

## Similarity Concepts for the Planning of Organic Reactions and Syntheses

Johann Gasteiger,\* Wolf-Dietrich Ihlenfeldt, Ralf Fick, and John Royce Rose

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

Received July 16, 1992

The similarity of chemical structures has been defined by generalized reaction types and by gross structural features. These definitions correspond to structural transformations that can be made for an entire dataset of structures prior to a search query. This allows an efficient and rapid search of databases of structures. The similarity of reactions is defined by physicochemical parameters of the atoms and bonds at the reaction site. These definitions provide criteria for searching databases of reactions and for drawing conclusions on reaction conditions. Applications for the organization of databases of structures, for reaction prediction, for reaction-planning, for synthesis design, and for the automatic acquisition of knowledge about chemical reactions are shown.

### INTRODUCTION

In the absence of clear mathematical equations for the prediction of properties of compounds or reactions, the chemist draws many conclusions by analogy from known information. To be able to choose the appropriate information for drawing analogy inferences, one must be able to perceive some sort of similarity between the desired and the selected information.

In this sense, the definition of similarity is heavily dependent on the kind of property that one wants to predict; the bioactivity of a compound asks for a similarity measure that is different from the one chosen for estimating the boiling point or the chemical reactivity of a compound.<sup>1,2</sup> This is underscored by the wide variety of similarity measures in the other papers in this issue.

The similarity measures explored in this paper have all been developed for applying them to chemical reactions. Some can be used for analyzing the relationships between the starting materials or products of a reaction or a sequence of reactions; others directly center on the changes in the bond and electron distribution occurring in a chemical reaction.

The first types of similarity measures can be applied to searches in databases of compounds; whereas, the second type is applicable to searches in databases of chemical reactions.

They can be used in the following problem areas

- synthesis design
- reaction prediction
- reaction planning
- acquisition of knowledge on reactions

The concept of similarity has gained much interest in recent years, strongly stimulated by the increased availability of databases on chemical compounds and reactions. However, this area of research is still in its infancy; a lot of research has yet to be done. Similarity measures are always highly subjective concepts, their importance and merits being measured by the extent they allow one to draw conclusions of analogy and lead to new insights. To assist in the development of the field, we have developed a series of similarity measures in order to explore their merits in information processing. In addition, we have chosen such measures that make it clear to the user why certain structures or reactions are perceived as similar and give them control over the concepts they want to use to define similarity.

### SIMILARITY BASED ON STRUCTURE OF MOLECULES

Several approaches have already been developed for defining the similarity of molecules based on constitutional information

as contained in a connection table.<sup>1,2</sup> The two structures to be compared are broken into fragments, and the fragments of each molecule are matched with each other. The number of fragments, *C*, that are common to both structures is compared with the number of fragments, *A* and *B*, in the individual two structures. The Tanimoto index, *T*, is taken as a quantitative measure of similarity:

$$T = C / (A + B - C)$$

Clearly, this similarity definition is strongly dependent on the set of fragments chosen. Quite often the fragments are the same as those selected as screens for substructure searching. As these differ from system to system, quite different similarity rankings are obtained by the various systems.

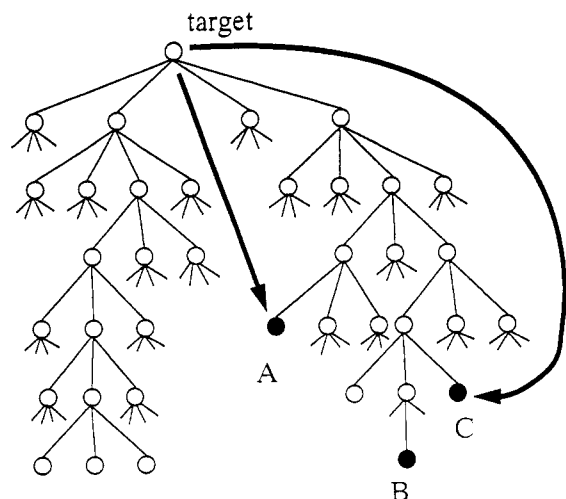
Such similarity measures may find new structural relationships and provide food for thought. However, they give no information on how to explain such similarity.

Our interest focused on defining similarity measures that can be used for reaction and synthesis planning.<sup>3</sup> The design of organic syntheses looks to find relationships between a target molecule and available starting materials. The recognition of structural similarity between the target and some starting materials can lead to strategies that greatly simplify the search space of a synthesis and find a rather direct route from those starting materials to the desired target compound (Figure 1).<sup>4-6</sup>

Our aim was not only to recognize those similarities that can be used in the efficient planning of syntheses but also to give hints as to how a perceived similarity can be utilized for a synthesis (which reactions or sequence of reactions to choose to convert a starting material into the target).

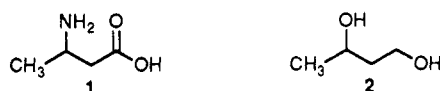
To give an example, let us assume that we are interested in a synthesis of 3-aminobutyric acid, **1**. Let us further assume that we have 1,3-butanediol, **2**, in our depository of available compounds (Scheme I).

After some consideration, a chemist will conclude that 1,3-butanediol might be a reasonable starting material for the synthesis of 3-aminobutyric acid. The conversion of 1,3-butanediol into 3-aminobutyric acid would need a selective oxidation of the primary alcohol to the carboxylic acid, followed by a replacement of the remaining OH groups by an amino group. Neither conversion is straightforward, needing protecting or activating groups (e.g., as tosylate), respectively. But once a chemist has perceived the similarity of 1,3-butanediol to 3-aminobutyric acid, he can draw from his rich experience in reaction control to achieve this conversion.



**Figure 1.** Perception of structural similarity can lead to efficient synthesis strategies and greatly limit the number of reactions to be explored. A, B, and C are available compounds.

**Scheme I**



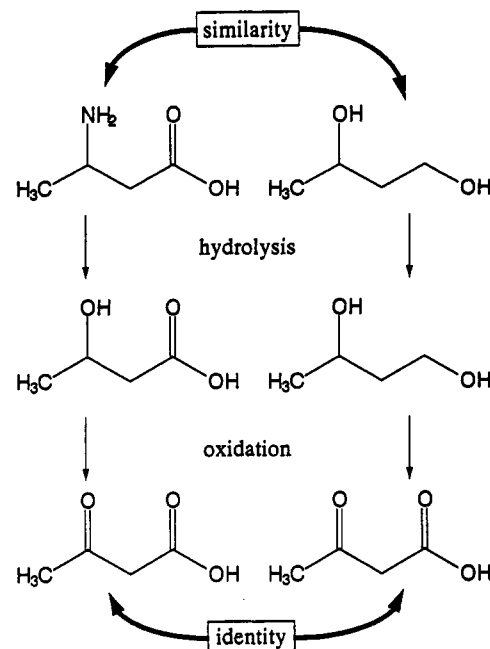
How can the computer be told to recognize the relationship between 3-aminobutyric acid and 1,3-butanediol? A substructure search with a four-carbon unit and a variable number of substituents at positions 1 and 3 of both carbon skeletons would be one answer to the problem. However, such a substructure search would not show how to achieve the conversion of 2 to 1 (which reactions or types of reactions to choose to make this transformation).

Our approach centers directly on reactions that can achieve such conversions. In effect, we submit chemical structures to generalized reaction types, thus, converting them to modified structures that are then compared. For example, one such sequence of events is a hydrolysis followed by an oxidation. Applying this sequence to 3-aminobutyric acid and 1,3-butanediol gives in both cases acetoacetic acid (Figure 2).

The very fact that 3-aminobutyric acid and 1,3-butanediol both lead to the same structure as perceived by our hash-coding algorithm<sup>4</sup> is taken as evidence that the initial structures, 3-aminobutyric acid and 1,3-butanediol, have something in common and are to be taken as similar. In fact, we can even specify why they are similar. They are similar because 3-aminobutyric acid can be converted into 1,3-butanediol by a sequence of hydrolysis and redox reactions.

Thus, advice is obtained on how to convert one structure into the other structure that has been perceived as similar. This perception of similarity based on a reaction type can be directly translated into a sequence of reactions and, thus, into a strategy for synthesis planning.

**Similarity Measures Based on Generalized Reactions.** To have a broad scope of similarity measures useful for synthesis design, a series of transformations based on general reaction types was defined and implemented. In generalizing reactions, we wanted to put reactions together that proceed under similar reaction conditions and use similar types of reagents. Substitution, elimination, reduction, oxidation, and hydrolysis reactions are such general reaction types. The following list contains the transformations defined by general types of reaction that are taken as a basis for similarity definitions:



**Figure 2.** Perception of structural similarity by a sequence of structural transformations followed by an identity match.

