Models, Concepts, Theories, and Formal Languages in Chemistry and Their Use as a Basis for Computer Assistance in Chemistry[†,‡]

Ivar Ugi,* Johannes Bauer, Carola Blomberger, Josef Brandt, Andreas Dietz, Eric Fontain, Bernhard Gruber, Annette v. Scholley-Pfab, Antje Senff, and Natalie Stein

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

Received September 1, 1993[⊛]

The present state of the work on the logic-oriented approach to the computer handling of chemical information is discussed. About 2 decades after the advent of the constitution-oriented algebra of the *be*- and *r*-matrices, and the stereochemical theory of the chemical identity group, new chemical definitions, concepts, and perspectives have expanded the field of applicability of this mathematical model. This paper describes how the logical structure of chemistry can be used in the field of reaction generation, reaction classification, reaction documentation, and the elucidation of reaction mechanisms. An introduction into new aspects of formal stereochemistry and the representation of configurational and conformational information is given. A new data structure that allows handling of stereochemical features and delocalized electron systems by the so-called,extended *be*- and *r*-matrices is presented.

1. EARLY DEVELOPMENT OF COMPUTER CHEMISTRY

As soon as computers were available, they were used in various fields of chemistry, especially in physicochemistry and in quantum chemistry. The evaluation of spectroscopic data and chemical documentation began about 1960.[1,2]

In 1965 Lederberg and his colleagues initiated the DENDRAL Project with the goal of developing a computer program that determines molecular features from spectroscopic data and assembles these substructures into the complete structure.[3-5] The underlying concept of DENDRAL can be considered as a milestone in the development of computer chemistry. For the first time, formal methods were applied to solve chemical problems. DENDRAL also strongly influenced the development of artificial intelligence.

From 1960 on Vladutz has introduced the logically organized storage and systematic retrieval of molecular structures and chemical reactions by computers.[6] Subsequently Vladutz has proposed the computer-assisted design of multistep organic syntheses (CAOS: computer-assisted organic syntheses). This concept relies on the use of databases of known chemical reactions. The synthesis of a desired target compound is then proposed by generating appropriate sequences of conventional reactions or their analogues.[7] During his discussions Vladutz had mentioned many times that the chemical horizon of these programs is limited to those parts of chemistry that are stored in the databases. He recognized that such computer programs can never be able to generate fundamentally new chemical reactions and syntheses.

The memory of G. Vladutz should be honored like that of the most renowned colleagues, who were considered as the leading scientists before World War II. Then progress in the natural sciences was defined through fundamentally new ideas and concepts. Nowadays scientists seem to believe that progress lies in the generation and the organization of large amounts of analogous and improved scientific methods and results. For this kind of research it is often easier to recruit more co-workers and to get more financial support.

In 1967 the ideas of Vladutz led Corey, Wipke, and their co-workers at the University of Harvard to the development of the computer-assisted synthesis design program LHASA.[8] LHASA is based on systematic degradations of target molecules into their precursor molecules, that ultimately lead to commercially available starting compounds.[9,10] The preparative pathways of organic syntheses are based on a collection of preferred chemical reactions or their analogues. After Wipke had left the LHASA project, he and his collaborators began to develop the improved synthesis design program SECS at the Princeton University.[11] He continued this later at University of California at Santa Cruz. Wipke's SECS computer program was further revised to the computer design program CASP, a joint project of the eight major German and Swiss chemical and pharmaceutical companies. It is a systematically improved synthesis design system with a database system of about 20 000 preparative organic reactions. Wipke derived from his SECS the extremely successful molecular data system MACCS and his chemical reaction database REACCS, which were later sold to the MDL Maxwell Co. The databases MACCS and REACCS have now been combined into the ISIS collection.

Gelernter published the "realization of a geometry theorem proving machine" in 1960.[12,13] From 1970 on, Gelernter et al. developed at the State University of New York, Stony Brook, the chemical database synthesis design program SYNCHEM with an analogous logical approach as the geometry theorem proving computer program.[12,14] Hendrickson developed his "mathematical representation of the structural features of organic molecules and their interconversions".[15] The corresponding computer program for the planning of organic syntheses was developed by Moreau at Paris;[16] at that time it was very impressive. Hendrickson and Toczko subsequently also produced a computer program[17,18] on the basis of the latter theory.

At Bayer AG in 1965-7 the mathematician Kaufhold and the chemist Ugi[19] began to investigate which new types of chemical problems can be solved with the aid of mathematics and the then new computer systems in chemistry. The extremely complex reaction mechanism and the conditions for the optimization of syntheses of peptide derivatives by four component condensations (4CC, Ugi reaction[20]) could

[†] Presented at The Third International Conference "Chemical Structures, The International Language of Chemistry" Leeuwenhorst Congress Center, Noordwijkerhout, The Netherlands, June 6-10, 1993.

[‡] Dedicated to the 75th birthday of Sir Derek Barton.

[⊛] Abstract published in *Advance ACS Abstracts,* December 15, 1993.

be analyzed by these methods. They also developed a computer program for the bilateral design of peptide syntheses, a program that generates combinations of conventional peptide syntheses and 4CCs leading from known starting materials to the protein target. The probabilities of the individual known reactions are evaluated by the estimated reaction yields of connecting pairs of amino acids by conventional amine acylations or estimated yields of tripeptide syntheses by 4CCs.[21-23]

This early computer-assisted planning of chemical research demonstrated that complex problems can be advantageously solved with the help of computers. However, it was also observed, that it is often more effective not to confine the investigations just to chemistry. It was recognized, that the introduction of new or improved mathematical methods into computer-assisted research can become much more effective and economical, if the mathematical background itself is improved. In chemistry the scientific, preparative and experimental methods can in most cases be optimized by the application of basic logic and mathematical reasoning.

