthetic Organic Chemistry via the Methodology of Inductive and Deductive Machine Learning. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 492–504.
- (8) Rose, J. R.; Gelernter, H. ISOLDE: A System for Learning Organic Chemistry through Induction. In *EKAW-89: Third European Workshop on Knowledge Acquisition for Knowledge-Based Systems*, Paris, France, 1989; Boose, J. H., Gaines, B. R., and Ganascia, J. G., Eds.; European Coordinating Committee for Artificial Intelligence: Paris, 1989.
- (9) Rose, J. R.; Gelernter, H. Unsupervised Learning of Context-Sensitive Graph Rewriting Rules. To be published in *Machine Learning*.
- (10) Hendrickson, J. B.; Miller, T. M. Reaction Indexing for Reaction Databases. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 403–408.
- (11) Weise, A. Synthesis Simulation by Synthon Substitution. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 490–491.
- (12) Blurock, E. S. Computer-Aided Synthesis Design at RISC-Linz: Automatic Extraction and Use of Reaction Classes. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 505–510.
- (13) Funatsu, K.; Sasaki, S. Computer-Assisted Synthesis Design and Reaction Prediction System AIPHOS. *Tetrahedron Comput. Methodol.* **1988**, *1*, 27–38.
- (14) Gasteiger, J.; Röse, P.; Hondelmann, U.; Witzelbichler, W. Computer-Assisted Degradation of Chemicals: Hydrolysis of Amides and Benzoylphenylureas. Manuscript in preparation.
- (15) Jorgenson, W. L.; Laird, E. R.; Gushurst, A. J.; Fleischer, J. M.; Goethe, S. A.; Helson, H. E.; Paderes, G. D.; Sinclair, S. CAMEO: a program for the logical prediction of the products of organic reactions. *Pure Appl. Chem.* **1990**, *62*, 1921–1932.
- (16) Gelernter, H.; Miller, G. A.; Larsen, D. L.; Berndt, D. J. Realization of a large expert problem-solving system: SYNCHEM2, a case study. *IEEE 1984 Proceedings of the First Conference on Artificial Intelligence Applications*; IEEE Computer Society Press: Silver Spring, MD, 1984.
- (17) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity—A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219–3228.
- (18) Gasteiger, J.; Saller, H. Calculation of the Charge Distribution in Conjugated Systems by a Quantification of the Resonance Concept. *Angew. Chem.* **1985**, *97*, 699–701; *Angew. Chem., Int. Ed. Engl.* **1985**, *24*, 687–689.
- (19) 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.
- (20) Gasteiger, J.; Hutchings, M. G. Quantification of Effective Polarizability. Applications to Studies of X-Ray Photoelectron Spectroscopy and Alkylamine Protonation. *J. Chem. Soc., Perkin Trans.* **1984**, *2*, 559–564.
- (21) Gasteiger, J.; Saller, H.; Löw, P. Elucidating Chemical Reactivity by Pattern Recognition Methods. *Anal. Chim. Acta* **1986**, *191*, 111–123.
- (22) Gasteiger, J.; Marsili, M.; Hutchings, M. G.; Saller, H.; Löw, P.; Röse, P.; Rafeiner, K. Models for the Representation of Knowledge about Chemical Reactions. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 467–476.
- (23) Gasteiger, J.; Weiske, C. *Proceedings of the 13th International Online Information Meeting*; Learned Information: Oxford, U.K., 1989; pp 147–154.
- (24) Gasteiger, J.; Hutchings, M. G.; Christoph, B.; Gann, L.; Hiller, C.; Löw, P.; Marsili, M.; Saller, M.; 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.
- (25) Giese, B.; Farshchi, M.; Hartmanns, J.; Metzger, J. O. Isoselectivity Correlation for the Stereoselectivity of the Hydrogen Atom Transfer to Cyclic Alkyl Radicals. *Angew. Chem.* **1991**, *103*, 619–620; *Angew. Chem., Int. Ed. Engl.* **1991**, *30*, 600–601.