1. Exchange of heteroelements of a group in the periodic system (e.g., halogen atoms). Clearly this corresponds to a substitution reaction.
2. Exchange of heteroelements (C-X) and of heteroelements including hydrogen atoms (C-YH<sub>n</sub>) that are not in the same group of the periodic system. These, too, correspond to substitution reactions, but on a wider scope.
3. Oxidation of carbon atoms carrying heteroatoms.
4. Reduction, i.e., hydrogenation by cleavage of hetero-heterobonds, and saturation of multiple bonds and aromatic rings with hydrogen atoms.
5. Reduction, but not of aromatic rings.
6. Elimination of H-Hal, H<sub>2</sub>O, NH<sub>3</sub>, etc. from adjacent carbon atoms.
7. Elimination followed by reduction.
8. Hydrolysis of acetals, esters, heterosubstituents at multiple bonds, etc.
9. Hydrolysis, but not if substituent is on an aromatic ring. This feature takes into account the fact that substituents on an aromatic ring are harder to exchange by nucleophilic substitution.
10. Hydrolysis, but not if substituent is on a benzenoid aromatic ring. Substituents on heteroaromatic rings, on the other hand, are hydrolyzed. This option recognizes that a nucleophilic substitution is particularly difficult on a benzenoid aromatic ring.
11. Hydrolysis followed by oxidation.

This list highlights our intention of defining similarity while simultaneously giving information on which types of reactions (substitution, elimination, hydrolysis, oxidation, reduction) can be used to interconvert similar compounds. As these similarity measures are intended for strategic considerations, no details on the reactions or on reagents to be used are incorporated. We intentionally leave it to the user to decide which reagent to use to perform an oxidation, for example.

However, some general observations on the feasibility of reactions are included in the definition of the reaction types used in the transformations. Thus, the reduction transfor-

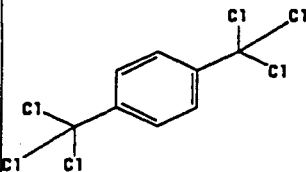
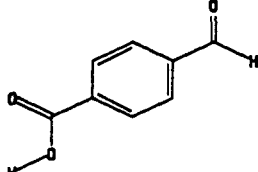
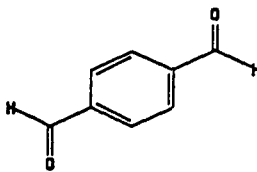
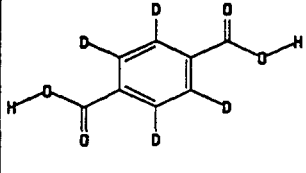
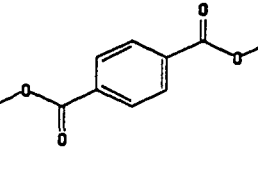
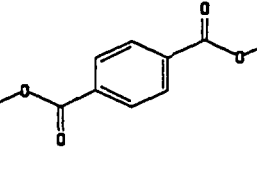
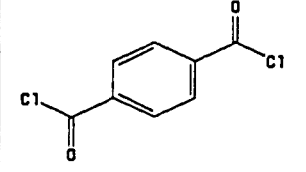
 <p>1146 JANSEN 1541791</p> <p>ALPHA,ALPHA,ALPHA,ALPHA',ALPHA</p>	 <p>1701 JANSEN 1545835</p> <p>4-CARBOXYBENZALDEHYD 97%</p>	 <p>7476 JANSEN 1378915</p> <p>TEREPHTHALDIALDEHYD 98%</p>
 <p>7477 JANSEN 1875231</p> <p>TEREPHTHALSAEURE-D4 98 ATOM %</p>	 <p>7478 JANSEN 1807230</p> <p>TEREPHTHALSAEURE 98%</p>	 <p>7479 JANSEN 1379016</p> <p>1379016 100-21-0</p>
 <p>7480 JANSEN 1531081</p> <p>TEREPHTHALSAEUREDICHLORID 97%</p>		

Figure 3. Results of a search in the Janssen Chimica Catalog with terephthalic acid as the query using the combined hydrolysis/oxidation transformation.

mation saturates multiple bonds with hydrogen atoms but also cleaves hetero-hetero bonds. This takes into account the fact that hetero-hetero bonds are often cleaved in reduction reactions.

Another point to be noted is our inclusion of the different implementations of the hydrolysis reaction. The exchange of substituents on the carbon skeleton (e.g., of an  $\text{NH}_2$  group for an OH group) may be allowed in all cases, not allowed if the carbon atom is part of any aromatic system, or not allowed if it is part of a benzenoid aromatic system. This takes into account that nucleophilic exchange is more difficult if a group is on an aromatic ring, and more so if the ring is terephthalic.

As an example, the results of a search with terephthalic acid in the Janssen Chimica Catalog (version 1987) using the combined hydrolysis and oxidation similarity transformation is given in Figure 3.

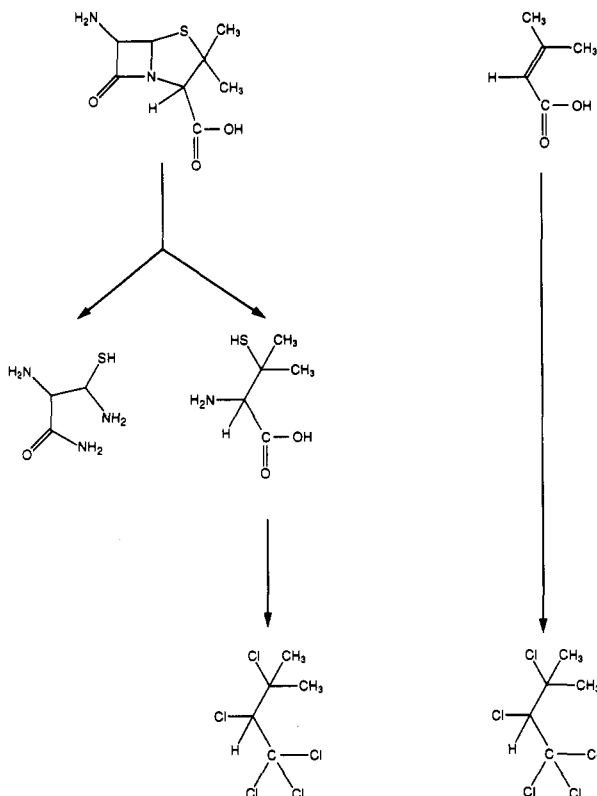
The printout shows that terephthalic acid is available in two different commercial purities and in the deuterated form. The other compounds can be converted into terephthalic acid by hydrolysis or oxidation reactions.

**Similarity Measures Based on Substructures.** Another set of transformations isolates a certain substructure from a molecule. If two molecules have the same substructure, they are considered to be similar. In defining these transformations we again had synthesis design as an application in mind. Thus, transformations to substructures that are of importance in a synthesis were conceived. Because of the importance of the construction of the carbon skeleton in an organic synthesis, most of these transformations will not break a carbon-carbon bond but will maintain the carbon skeleton. The following list contains some of the transformations that isolate a certain substructure from the molecule investigated.

1. Identity match with or without consideration of stereochemistry
2. Carbon skeleton
3. Ring system only
4. Ring system and carbon skeleton of substituents
5. Ring system with substitution pattern
6. Aromatic ring including  $\alpha$ -atoms
7. Carbon skeleton and  $\alpha$ -heteroatoms
8. Carbon skeleton and  $\alpha$ -heteroatoms including stereochemistry
9. Carbon skeleton and substitution pattern.  
This transformation has several variants:  
(a) multiple bonds are considered as substituents or not  
(b) the bond order to substituents is counted or not  
(c) multiple substituents are counted multiply or only once
10. Carbon skeleton and aromatic rings with  $\alpha$ -heteroatoms with and without consideration of stereochemistry

A skeleton is defined as an arrangement of carbon atoms that are directly bonded to each other. A molecule is cleaved at heteroatoms. Some of the similarity criteria differ only in whether or not they include the heteroatom in the transformed product. In cases where the molecule contains several fragments, only the one with the highest number of atoms is selected.

When two molecules are compared for similarity, both structures are modified according to the rules inherent in the transformation in question. A substructure similarity



**Figure 4.** Transformation of 6-aminopenicillanic acid (APA) and of 3,3-dimethylacrylic acid by the criterion that searches for the carbon skeleton and its substitution pattern. The double bond is here considered as the functionalization, hence a substituent. (The appearance of the chlorine atoms is explained in the text.)

search does not entail a substructure search. Rather the structures are changed by the transformation corresponding to the similarity criterion, and then the structures are compared for identity.