1.1. Attempt To Develop Chemical Computer Programs That Are Not Confined to Collections of Known Chemical Data or Analogous Data. The early synthesis-oriented computer programs were obviously limited to available chemical information. These limitations were already seen in the early protein synthesis design program[22] by Kaufhold and Ugi at Bayer AG and the SYNCHEM project by Gelernter et al.,[12,14] as well as the data-oriented synthesis design computer programs by Corey, Wipke, and their coworkers.[24-26] During a discussion about the generally admired LHASA program of Corey in London in 1972, Woodward made the observation that such computer programs can only generate syntheses from limited collections of published chemical and physicochemical data. Steven and Brownscombe[27] recognized already in 1972 that the possibility to correctly evaluate generated reactions is limited by the accessibility of the corresponding chemical energy data. The calculation of such data was later improved and generalized much farther by Gasteiger et al. in their synthesis design program EROS.[28,29] Also, Hendrickson's approach[15] was limited by the scarcity of the available molecular data.

1.2. Mathematical and Logical Basis for Computer Chemistry. Long before the development of quantum theory, various mathematical concepts and methods have played important roles in the early conceptual progress in chemistry.

Until 1970, the application of mathematics in chemistry belonged in essence to two areas: on the one hand the graph theoretical description of molecules and on the other hand the enumeration of isomers by group theory, which is still investigated.[30,31]

The historical discovery of the chemical elements is connected with their representation by various symbols that led to their description by combinations of one or two letters like the hydrogen *H*, helium *He*, lithium *Li*, etc. A famous international chemical conference took place at Karlsruhe, Germany, in 1860. It was then decided to represent chemical formulas by combinations of lines for chemical bonds and symbols for the chemically connected atoms. Nowadays chemical formulas advanced to a language for describing chemical molecules and their changes.[32]

In 1874 Le Bel and van't Hoff developed their early stereochemical concepts[33,34] that were based on group theory and permutations, and about the same time the mathematician Cayley[35] began to describe molecules by graph theoretical concepts.

This is still proceeding further. During the past 25 years in Bucharest, Balaban[36] has developed graph theoretical molecular chemistry. A successful combination of graph theoretical and group theoretical methods in permutational chemistry and computer chemistry has been achieved in the past 10 years by Kvasnička and co-workers[37,38] in Slovakia. In Moscow, Zefirov, Tratch, et al.[39] have also developed some comparable mathematical approaches in chemical and computer oriented methods.

From 1935 on, Pólya[40] applied permutation group theory in order to count the number of constitutional and stereochemical isomers. Since that time many mathematicians and chemists have developed such counting procedures. In particular Kerber and Thürlings[41] have been very productive in the counting of isomers by permutational methods for more than 15 years.

Ruch and Ugi have improved the treatment of reactions of enantiomers by permutational group theory in 1965–9.[42,43]

In 1970–3 Dugundji and Ugi[44] came to the conclusion that also some other types of mathematical approaches can lead to further progress in the treatment of constitutional chemistry and stereochemistry. It was then also realized that various definitions in chemistry as well as their logical approaches should be reformulated.

In the beginning of computer chemistry, the constitutional structures of molecules was represented by their *adjacency matrices* and *topological matrices*,[45] later also by their *connectivity matrices*.[46] These matrices are useful to the computer-assisted representation of molecular constitutions, but reactions are not representable by matrix transformations in a chemically meaningful way. However, interconversions of constitutional isomers and isomeric ensembles of molecules (EM) can be represented by the elementwise difference of the entries of the bond- and electron-matrices (*be*-matrices) of the reaction product and the educt which make up the reaction matrix (*r*-matrix).[44]

It was necessary to develop a logical representation and description of chemistry, which should be as simple and applicable as possible. In consequence the following notions were redefined: Molecules consist of atomic cores (atomic nuclei and electrons of the inner shells) and valence electrons. They are held together by covalent or ionic bonds. A covalent bond corresponds to a pair of valence electrons that simultaneously belongs to two neighboring atoms. A chemical reaction is the conversion of an EM into an isomeric EM by redistribution of valence electrons. During a chemical reaction the atomic cores and the total number of valence electrons remain the same.

1.3. DU-Model. The DU-model[44] serves as the theoretical foundation for the computer-assisted deductive solution of chemical problems by computer programs.[47] The fundamental equation

$$\mathbf{B} + \mathbf{R} = \mathbf{E} \qquad (1)$$

of this model represents chemical reactions. When n atoms participate in a reaction, the $n \times n$ symmetric *be*-matrices $\mathbf{B} = \langle b_{ij} \rangle$ and $\mathbf{E} = \langle e_{ij} \rangle$ describe the chemical constitution of the reacting isomeric ensembles of molecules EM(**B**) and EM(**E**) at the beginning and at the end of the reaction, and the addition of the $n \times n$ *r*-matrix $\mathbf{R} = \langle r_{ij} \rangle$ transforms **B** into **E**; **R** expresses an electron-redistributing scheme, an "electron-pushing" pattern.

2. REACTION GENERATORS OF RAIN AND IGOR

The solutions of eq 1 can be found in two ways:

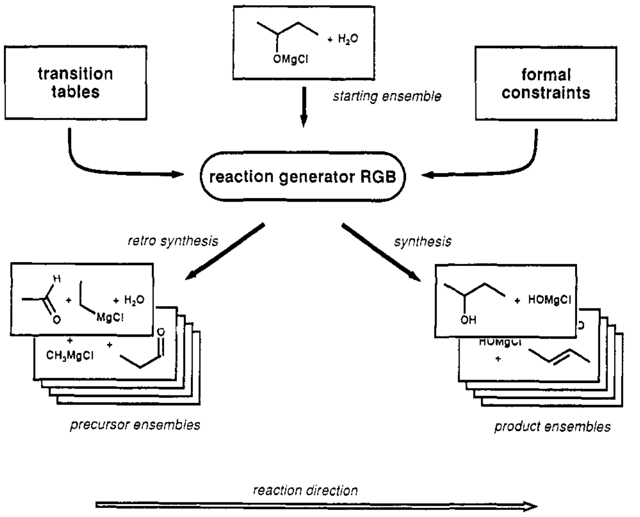


Figure 1. RAIN's reaction generator RGB.

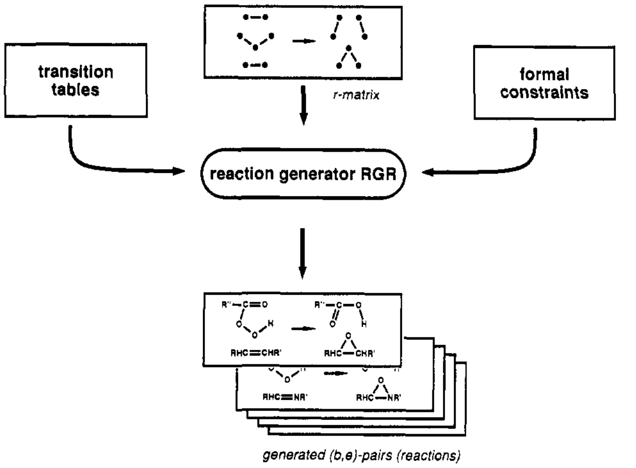


Figure 2. IGOR's reaction generator RGR.

I. When a *be*-matrix **B** is given, the pairs (**R**,**E**) that satisfy eq 1 under the given boundary conditions are the solutions of eq 1.

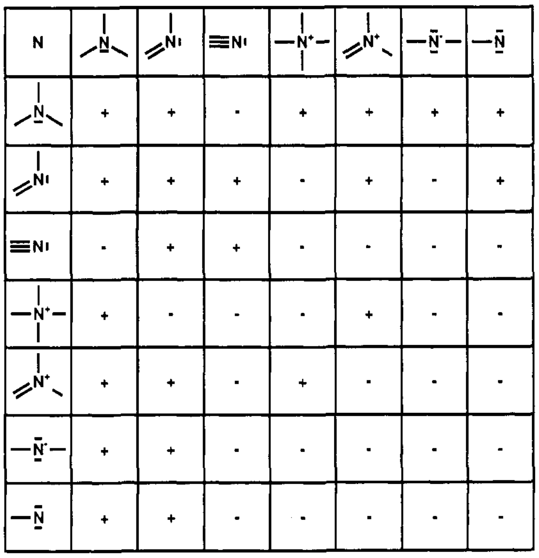
II. When the *r*-matrix **R** is given, the solutions to eq 1 are the conceivable pairs (**B**,**E**).

Solutions of type I are called *b*-solutions. They are found by reaction generators (RGs)[48] of the type RGB; Figure 1 illustrates the action of an RGB as it is incorporated in the program *RAIN* (*r*eactions *a*nd *i*ntermediates *n*etworks).[49-51]

Solutions of type II are called *r*-solutions. They are found by reaction generators of the type RGR; Figure 2 illustrates the action of an RGR as it is incorporated in the program *IGOR* (*i*nteractive *g*eneration of *o*rganic *r*eactions).[52,53]

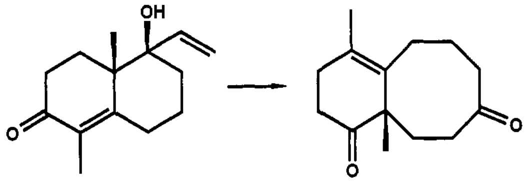
RAIN's RGB as well as IGOR's RGR are guided by transition tables (TTs) that record the allowed valence chemical behavior for each considered chemical element or pseudoelement. The rows of a transition table correspond to the valence schemes[54,55] of the atoms of the educts of a reaction and the columns to those of the atoms of the products of a reaction. An entry + or – at the intersection of a row and a column indicates whether or not the transition within a reaction, or a reaction step, is permitted, as seen in Figure 3.

Furthermore, RAIN and IGOR are controlled by sets of formal constraints that, e.g., describe the occurrence of formal charges. In both programs lists of required or forbidden substructures can be defined. The RGB of RAIN is constrained by rules for the maximum topological complexity of electron redistribution. The RGR of IGOR allows predefinition of certain required skeletal units in the educts or the products.



N	$>\underline{N}-$	$=N\vert-$	$\equiv N\vert$	$>N^{+}<$	$=N^{+}<$	$-\overline{\underline{N}}^{\cdot}-$	$-\overline{\underline{N}}$
$>\underline{N}-$	+	+	-	+	+	+	+
$=N\vert-$	+	+	+	-	+	-	+
$\equiv N\vert$	-	+	+	-	-	-	-
$>N^{+}<$	+	-	-	-	+	-	-
$=N^{+}<$	+	+	-	+	-	-	-
$-\overline{\underline{N}}^{\cdot}-$	+	+	-	-	-	-	-
$-\overline{\underline{N}}$	+	+	-	-	-	-	-

Figure 3. User-defined transition table of nitrogen.



1 2

Figure 4. Reaction of **1** → **2**.

The main advantage of using formally constrained RGs in contrast to library search of collected transforms is that there obviously exists no limit to the creativity of such a reaction generating system. Of course, in order to avoid combinatorial explosions, the supplied sets of boundary conditions have to be intensively and judiciously used.

3. OPTIMIZED GENERATION OF REACTION SEQUENCES WITH RAIN

The ensembles EM(**B**) and EM(**E**), representing a chemical multistep reaction of *n* atoms, can be regarded as two points in an n^2-dimensional hyperspace. The smallest L_1-distance between these points is called the minimum chemical distance (MCD).[37,56,57] It is equal to the double of the minimum number of relocated valence electrons.