Figure 4 shows the changes that are made with 6-aminopenicillanic acid (APA) and with 3,3-dimethylacrylic acid when the similarity transformation that searches for the carbon skeleton and its substitution pattern is used. First, the carbon heteroatom bonds are broken, and the heteroatom is assigned to each carbon fragment. APA gives two fragments. The fragment with the higher number of atoms is processed further. Then, all heteroatoms are replaced by chlorine atoms; heteroatoms that are double-bonded to carbon are replaced by two chlorine atoms. Furthermore, multiple bonds are saturated by chlorine atoms.

In this manner, both APA and 3,3-dimethylacrylic acid lead to the same transformed structure, 2-methyl-2,3,4,4-tetrachlorobutane. This supports the conclusion that both structures are similar. In fact, it is quite conceivable that dimethylacrylic acid could serve as a precursor in a synthesis of 6-aminopenicillanic acid. Epoxidation followed by two successive nucleophilic substitutions with a sulfur and a nitrogen nucleophile would lead to 2-amino-3-methyl-3-mercaptopentanoic acid.

Figure 5 shows part of the output list that was obtained in a search of the Janssen Chimica Catalog for compounds that have the same ring skeleton and substitution pattern as norbornyl bromide. In this search multiple bonds were considered not to be substituents but treated as though they were single bonds.

Note that the nature of the substituent is not considered. The isolation of the ring system is one of the few cases where a C-C bond is broken. The stereochemistry at the substitution site is also not taken into account for this similarity criterion.

## IMPLEMENTATION

The similarity criteria presented in the previous sections are particularly valuable for searching databases of compounds. One important application is their use in synthesis design, when appropriate starting materials for the synthesis of the target structure or the precursor molecules are sought in a database of available compounds.

The development of a hash-coding algorithm<sup>4</sup> that can compress the entire description of the structure of a compound into a 32-bit integer, one computer word, made it possible for searches for similar structures to be very efficient and rapid.

With such a concise structure representation as given by a 32-bit integer, one can store not only the structures of a dataset of compounds but also the structures that are generated from them by the various transformations. Thus, the various transformations of the entire dataset of structures can be performed in advance and the transformed structures stored in the form of hash-codes together with the initial structures (Figure 6).

In this way, a database can be prepared for a rapid similarity search. Then, only the query structure (target or synthesis precursor) has to be transformed according to the rules of a specific transformation; the structure thus obtained is hash-coded, and the corresponding column of the matrix of hash-codes of initial and transformed structures can be scanned (Figure 7).

As only the query structure transformation, hashing and comparison of 32-bit integer numbers have to be done; such a similarity search is quite rapid. This is particularly true for small or medium-sized datasets. In such cases, the entire dataset of transformed structures can be kept in computer memory, and the search can be performed directly in memory.

In our implementation we have three datasets: 7849 structures from the Janssen Chimica Catalog, 3561 structures from the Merck-Schuchardt Catalog, and 2211 structures from the CHIRON<sup>7</sup> database. Even from the largest database, the Janssen Catalog, the hash-codes of several transformations of the entire dataset can be simultaneously stored in memory.

These methods for similarity searching have been integrated into our WODCA system (Workbench for the Organization of Data for Chemical Applications).<sup>3,8</sup>

The perception of structural similarity is intended to stimulate analog reasoning in the chemist. It should give him new ideas for solving given problems, in particular, for the design of syntheses. To assist in this process, the structures that result in a search are simultaneously graphically displayed on the screen. By showing several structures simultaneously, the associative power of the human mind is stimulated.

In addition, while giving the hit list of structures in one window, other information, e.g., availability and price of the compounds, can be shown in another window. In Figure 8 a typical example of a screen during a similarity search is shown.

The query structure,  $\beta$ -bromopropionic acid, is shown at the lower left-hand corner of the screen. The list of structures that is obtained with this query structure by the transformation that allows an elimination and a reduction to occur covers most of the screen. The pointer on L-(+)-lactic acid (Milchsäure) asks for more information on this compound; this information is given in the window at the lower right-hand corner.

## APPLICATIONS OF SIMILARITY OF STRUCTURES

In the following, a series of examples for different types of applications of the similarity concepts is given. All similarity

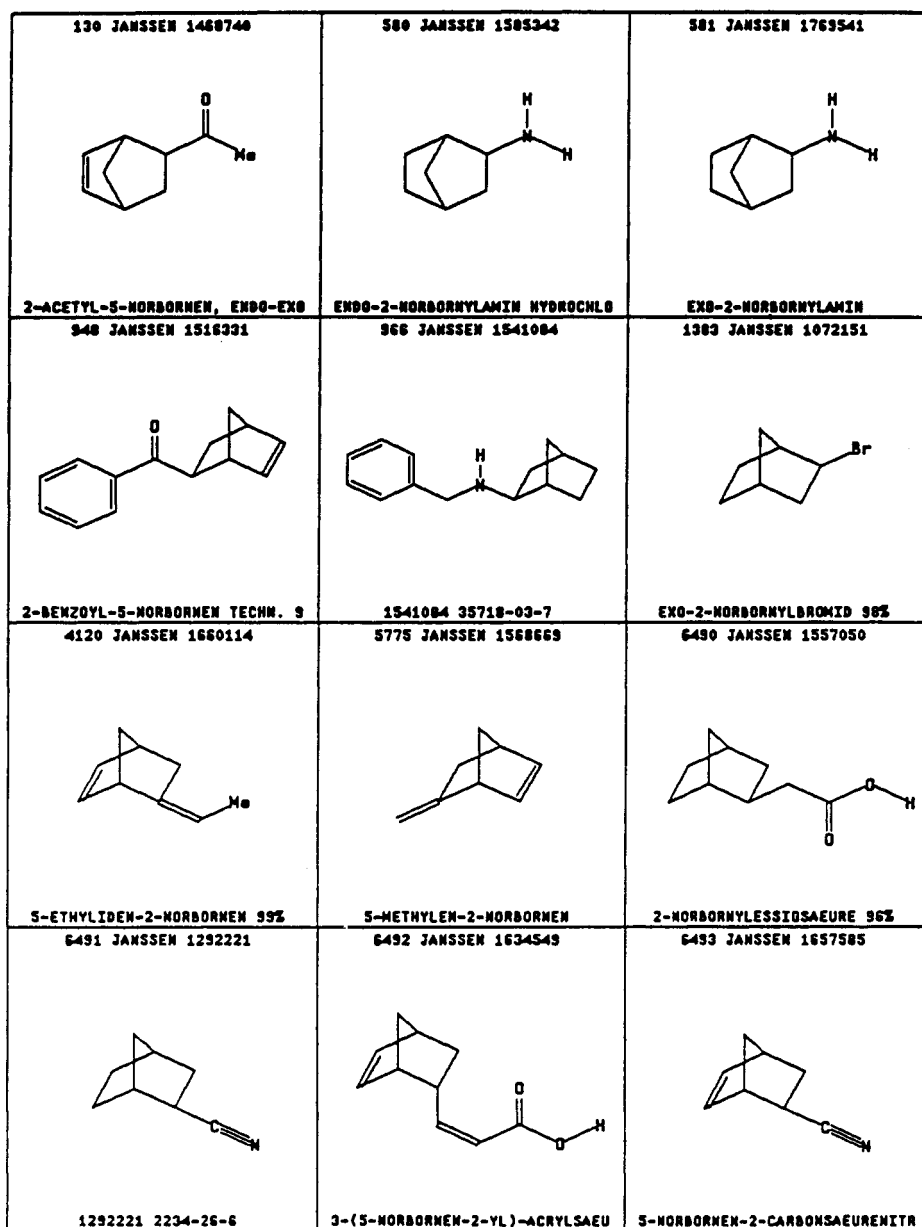


Figure 5. Part of the list of the structures that have the same (norbornyl) skeleton and substitution pattern.

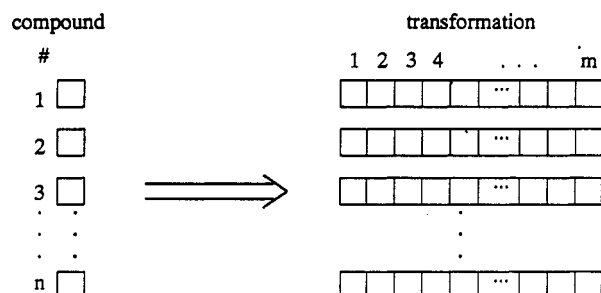


Figure 6. Preparation of a database of structures for similarity searches by prior transformation and storing of the hash-codes of the transformed structures.

searches were performed on the Janssen Chimica Catalog containing the 7849 structures of available compounds.

It should be kept in mind that the same hit list of structures in each of these similarity searches would be obtained from any one of the individual structures as from the query structure. All the structures are classed as similar because they are transformed into the same base structure (Figure 9). Thus, an entry to this manifold of structures is obtained from any one of the structures contained therein. The base structure

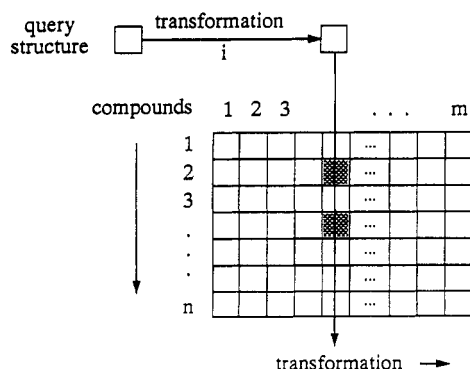


Figure 7. Strategy for a similarity search; only the query structure has to be transformed and hash-coded.

generated by the transformation may or may not be part of the hit list.

**Compression of Databases of Structures.** Similarity criteria can be used to organize databases of structures. Structures that are similar contain, to a certain extent, the same kind of information. Only those structures that are dissimilar to all the other structures in a database bring essential new

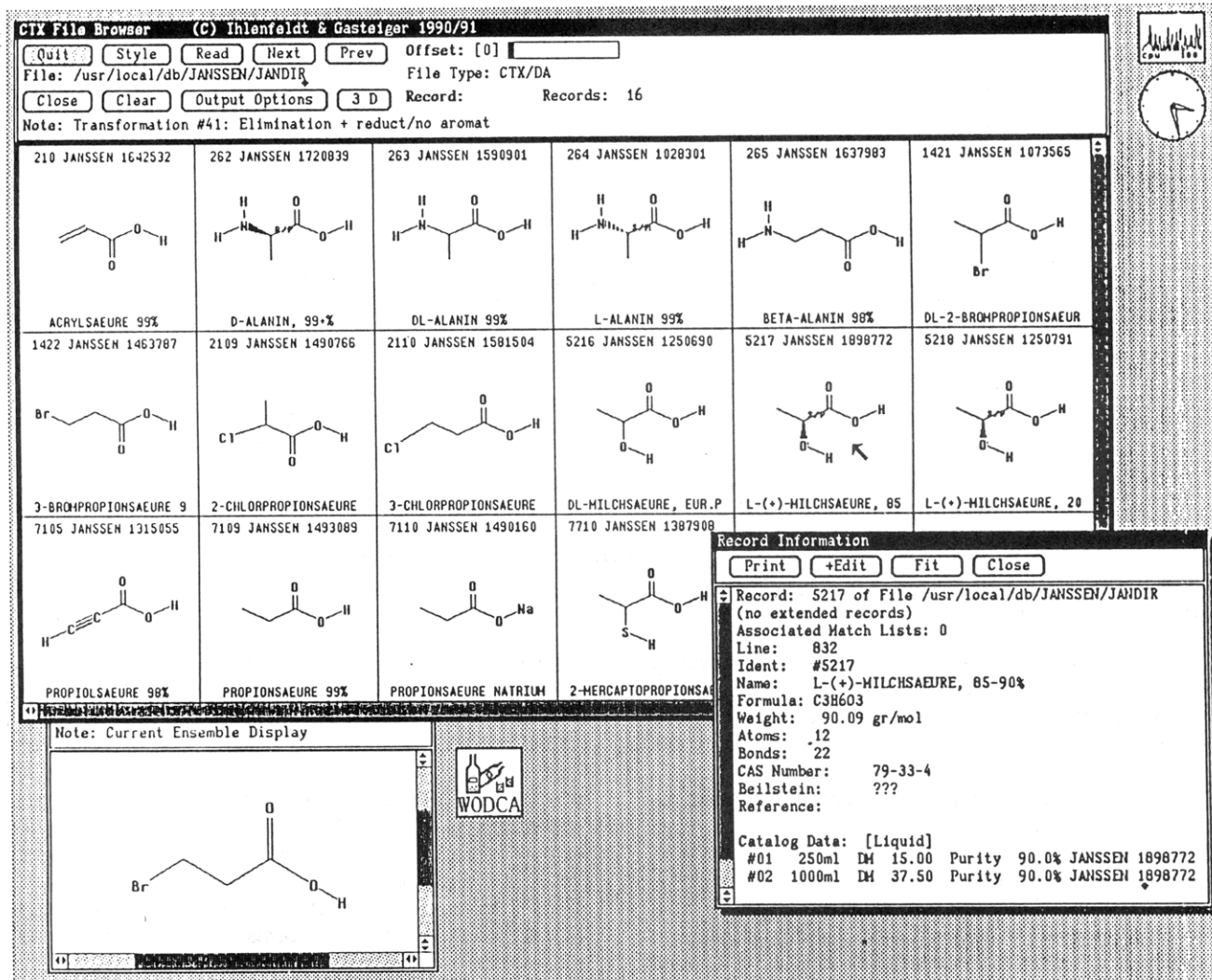


Figure 8. Screen during a similarity search showing various windows with supportive information.

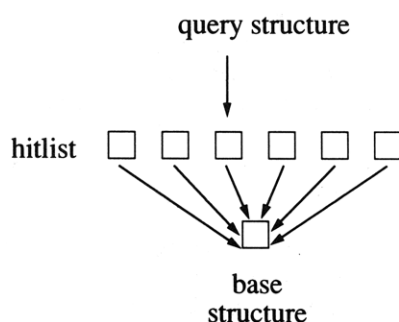


Figure 9. Same hit list of structures will be obtained from any one of these structures taken as a query because all are transformed into the same base structure.

information. Thus, one could reduce a database of structures to one member of each class of similar structures and only add a new structure if it is not similar to any of the structures already contained in the database.

We have explored the extent to which various similarity criteria would support the reduction in size of a database of structures. All numbers reported here have been derived from searches in the Janssen Chimica database that contained 7849 structures. Table I gives the number of different structures obtained with the mentioned similarity transformation together with the percentage in the reduction in the number of molecules.

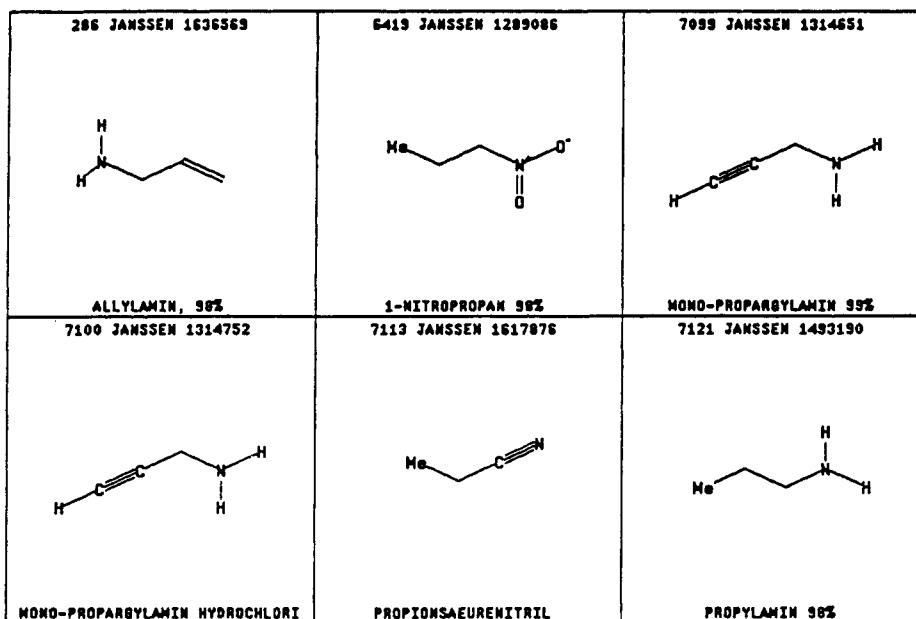
The reduction in the number of molecules is an indication of the degree of generalization inherent in a transformation.

Table I. Reduction in Number of Molecules in a Database of Structures (7849) by Various Transformations

transformation	no. of structures	reduction, %
identity (none)	7849	0
element exchange	7095	10
element and XH exchange	6088	22
oxidation	6748	14
reduction	6193	21
elimination	7237	8
hydrolysis	3972	49
hydrolysis followed by oxidation	3320	58
carbon skeleton	906	89
carbon skeleton and $\alpha$ -heteroatoms	4708	40
rings and carbon skeleton	1598	80

Whereas the exchange of elements within a group of the periodic system does not greatly reduce the number of structures, the number of different carbon skeletons in the database only amounts to 11% of the entire dataset.