A reaction sequence that connects EM(**B**) with EM(**E**) with minimum structural changes should only pass EMs that are located on or near a more or less straight line between EM(**B**) with EM(**E**) in n^2-dimensional hyperspace.

This distance metric can be used for the optimized generation of reaction sequences between starting materials and products of a chemical reaction.[48,58,59] The following illustrative example is constructed in order to demonstrate the usage of the MCD in the tree growing process.

The program RAIN was used to investigate formally possible reaction paths between the vinyl carbinol **1** and the bicyclic dione **2** (Figure 4).

A genetic algorithm based method[58,59] calculated a MCD value of 16 for these molecules. In order to build reaction sequences in a bilateral manner, the RGB of RAIN produced 23 EMs that emerge from **1** (ensembles **3–25**), and 25 EMs that are immediate precursors of **2** (ensembles **6**, **26–49**); Figure 5 shows a two-dimensional "road map" of the generated

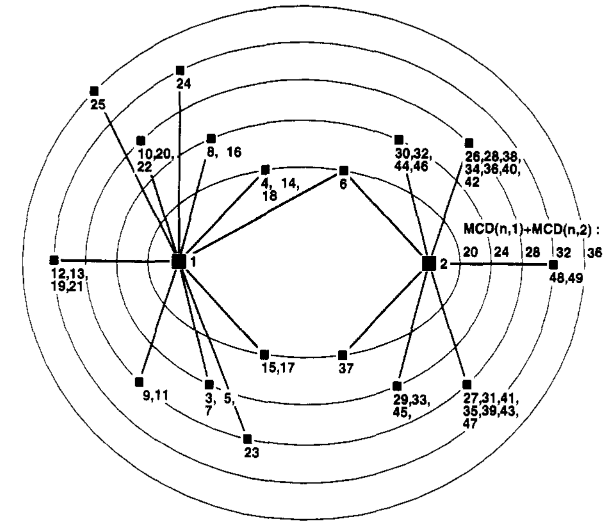


Figure 5. "Road map" of generated reaction trees.

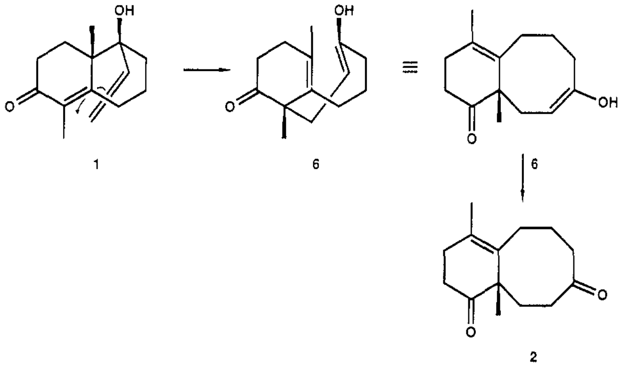


Figure 6. Reaction of **1** → **6** → **2**.

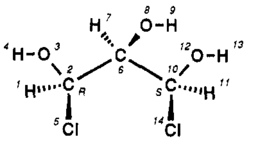
reaction trees. The EMs are located in a way such that their geometric distances to the locations of **1** and **2** correspond to their MCD values to the EMs **1** and **2**. EMs with the same sums of MCD values are located on ellipses with **1** and **2** as focus points.

It is obvious that most of the generated reaction steps do not lead into the direction of the opposite "goal". They could easily be discarded by a selection rule that restricts allowed intermediates to lie within the ellipse with MCD sum of 20 or 24. This would effectively reduce the amount of produced structures and would attract the user's attention to the main point of interest.

One EM (**6**) (Figure 6), lying on the ellipse with MCD sum 20, forms a bridge between **1** and **2**. The corresponding molecule represents an intermediate of a two-step reaction sequence leading from **1** to **2**. The first reaction step is a concerted (3,3) sigmatropic rearrangement of the Oxy-Cope moiety. The second step is a 1,3 tautomeric hydrogen shift. This is in excellent correspondence with the experimental results.[60]

4. CANONICAL NUMBERING OF ATOMS IN ENSEMBLES OF MOLECULES

Formal representations of molecules and chemical reactions form the basis for problem solving algorithms. The representation of a molecule has to be unique, in the sense that no other chemically distinct molecule is represented in the same way. However, there is the problem that one molecule has more than one representation.



50

Figure 7. Structure of **50** numbered arbitrarily.

In order to avoid time consuming algorithms for determining graph isomorphisms, it is helpful to find a normalized representation for each molecule. The algorithm CANON[61] employs subsequent partitioning of the set of atoms of the EM into equivalence classes in order to achieve an unique labeling of the atoms. An extended version of CANON considers constitutional aspects as well as configurational properties, no matter how many stereoisomers exist. Thus, CANON is superior to the MORGAN algorithm[62] which is used with CAS.

The following example shows the main rules of the extended CANON algorithm.

4.1. Considering Constitutional Aspects. The atoms of molecule **50** (Figure 7) are numbered arbitrarily. These numbers are used for constructing a connectivity list (Table 1, column a).

Rule 1: Each kind of atom is assigned an index ranked by its atomic number, starting with 1 for the atoms with the highest atomic number. In molecule **50** the indices for chlorine atoms will be 1, for oxygen 2, for carbon 3, and for hydrogen 4.

Rule 2: The indices of the neighbor atoms of each atom are sequentially numerically ordered. This sequence combined with the atom index of a considered atom results in the atom descriptor. In molecule **50** the atom descriptor of atom 2 is 3:1.2.3.4, where 3 is the atom index of atom 2 and 1, 2, 3, and 4 are the atom indices of its neighbor atoms (Table 1, columns b).