**Reaction Prediction.** The similarity measures reported here have been primarily designed for the perception of structural similarity that can be exploited for the design of organic syntheses. However, because many of these criteria use transformations based on general reaction types, these similarity concepts can also be applied in problems where the products of reactions are of interest. These transformations have no built-in evaluations of chemical reactions and cannot, therefore, really predict the products of chemical reactions like the reaction prediction system EROS.<sup>9-12</sup> However, they



**Figure 10.** Similar structures based on the maximum reduction transformation applied to a compound consisting of a chain of three carbon atoms and a nitrogen atom.

can give information on what products might be expected. Thus, these methods can be used for the rapid scanning of databases, can help in the analysis of analytical data, or can give an indication of what degradation or metabolic products might be formed.

Figure 10 shows products that were found similar when the transformation of maximum reduction was used.

The final structure obtained by this transformation from any one of the structures contained in this hit list is 1-aminopropane, which is also contained in this list. Indeed, any one of these compounds will give 1-aminopropane on complete reduction. However, depending on which compound provided the query structure, one might also find intermediates of the reduction process.

In the Introduction, the similarity concept was explained with the example of 3-aminobutyric acid and 1,3-butanediol. Figure 11 gives the full list of structures obtained for these compounds using the combined transformation hydrolysis and oxidation. The base structure for this similarity comparison is acetoacetic acid. It is not found in the hit list because it is an unstable compound. With respect to similarity however, this does not matter.

On closer inspection, a variety of reaction pathways connecting the structures of Figure 11 will be found. Just a few examples are

3-Aminocrotonitrile can be converted by hydrolysis and reduction into 1,3-butanediol.

The dimer of ketene, 4-methyleneoxetan-2-one, reacts with ethanol to give ethyl acetoacetate.

The same compound can be transformed with ammonia into 3-aminocrotonamide or 3-aminocrotonitrile or further, on reduction, into 3-aminobutyric acid.

Thus, one single similarity search sheds light onto a manifold of compounds that can be interconverted by standard reactions. This underscores the merit of using the transformations based on general reaction types to encapsulate reactions useful for interconverting organic compounds.

Not only those transformations based on generalized reactions but also those transformations isolating a specified substructure can be useful in pointing out reactions within a set of compounds.

Figure 12 shows part of a hit list obtained by the

combination-transformation that first isolates the skeleton and the  $\alpha$ -atoms and then converts the substituted carbon atoms to their maximum oxidation state.

The first step in the transformation, which isolates the skeleton and the  $\alpha$ -atoms of its substituents, cleaves the structures at the heteroatoms ( $\alpha$ -atoms). Clearly, this corresponds, in this case, to a hydrolysis reaction. A special built-in feature of this transformation converts trihalomethyl groups into a carboxylic acid, accounting for the occurrence of this reaction step on hydrolysis under forced conditions. The base structure of this transformation is pyruvic acid, which is also contained in Figure 12.

**Synthesis Design.** The results of the previous searches (Figures 10–12) can also be used in the planning of organic syntheses.

For example, if any one of the compounds contained in Figure 11 had been the target of a synthesis, this similarity search would have suggested that, among other possibilities, 1,3-butanediol could serve as a starting material. Results with other similarity transformations using 1,3-butanediol have already been reported.<sup>5</sup>

Next, searches for starting materials for a synthesis of 6-aminopenicillanic acid (APA) are performed. The results reported in Figure 13 were obtained with APA as query and using the transformation that isolates the carbon skeleton together with the  $\alpha$ -atoms of the substituents.

This is a rather restrictive search criterion since it asks for all the heteroatoms to be present on the carbon skeleton of the precursor molecules. Therefore, all the structures found are variants of one basic compound, 2-amino-3-methyl-3-mercaptoputanoic acid or penicillamine. This compound is in the catalog in three different stereoisomers (D, L, and the racemic D,L form) as well as an acetyl derivative, the acetone adduct, and the disulfide form. These compounds might indeed be used as starting materials for the synthesis of APA. In any case, penicillamine is a precursor in the biosynthesis of APA.

Clearly, the same hit list could have been obtained by a substructure search with a query consisting of the carbon skeleton and the  $\alpha$ -heteroatoms, with the rest as free sites. However, this would have required manual preparation of this exact query. In a transformation similarity search, the query profile is prepared automatically from the input

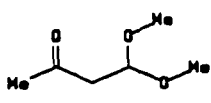
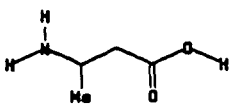
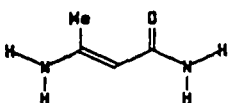
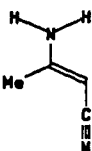
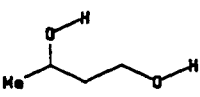
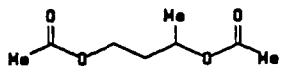
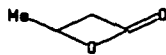
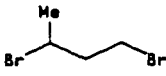
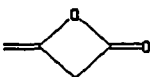
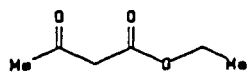
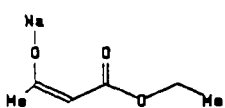
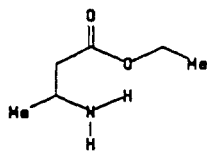
<p>72 JANSSEN 1024358</p>  <p>ACETYLACETALDEHYD-DIMETHYLACET</p>	<p>367 JANSSEN 1032745</p>  <p>DL-3-AMINOBUTTERSÄURE 99%</p>	<p>405 JANSSEN 1798742</p>  <p>3-AMINOCROTONSÄUREAMID 98%</p>
<p>406 JANSSEN 1524516</p>  <p>3-AMINOCROTONSÄURENITRIL, 96-</p>	<p>1505 JANSSEN 1076292</p>  <p>1,3-BUTANDIOL 99%</p>	<p>1508 JANSSEN 1917667</p>  <p>1917667 1117-31-3</p>
<p>1636 JANSSEN 2042858</p>  <p>BETA-BUTYROLACTON 98.5%</p>	<p>2693 JANSSEN 1688204</p>  <p>1,3-DIBROMBUTAN, 97.5%</p>	<p>3258 JANSSEN 1152781</p>  <p>1152781 2130-41-8</p>
<p>3956 JANSSEN 1179760</p>  <p>ACETESSIGSÄUREETHYLESTER 99%</p>	<p>3957 JANSSEN 1727711</p>  <p>ACETESSIGSÄUREETHYLESTER NATR</p>	<p>3969 JANSSEN 1180164</p>  <p>1180164 5303-65-1</p>

Figure 11. Structures found similar by the combined hydrolysis and oxidation transformation.

structure, in this case from APA, by only specifying which kind of transformation is desired.

To slightly broaden the search profile for compounds similar to APA, the transformation that selects the carbon skeleton together with the substituents as present in the query structure was chosen, using the variant that also considers double bonds as substituents. Figure 14 shows the results.

Now, in addition to the structures already found in the previous search (Figure 13), 3,3-dimethylacrylic acid and two of its derivatives were found. As already discussed above (cf. Figure 4), 3,3-dimethylacrylic acid can be considered as a starting material for the synthesis of penicillamine and thus also for APA.

#### SIMILARITY MEASURES BASED ON DESCRIPTIONS OF REACTION SITE

Quite a few definitions of the similarity of reactions are conceivable. A rather pragmatic approach is to define those reactions as similar that proceed under similar reaction conditions. As reaction conditions are presently incompletely specified in reaction databases, the similarity of reaction conditions has to be translated into other search criteria.

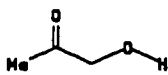
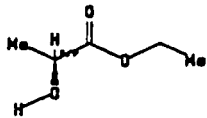
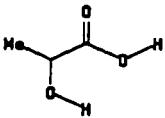
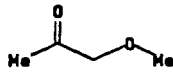
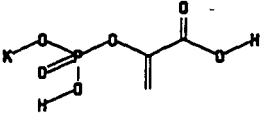
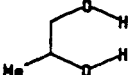
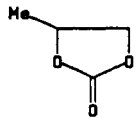

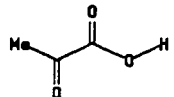
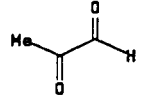
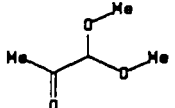
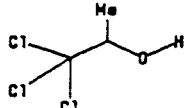
Reaction conditions are largely dependent on the mechanism of a reaction, and vice versa. The mechanism of a reaction for its part is determined by electronic and energy effects at the reaction center: bonds with low bond dissociation energies are prone to undergo radical cleavage, whereas highly polar bonds tend to break heterolytically and are thus favored by polar solvents and accelerated by acid or base catalysis. This suggests the use of parameters describing electronic and energy effects at the reaction site for the definition of the similarity of reactions.

**Electronic and Energy Effects.** In the last 10–15 years, we have developed a series of empirical methods for calculating all-important chemical factors like charge distribution,<sup>13,14</sup> inductive,<sup>15</sup> resonance,<sup>14</sup> and polarizability<sup>16</sup> effects and bond dissociation energies.<sup>17</sup> These procedures are fast enough to be applicable to large datasets of chemical reactions with reasonable response times.

Table II shows bond dissociation energies (BDE) of a series of ester and ether cleavages.

It can be seen that the variations of the bond dissociation energy of the R–O bond within the ester or within the ether series are much larger than those between esters and ethers



<p>49 JANSEN 1603429</p>  <p>HYDROXYACETON TECHN. 5392 JANSEN 1757316</p>	<p>4129 JANSEN 1186632</p>  <p>L(-)-MILCHSAEUREETHYLESTER 98% 6953 JANSEN 2291010</p>	<p>5216 JANSEN 1250690</p>  <p>DL-MILCHSAEURE, EUR. PHARM., 8. 7091 JANSEN 1587261</p>
 <p>METHOXYACETON 99% 7124 JANSEN 1315661</p>	 <p>PHOSPHOENOLBRENZTRAUBENSAEURE 7125 JANSEN 1496224</p>	 <p>1,2-PROPANDIOL, 99% 7246 JANSEN 1321422</p>
 <p>1,2-PROPYLENCARBONAT 99% 7248 JANSEN 1757922</p>	 <p>PROPENOXID 99% 7249 JANSEN 1738017</p>	 <p>BRENZTRAUBENSAEURE 95% 7910 JANSEN 1599688</p>
 <p>METHYLGLYOXAL 40 WT% LOESUNG I 1599688 76-00-6</p>	 <p>METHYLGLYOXALDIMETHYLACETAL 98</p>	 <p>1599688 76-00-6</p>

**Figure 12.** Results of a combined transformation that isolates the carbon skeleton with the  $\alpha$ -atoms of the substituents and then performs an oxidation.

with the same R groups. Particularly remarkable is the rather low value of the BDE for the benzyl derivative in both the ester and ether case. This suggests that there might be a way to cleave a benzyl ester that is more closely related, more similar, to the cleavage of a benzyl ether than to any of the other ester cleavages. Indeed this is the case: both a benzyl ester and a benzyl ether can rather easily be cleaved by hydrogenation under catalysis by palladium (Scheme II).

The mechanism of the cleavage of benzyl ester is quite different from most of the other cleavages of esters that proceed under acid or base catalysis. And, in fact, the rather easy hydrogenolysis of benzyl esters and benzyl ethers is founded on the low bond dissociation energy of both compounds.

This example shows that important conclusions on the conditions (and on the mechanism) of a reaction can be drawn from consideration of an energy parameter. Furthermore, it shows how reactions that are put into different classes (ester or ether cleavage) might be seen to be quite similar when physicochemical parameters at the reaction site are considered and found to have similar values.

In the preceding example, a reaction is predominantly characterized by a single parameter, the dissociation energy

of a bond. This is more the exception than the rule. Most chemical reactions are simultaneously influenced by several effects. Even then the electronic and energy parameters can be used for the definition of the similarity of reactions, but they have to be used in concert.

This will be explored with a series of reactions that have the same reaction center in common, the breaking of a CC double bond and a CH bond and the making of a CC and a CH bond (Figure 15).

These reactions encompass Friedel-Crafts alkylations by olefins, the free radical addition of ketones to olefins, and Michael additions. The task is now to differentiate these various types of reactions and simultaneously to recognize the similarity of the instances within each of these classes. The reactions were classified by BRANGÄNE, an unsupervised machine-learning program that bases its classification on topological similarity.<sup>18</sup> With an eye toward extending BRANGÄNE so that it can exploit physicochemical parameters for classification, parameters were calculated for the reaction in this dataset.

As a first entry, for several free radical reactions and Michael additions, the differences in the total charge in the two reacting

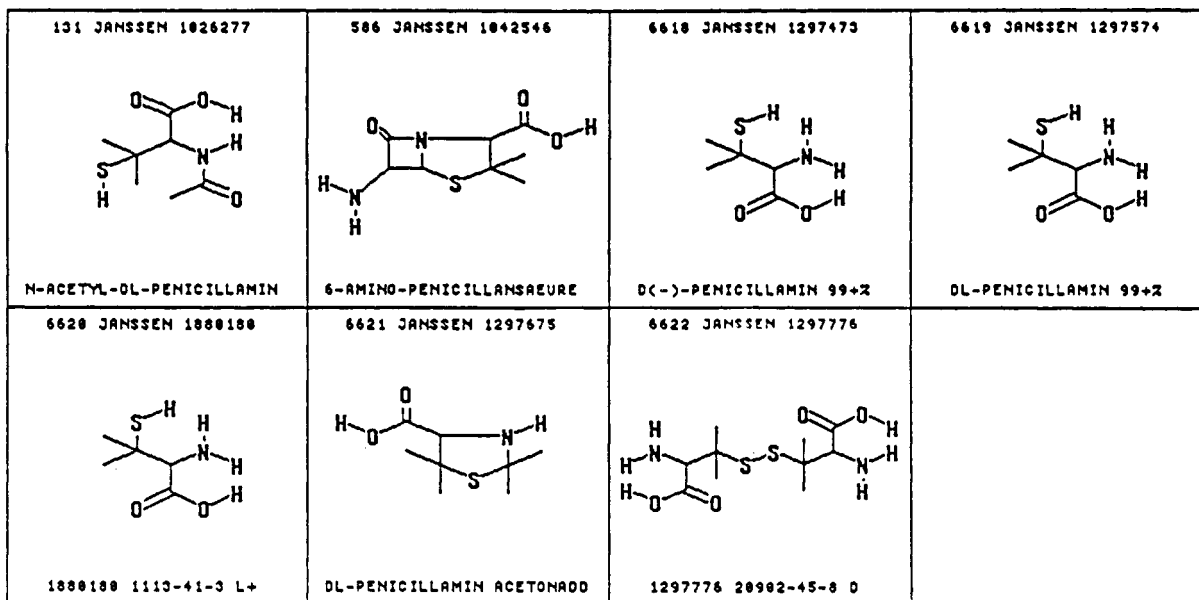


Figure 13. Structures found similar to 6-aminopenicillanic acid (APA) using the transformation that selects the carbon skeleton together with the  $\alpha$ -atoms of the substituents.