Rule 3: The set of atoms is ordered lexicographically according to the atom descriptors (Table 1, columns c).

Rule 4: Rules 2 and 3 are repeated until no more changes occur.

In case the atoms are not sorted in a unique way, the ring descriptor[61] must be determined. This descriptor is formed by the sequence of the sizes of the smallest rings in which that atom occurs. Atoms which cannot be distinguished by these criteria are constitutionally equivalent.

4.2. Considering Stereochemical Aspects. In some cases constitutionally equivalent atoms may be configurationally distinct and thus not chemically equivalent. Recently, CANON has been extended by a fifth rule which considers configurational properties:

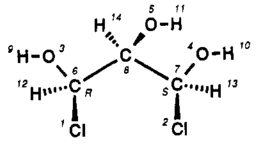
Rule 5: In each class of constitutionally equivalent atoms the configurationally relevant atoms are ordered according to the CIP-rules[63] R > S, proR > proS, and Z > E (Table 2, column d). If the application of rule 5 did not yield a further partitioning and there are still equivalence classes containing more than one atom, the algorithm has to be continued at rule 2.

Table 1. CANONical Indices of **50** without Considering Stereochemical Aspects

connectivity list	ordering by atomic no.	ordering by constitution						
(a)	(b_1)	(c_1)	(b_2)	(c_2)	(b_3)	(c_3)	(b_4)	(c_4)
1- 2	1- 4:3	5- 1:	5- 1:3	5- 1:	5- 1:4	5- 1:	5- 1:4	5- 1:
2- 1,3,5,6	2- 3:1.2.3.4	14- 1:	14- 1:3	14- 1:	14- 1:4	14-1:	14- 1:4	14- 1:
3- 2,4	3- 2:3.4	3- 2:	3- 2:3.5	3- 2:	3- 2:4.6	3- 2:	3- 2:4.6	3- 2:
4- 3	4- 4:2	8- 2:	8- 2:4.5	12- 2:	12- 2:4.6	12- 2:	12- 2:4.6	12- 2:
5- 2	5- 1:3	12- 2:	12- 2:3.5	8- 3:	8- 3:	8- 3:	8- 3:	8- 3:
6- 2,7,8,10	6- 3:2.3.3.4	2- 3:	2- 3:1.2.4.6	2- 4:	2- 4:1.2.5.7	2- 4:	2- 4:1.2.5.8	2- 4:
7- 6	7- 4:3	10- 3:	10- 3:1.2.4.6	10- 4:	10- 4:1.2.5.7	10- 4:	10- 4:1.2.5.8	10- 4:
8- 6,9	8- 2:3.4	6- 4:	6- 4:	6- 5:	6- 5:	6- 5:	6- 5:	6- 5:
9- 8	9- 4:2	4- 5:	4- 5:2	4- 6:	4- 6:2	4- 6:	4- 6:2	4- 6:
10- 6,11,12,14	10- 3:1.2.3.4	9- 5:	9- 5:2	9- 6:	9- 6:3	13- 6:	13- 6:2	13- 6:
11- 10	11- 4:3	13- 5:	13- 5:2	13- 6:	13- 6:2	9- 7:	9- 7:	9- 7:
12- 10,13	12- 2:3.4	1- 6:	1- 6:3	1- 7:	1- 7:4	1- 8:	1- 8:4	1- 8:
13- 12	13- 4:2	7- 6:	7- 6:4	11- 7:	11- 7:4	11- 8:	11- 8:4	11- 8:
14- 10	14- 1:3	11- 6:	11- 6:3	7- 8:	7- 8:	7- 9:	7- 9:	7- 9:

Table 2. CANONical Indices of **50** Considering Stereochemical Aspects

Ordering by R > S	ordering by configuration			
(d)	(b_5)	(c_5)	(b_6)	(c_6)
5- 1:	5- 1:4	5- 1:	5- 1:	5- 1:
14- 1:	14- 1:5	14- 2:	14- 2:	14- 2:
3- 2:	3- 2:4.7	3- 3:	3- 3:	3- 3:
12- 2:	12- 2:5.7	12- 4:	12- 4:	12- 4:
8- 3:	8- 3:	8- 5:	8- 5:	8- 5:
2- 4:	2- 4:	2- 6:	2- 6:	2- 6:
10- 5:	10- 5:	10- 7:	10- 7:	10- 7:
6- 6:	6- 6:	6- 8:	6- 8:	6- 8:
4- 7:	4- 7:2	4- 9:	4- 9:3	4- 9:
13- 7:	13- 7:2	13- 9:	13- 9:4	13- 10:
9- 8:	9- 8:	9- 10:	9- 10:	9- 11:
1- 9:	1- 9:4	1- 11:	1- 11:	1- 12:
11- 9:	11- 9:5	11- 12:	11- 12:	11- 13:
7- 10:	7- 10:	7- 13:	7- 13:	7- 14:



51

Figure 8. Structure of **50** with CANONical indices.

Rule 6: The breaking of the equivalence classes is carried out as described in ref 61.

5. FORMAL DESCRIPTION OF STEREOCHEMISTRY

Even at temperatures close to absolute zero, molecules are permanently in motion. Therefore, an isomer is a collection of a multitude of molecules that differ in shape and interconvert with each other spontaneously. The molecules of an isomer are considered to be chemically identical.[64] The (stereo-)chemical features of an isomer are determined by the properties of all molecules. This is the reason why geometrically-oriented approaches cannot give an adequate description of stereochemical phenomena; rather they are representing a chemical compound by a single rigid molecular model.