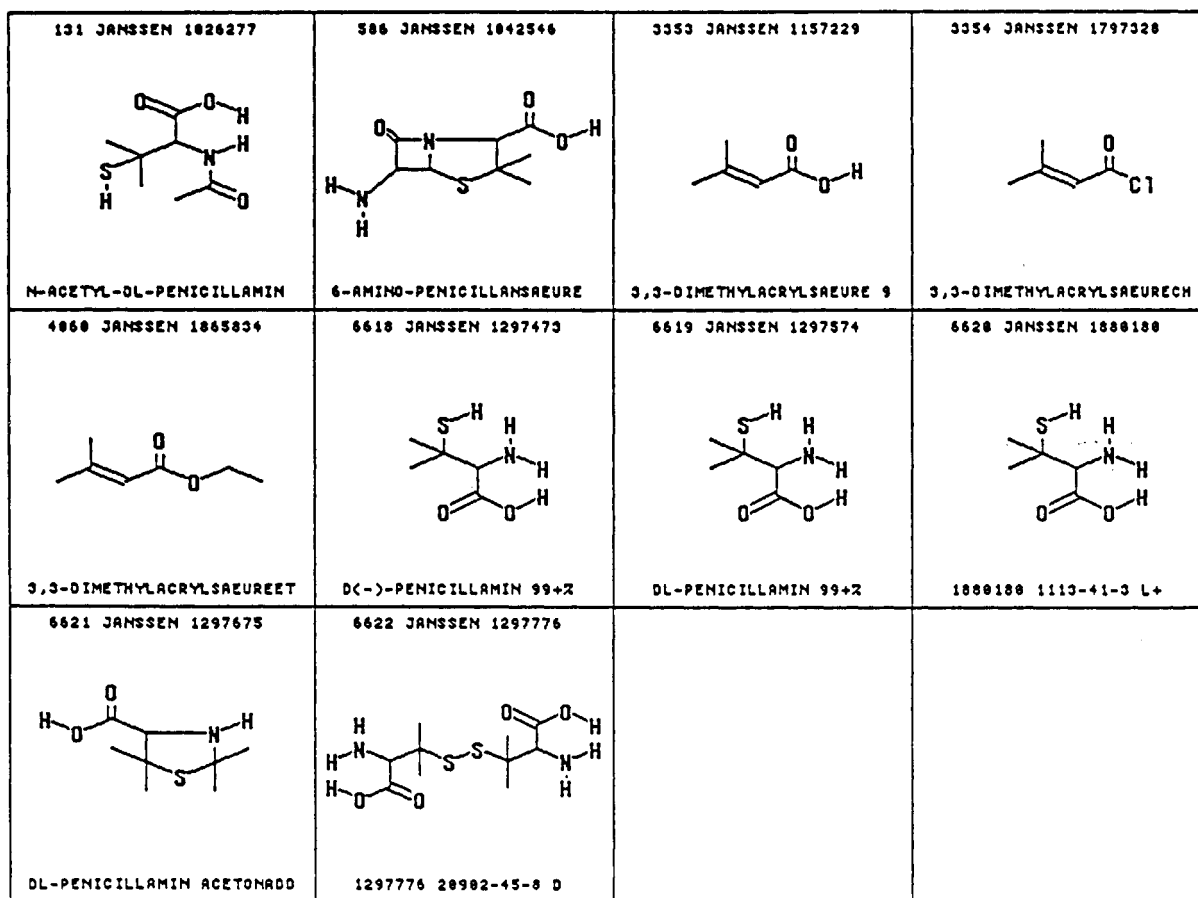


Figure 14. Structures found similar to APA by the transformation that selects the carbon skeleton and the substitution pattern with double bonds being considered as substituent.

bonds, the C-H and the C=C bonds, are plotted against each other (Figure 16).

Whereas the charge difference of the CH bond taken as the vertical axis does not separate the two types of reactions, the free radical and the Michael addition, the charge difference of the C=C bond taken as the horizontal axis just about separates the two types of reactions. Clearly, a better separation is obtained when both charge differences are used and a diagonal separating line is drawn through the plot of Figure 16.

However, when members of all three types, Friedel-Crafts reactions, free radical reactions, and Michael additions, are taken, the two charge differences no longer suffice to separate these three types of reactions. In this situation, additional parameters were used to describe the two bonds of the reaction center.

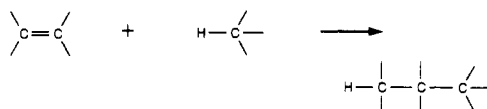
These included structural parameters like the number of non-hydrogen atoms on the atoms of the reaction center, whether they are part of an aromatic system or not, and the size of the ring in which they are embedded. Furthermore,

**Table II.** Bond Dissociation Energies (BDE) Calculated for Various Esters and Ethers (cf. Ref 17)

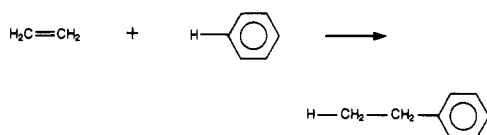
BDE (O-R)	CH <sub>3</sub> -CO-O-R	CH <sub>3</sub> O-R
CH <sub>3</sub>	346.2	344.9
C <sub>2</sub> H <sub>5</sub>	348.4	347.2
CH(CH <sub>3</sub> ) <sub>2</sub>	340.2	339.0
C(CH <sub>3</sub> ) <sub>3</sub>	334.6	333.4
C <sub>6</sub> H <sub>5</sub>	392.6	422.6
CH <sub>2</sub> -C <sub>6</sub> H <sub>5</sub>	295	293.7

<sup>a</sup> In kJ/mol.

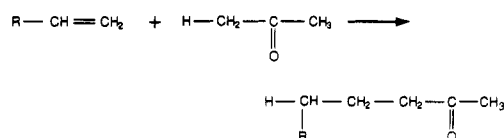
Reaction Site



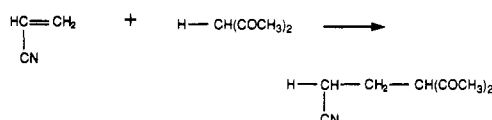
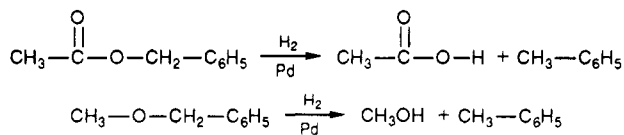
Friedel - Crafts Reaction



Free Radical Reaction



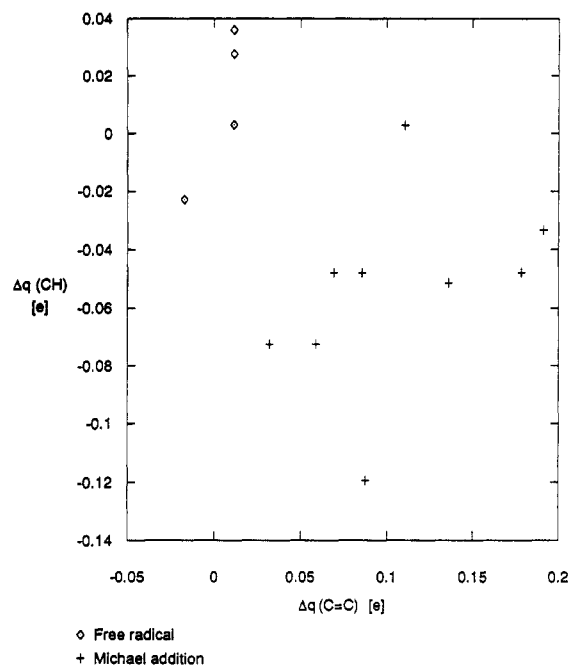
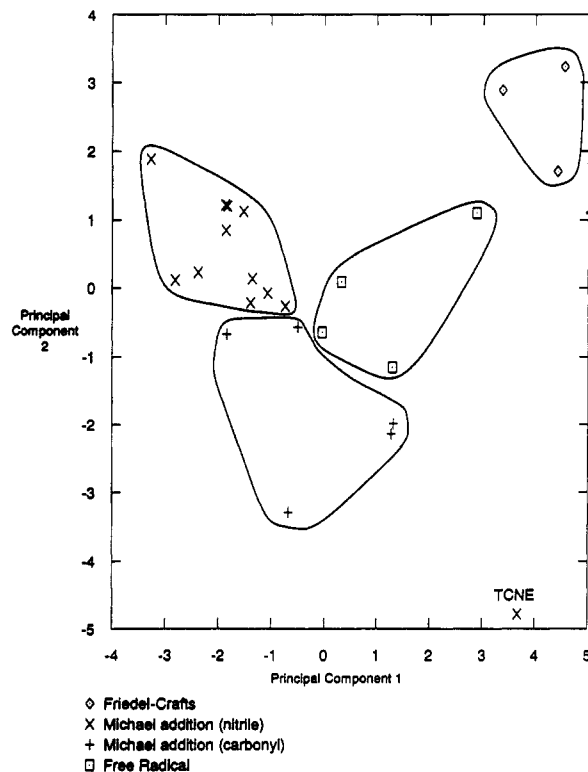
Michael Addition

**Figure 15.** Reactions with a common reaction center.**Scheme II**

the bond dissociation energy and electronic parameters like partial charges, difference in  $\pi$ -electronegativity, and the resonance stabilization of charges generated on breaking the bonds were used. Altogether there were 15 parameters.

Then a principal component analysis (PCA) was performed. The first two components obtained in this analysis are plotted against each other in Figure 17. Significantly, this plot agrees almost completely with BRANGÄNE's classification even though members from different classes lie very close together.

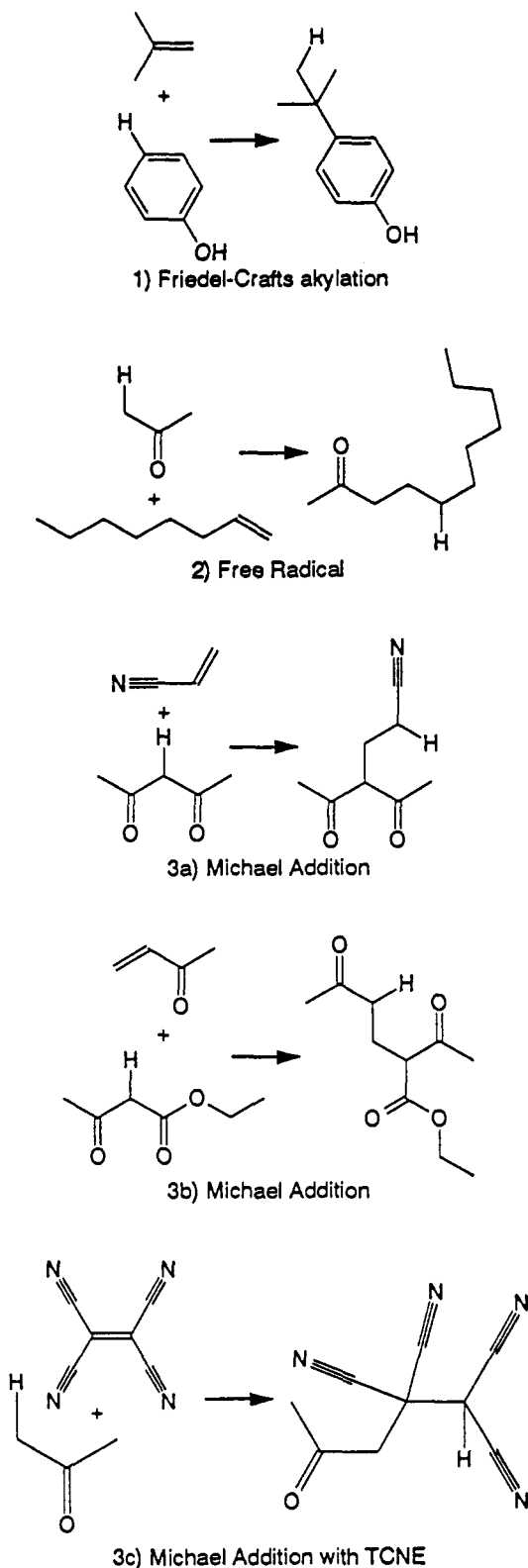
As indicated, the three types of reactions separate into individual clusters. In fact, the instances of the Michael addition can be further differentiated into those that have a nitrile group at the double bond and those that have not, which BRANGÄNE did. The reaction that involves an addition to tetracyanoethylene (TCNE) is not clustered with any of the other reactions, in contrast to BRANGÄNE's clustering, because it has two nitrile groups on both ends of the double bond. This shows how electronic and energy parameters can

**Figure 16.** Plot of the difference in the total charge in the CH bond (vertical axis) against the charge difference in the CC double bond (horizontal axis) for a series of free radical reactions and Michael additions.**Figure 17.** Plot of the first two components from a principal component analysis of various electronic and energy effects of the two bonds in the insertion of a CC double bond into a C-H bond.

be used to improve a classification based purely on topological features.

Figure 18 shows a representative example from each of the clusters obtained in the plot of the first two factors of the PCA.

This example shows that the various electronic and energy parameters used are able, in concert, on the one hand, to differentiate the various mechanisms for the insertion of an olefin into a C-H bond and, on the other hand, to cluster



**Figure 18.** Instances of reactions for each of the four clusters of the plot of Figure 17.

instances of the same mechanism or reaction type into groups of similar reactions.

#### APPLICATIONS OF SIMILARITY OF REACTIONS

**Reaction Planning.** The perception of the similarity of reactions based on physicochemical parameters of the reaction site can be used for the planning of reactions, for predicting which course a reaction will take, and for selecting reaction conditions to initiate a desired reaction. In the case of the

cleavage of esters and ethers (Table II), one parameter, the bond dissociation energy, suffices to make a decision: If the value of the BDE is low, the cleavage of an ester or ether can be achieved by a homolytic process. The palladium catalyzed hydrogenation is such a reaction.

From the example of the insertion of a CC double bond into a C-H bond (Figures 15-18), the more general form of a procedure for the classification of a reaction by similarity can be derived.

1. Search for all those reactions that have the same reaction center as the reaction to be investigated (the query).
2. Calculate parameters on the electronic and energy effects at the reaction site.
3. Perform a principal component analysis (PCA) on the dataset of retrieved reactions that are characterized by the physicochemical parameters. Alternatively, if a supervised learning method is desired, a linear discriminant analysis (LDA) can be carried out.
4. Locate the query structure within one of the various clusters of the PCA or LDA and thus define similar reactions.
5. Derive the appropriate reaction conditions from those of the similar reactions.

In the case of the insertion of a CC double bond into a C-H bond, the location of the query reaction within one of the four clusters of reactions (see Figure 17) can be used to decide whether the reaction should be run under catalysis by an acid or a Lewis acid (Friedel-Crafts alkylation), by a radical initiator (free radical addition), or by a base (Michael addition).

**Acquisition of Knowledge about Reactions.** The perception of similarity between organic reactions based on physicochemical parameters makes it possible to draw conclusions on the kind of electronic and energy effects that influence a particular reaction and to what extent these effects are of importance.

In this way, knowledge on the driving forces and the mechanisms of organic reactions can be derived. These approaches can also be incorporated into systems for the automatic acquisition of knowledge about chemical reactions.<sup>18</sup> We are presently exploring the benefits of the definition of similarity of reactions based on physicochemical parameters in the application of machine learning techniques to organic reactions.

#### CONCLUSIONS

The definition of structural similarity based on general reaction types and on a single gross structural feature criterion can be used for the analysis of the information content of databases of structures, for predicting reactions, and for designing syntheses. The representation of these similarity criteria as structural transformations, in combination with a hash-coding algorithm, makes it possible to perform the search for structural similarity directly in memory. This provides for efficient and rapid searches of databases of structures.

The similarity of reactions has been defined by physicochemical parameters calculated for the atoms and bonds of the reaction center. The perception of the similarity of reactions allows one to draw conclusions on the mechanism of a reaction and appropriate reaction conditions and can help in the planning of reactions. Furthermore, these concepts

can be used in the automatic acquisition of knowledge about chemical reactions by machine learning techniques.

#### ACKNOWLEDGMENT

Support of this work by Studienstiftung des Deutschen Volkes von Humboldt-Stiftung and Deutsche Forschungsgemeinschaft is gratefully acknowledged. We thank Janssen Chimica, Beerse, Belgium, and Merck, Darmstadt, Germany, as well as Prof. S. Hanessian for making their databases of structures available to us.

#### REFERENCES AND NOTES

- (1) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
- (2) Warr, W. A. *Chemical Structures II*; Springer-Verlag: Berlin, 1991.
- (3) Gasteiger, J.; Ihlenfeldt, W. D.; Röse, P. A collection of computer methods for synthesis design and reaction prediction. *Recl. Trav. Chim. Pays-Bas*. **1992**, *111*, 270–290.
- (4) Ihlenfeldt, W. D.; Gasteiger, J. In *Software-Entwicklung in der Chemie 2*; Gasteiger, J., Ed.; Springer-Verlag: Heidelberg, 1988; pp 13–33.
- (5) Ihlenfeldt, W. D.; Gasteiger, J. In *Software-Development in Chemistry 5*; Gasteiger, J., Eds.; Springer-Verlag: Heidelberg, 1991; pp 187–195.
- (6) Gasteiger, J.; Ihlenfeldt, W. D. In *Chemical Structures II*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1991.
- (7) Hanessian, S.; Franco, J.; Gagnon, G.; Laramée, D.; Larouche, B. Computer-Assisted Analysis and Perception of Stereochemical Features in Organic Molecules Using the Chiron Program. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 413–425.
- (8) Gasteiger, J.; Ihlenfeldt, W. D. In *Software-Development in Chemistry 4*; Gasteiger, J., Ed.; Springer-Verlag: Heidelberg, 1990; pp 57–65.
- (9) Gasteiger, J.; Hutchings, M. G.; Christoph, B.; Gann, L.; Hiller, C.; Löw, P.; Marsili, M.; Saller, H.; Yuki, K. A New Treatment of Chemical Reactivity: Development of EROS, an Expert System for Reaction Prediction and Synthesis Design. *Top. Curr. Chem.* **1987**, *137*, 19–73.
- (10) Gasteiger, J.; Ihlenfeldt, W. D.; Röse, P.; Wanke, R. Computer-Assisted Reaction Prediction and Synthesis Design. *Anal. Chim. Acta* **1990**, *235*, 65–75.
- (11) Röse, P.; Gasteiger, J. Automated Derivation of Reaction Rules for the EROS 6.0 System for Reaction Prediction. *Anal. Chim. Acta* **1990**, *235*, 163–168.
- (12) Röse, P.; Gasteiger, J. In *Software-Development in Chemistry 4*; Gasteiger, J., Ed.; Springer-Verlag: Heidelberg, 1990; pp 275–288.
- (13) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity—A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219–3228.
- (14) 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.
- (15) 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.
- (16) Gasteiger, J.; Hutchings, M. G. Quantification of Effective Polarisability. Applications to Studies of X-ray Photoelectron Spectroscopy and Alkylamine Protonation. *J. Chem. Soc., Perkin Trans. 2* **1984**, 559–564.
- (17) 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.
- (18) 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. Comput. Sci.* **1990**, *30*, 492–504.