Stereochemistry is the science of the three-dimensional molecular structure and the observable consequences thereof. It comprises two aspects: the static aspect of stereochemistry deals with the number and kind of different stereoisomers that exist under given observation conditions. This aspect of stereochemistry is closely related to its dynamic aspect, dealing with the reactions in which the stereoisomers are formed or destroyed. The observation conditions determine which reactions or molecular motions occur and therefore which molecules are chemically identical under these conditions.

The theory of the chemical identity group[64] is based on the principle of permutational isomerism.[65] It allows the uniform treatment of static and dynamic stereochemical problems concerning both rigid and flexible molecules. In order to determine the stereoisomers and their mutual interconversions, the theory takes into account both internal and external molecular motions as well as isomerization mechanisms occurring under given observation conditions. No symmetry oriented geometric considerations are made. Each chemical object is assigned an adequate group theoretical counterpart by the theory of the chemical identity group. Thus, group theory and related fields of abstract algebra may be used for the description of relations between these objects and for the development of algorithms for the solution of modeled stereochemical problems.

Molecules are represented by *permuted models*.[64] The *reference model E* of a molecule *m* results from the imaginary dissection of *m* into a skeleton *s* and a set *L* of ligands. Ligands are those parts of the molecule which can be imaginarily rearranged on the skeletal sites. The whole set of permuted models produced by application of ligand permutations to the reference model *E* is called the *family of permuted models P*(*E*). If all the ligands are chemically distinguishable, the set of all permutations yielding chemically identical molecular models, when applied to the reference model, forms an algebraic group. This is the *chemical identity group*[64] of the reference model. Therefore, *permutational isomers*[65] are represented by equivalence classes of chemically identical molecular models. These classes are induced in the family of permuted models by the cosets of the chemical identity group. Permutational isomers can be stereoisomers or constitutional isomers.[64]

The *reaction schemes*[64] which are based on set valued maps are used to take into account the influence of ligand equivalences on the existing permutational isomers or in order to trace the mutual interconversions of permutational isomers through ligand preserving isomerizations: cosets of the chemical identity group representing potentially different permutational isomers merge into a single set of permutations due to an isomerization preserving the chemical identity of a molecular model or due to ligand equivalences.

In general the set of all permutations preserving the chemical identity of the reference model does not form an algebraic group, even if all ligands are chemically distinguishable.[66,67] For example, there may be intramolecular motions preserving the chemical identity of a molecule that are not expressible in terms of permutational groups. The reaction schemes mentioned above are limited to the combination of equivalence information available in the form of coset spaces of permutational groups, such as chemical identity groups or *ligand equivalence groups*.[64] But equivalence information on the elements of a family of permuted models may be represented by arbitrary equivalence class spaces, not just coset spaces. An universally applicable method to determine the set of all chemically identical molecular models in a family of permuted models is the *equivalence accumulation*.[66] This method can be regarded as a generalization of the reaction schemes, being suited for the combination of arbitrarily chosen equivalence class spaces.

The result of an equivalence accumulation of one equivalence class space with respect to another is yet another equivalence class space representing the logical disjunction of the two input equivalence relations. As the chemical identity of molecules depends on the observation conditions via both isomerization processes and molecular motions occurring under these conditions, the equivalence accumulation allows one to investigate the influence of the observation conditions on the existing permutational isomers.

The most important equivalence relation in a family of permuted models $P(E)$ arises from the notion of chemical identity. The *chemical identity class* $\mathrm{Id}(E)$ contains all permuted models in $P(E)$ that represent the same isomer as the reference model E. Analogously, the identity class $\mathrm{Id}(\lambda E)$ is composed of those molecules that are chemically identical to permuted model λE. The identity classes $\mathrm{Id}(\lambda E)$ and $\mathrm{Id}(\mu E)$ of any two permuted models λE and μE are either identical or disjoint, so the set of all different chemical identity classes is an equivalence class space in $P(E)$. It is called the *identity class space* $\mathrm{IdC}(E) = \{\mathrm{Id}(E), \mathrm{Id}(\lambda_1 E), \mathrm{Id}(\lambda_2 E), ..., \mathrm{Id}(\lambda_n E)\}$.

Another important equivalence relation is stereoisomerism. Two molecules are considered to be stereoisomeric if they have chemically equivalent constitutions.[67] A chemically meaningful definition of stereoisomers can be based on the definition of stereoisomeric molecules as follows:[66] Two isomers are considered as stereoisomers if there is at least one pair of stereoisomeric molecules, which contains one molecule from each isomer. The corresponding equivalence class space is called the *stereoisomerism class space* $\mathrm{StC}(E)$. It is obtained by performing an equivalence accumulation of the identity class space $\mathrm{IdC}(E)$ with respect to the *connectivity class space*.[66] The connectivity class space is made up of classes of permuted models which are interconvertible by permutations interchanging ligands within particular skeletal sites. Thus, the molecular models of one connectivity class are equivalent with respect to their chemical constitution.

The number of permuted models in a family $P(E)$ equals the number of distinguishable ligand redistributions on the skeletal sites. As the cardinality of a family of permuted models increases sharply with the number of ligands—for a permuted model with n ligands there are $n!$ distinguishable ligand redistributions—suitable methods for the structuring of this large amount of data have to be used in order to be able to handle it. The family of permuted models $P(E)$ is isomorphic to the symmetric group of permutations S_n, n being the number of ligands of the reference model E. Therefore, one can take advantage of group theoretical methods in order to structure $P(E)$.[66]

Chemically identical molecules interconvert spontaneously via molecular motions, such as external rotations and intramolecular motions. These ligand independent motions[66] of reference model E can be expressed in terms of the chemical identity group $S(E)$ of E which is a subgroup of S_n. The notion of the chemical identity group has slightly changed. It is no longer defined as the set of *all* permutations that preserve the chemical identity of E, but as the set of permutations representing ligand independent molecular motions.[66]

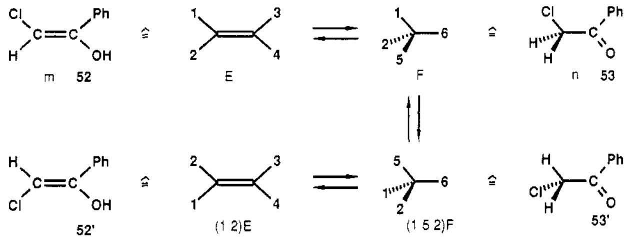


Figure 9. Interconversion of the Z- and E-forms of an enol via tautomerization.

The left coset space of $S(E)$ in S_n induces an equivalence class space in $P(E)$. Another equivalence class space in $P(E)$ is induced by the right coset space of the ligand equivalence group $\Sigma(L)$. If L is the ligand set of reference model E, then $\Sigma(L)$ consists of all those permutations that interchange chemically equivalent ligands on E. Consequently the application of $\Sigma(L)$ on E results in a set of molecular models chemically identical to E. If an equivalence accumulation of the left coset space of $S(E)$ in S_n with respect to the right coset space of $\Sigma(L)$ is carried out, the identity class space $\mathrm{IdC}(E)$ is obtained, provided that under the given observation conditions there are no further causes for the chemical identity of molecular models than ligand independent motions and ligand equivalences. In these cases, subsequent equivalence accumulations with respect to the corresponding equivalence class spaces have to be performed in order to obtain the identity class space $\mathrm{IdC}(E)$. The identity class space $\mathrm{IdC}(E)$ may also be affected by isomerization processes between $P(E)$ and another family $P(F)$. For example, the Berry pseudorotation may be modeled as a ligand preserving isomerization process between a family $P(E)$ of permuted models with a trigonal-bipyramidal skeleton and a family $P(F)$ with a quadratic-pyramidal skeleton.[64]

Modeling of chemical reactions which are not ligand preserving demands another approach: relations between two families of permuted models (with different sets of ligands) are defined with the help of *filters*.[68] These predicates require the skeletal sites of both the molecular models of the educt and the product of a chemical reaction to be occupied in a certain manner by ligands of the respective ligand set. Thus, requirements on the (physico-)chemical nature of educt and product of a reaction can be expressed. The ligand rearrangement specified by a filter is also part of the representation of the reaction mechanism. Each pair $(\lambda E, \mu F)$ of permuted models that satisfies the requirements imposed by a filter is selected from the families $P(E)$ and $P(F)$. It represents the chemical reaction of an educt represented by molecular model λE to a product represented by μF.

The following example may serve to illustrate the application of filters in the modeling of the stereochemical course of reactions.[66] It will be proved formally that the Z- and E-forms of an enol (**52** and **52′** in Figure 9) are chemically indistinguishable under certain observation conditions and therefore belong to the same isomer.

The first step is to create a reference model of the enol molecule m suitable to reflect its relevant stereochemical features, that is the Z/E-isomerism at the C=C double bond. Molecule m is conceptually dissected into the skeleton s_m and the ligand set $L_m = \{1, 2, 3, 4\}$, where 1 = Cl, 2 = H, 3 = Ph, and 4 = OH (Figure 10).

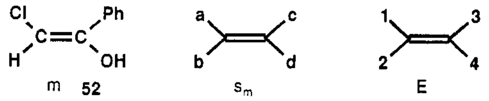


Figure 10. Reference model of the enol.

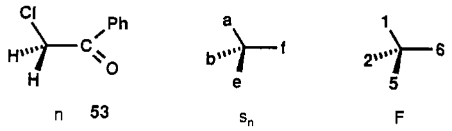


Figure 11. Reference model of the ketone.

The chemical identity group $S(E) = \{(\ ), (1\ 2)(3\ 4), (1\ 3)(2\ 4), (1\ 4)(2\ 3)\}$ of the reference model $E = (m, s_m, L_m)$ results from the following ligand independent molecular motions: an external 180°-rotation (1 2)(3 4) about the axis determined by the double bond, an external 180°-rotation (1 3)(2 4) about the axis lying perpendicular to the double bond in the molecular plane, and an external 180°-rotation (1 4)(2 3) about the axis perpendicular to the molecular plane. There are no intramolecular motions preserving the chemical identity of E. The family $P(E)$ of permuted models consists of $4! = 24$ permuted models. The cardinality of the identity group being 4, these models belong to six different identity classes: $\mathrm{IdC}(E) = \{S(E)E, (1\ 2)S(E)E, (1\ 3)S(E)E, (1\ 2\ 3)S(E)E, (1\ 4)S(E)E, (1\ 2\ 4)S(E)E\}$. The identity class $S(E)E$ represents the E-isomer **52** of the enol, whereas $(1\ 2)S(E)E$ represents the Z-isomer **52′**. We shall not be concerned here about the other identity classes, which represent the enolic forms of an acyl chloride ($(1\ 3)S(E)E$ and $(1\ 2\ 3)S(E)E$) and those of an aldehyde, respectively, ($(1\ 4)S(E)E$ and $(1\ 2\ 4)S(E)E$).

In a second step, a reference model for the ketone n is established. Molecule n is not very interesting from the stereochemical point of view. Only the carbon atom to which the chlorine atom is bonded could become asymmetric if it were substituted with other ligands. The details of the group R = COPh are not relevant to the modeling of the tautomerization, so the whole of R is abstracted into a single ligand. The reference model $F = (n, s_n, L_n)$ is depicted in Figure 11. The ligand set $L_n = \{1, 2, 5, 6\}$ (1 = Cl, 2 = H, 5 = H, 6 = R) comprises four ligands, just as L_m, but the tautomerization is not a ligand preserving process.

The chemical identity group $S(F) = \{(\ ), (1\ 2)(5\ 6), (1\ 5)(2\ 6), (1\ 6)(2\ 5), (2\ 5\ 6), (2\ 6\ 5), (1\ 5\ 6), (1\ 6\ 5), (1\ 2\ 6), (1\ 6\ 2), (1\ 2\ 5), (1\ 5\ 2)\}$ results from external 180°- and 120°-rotations. The two classes $S(F)F$ and $(1\ 2)S(F)F$ would correspond to the enantiomeric forms of the ketone, if all the ligands of L_n were chemically distinguishable. Because this is not the case here (ligands 2 and 5 both represent hydrogen atoms) the ligand equivalence group $\Sigma(L_n) = \{(\ ), (2\ 5)\}$ is not trivial. Performing an equivalence accumulation of $\mathrm{SC}(F)F = \{S(F)F, (1\ 2)S(F)F\}$ with respect to $\Sigma\mathrm{C}(L_n)F$, the equivalence class space induced in $P(F)$ by the right coset space of $\Sigma(L_n)$, one obtains the identity class space $\mathrm{IdC}(F) = \{\Sigma(L_n)\ S(F)F\} = \{P(F)\}$. There is only one isomer of the ketone, as the permuted models F from $S(F)F$ and $(2\ 5)F$ from $(1\ 2)S(F)F$ differ only by a permutation of chemically equivalent ligands. Thus, they represent chemically identical molecules and the classes $S(F)F$ and $(1\ 2)S(F)F$ merge into a single identity class.

Now the information represented by the two identity class spaces $\mathrm{IdC}(E)$ and $\mathrm{IdC}(F)$ has to be combined with information about the tautomerization process. The isomerization process depends on the ligand distribution on the skeletal sites. A molecular model from $P(F)$ tautomerizes to a molecular model from $P(E)$ if skeletal site c is occupied by a ligand representing Ph, skeletal site d by a ligand representing OH, e by a ligand representing H, and f by a ligand representing R. Thus the filter for the tautomerization can be established as $\Phi = (c = \mathrm{Ph} \wedge d = \mathrm{OH} \wedge e = \mathrm{H} \wedge f = \mathrm{R})$. Because of Φ, only those classes of $\mathrm{IdC}(E)$ that contain molecules of the form E_Φ and classes from $\mathrm{IdC}(F)$ containing molecules of the form F_Φ are relevant for the tautomerization (Figure 12).

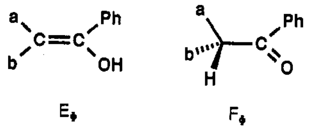


Figure 12. Prototypes of permuted models satisfying the requirements of the filter Φ.

As can be seen in Figure 9, the pairs (E, F) and $((1\ 2)E, (1\ 5\ 2)F)$, selected from $(P(E), P(F))$ by the filter Φ, represent tautomerization reactions. The molecular models F and $(1\ 5\ 2)F$ are connected via an external rotation, so there is a chain of isomerizations $E \leftrightarrow F \leftrightarrow (1\ 5\ 2)F \leftrightarrow (1\ 2)E$ through which the Z- and the E-forms of the enol interconvert spontaneously with each other. Thus it has formally been proven that under observation conditions allowing tautomerization, the (potential) Z- and E-isomers of the enol cannot be chemically distinguished. Therefore, only one isomer exists.

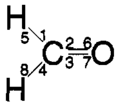
The group theoretical methods of the "classical" theory of the chemical identity group[64] in conjunction with the predicate logical specification of chemical reactions[68] make possible the universal and uniform modeling of stereochemical problems and their solution by algebraic means.

6. LINEARIZED AND HIGHLY STRUCTURED REPRESENTATION OF CONFORMATIONAL AND CONFIGURATIONAL INFORMATION

An interface between the formal descriptions of stereochemistry and constitutional chemistry has been introduced recently: the s- and r-vectors[67] constitute a new level of representation methodology for molecular systems. The algebra of s- and r-vectors covers constitutional aspects as well as stereochemical aspects and even the dynamics of molecules. They seem to be ideally suited as the basis for a data structure for existing and future computer support within the area of synthetic chemistry.

6.1. What to Represent. The representation of chemical data is the basis for any kind of computer support within the area of synthetic chemistry. The chemical data of a chemical system are as follows: atoms which are involved; information about the properties of atoms within the chemical system ("what kinds of bonds are valid", "how many valence electrons of an atom take part in the system"); relations between atoms, generally known as "chemical bonds", but also configurational and some conformational information ("what atoms are close"); information about the dynamics of the atoms respectively atomic groups, i.e., information about intramolecular rotations and intermolecular transitions; information about the dynamics of the chemical system, i.e., information about the conditions under which the relations between the atoms may change. Such a representation should be unique and unambiguous, that means, e.g., only one representation exists for a molecule and two molecules may not be represented in the same way.

The *be*-matrices do not meet these demands, for any permuted *be*-matrix represents the same molecule; stereoisomers on the other hand may not be distinguished because they are represented by the same *be*-matrix. Using group theoretical information about the *be*-matrices (*be*-matrices



54

Figure 13. Formaldehyde with numbered topologic positions.

Atom	C				H	O		H
A(A)	1	2	3	4	5	6	7	8
1					+			
2						+		
3							+	
4								+
5	+							
6		+						
7			+					
8				+				

Figure 14. Stereochemical *xbe*-matrix for formaldehyde.

form a free abelian group), an unique normal form could be found. However, the problem of the ambiguity would not be solved.

6.2. Stereochemical *be*- and *r*-Matrices. By the approach of Gruber[67] the *be*- and *r*-matrices were topologically enhanced to *stereochemical be*- and *r-matrices* and reduced to vectors afterward. Instead of the quality of the relations between individual atoms the stereochemical *be*- and *r*-matrices provide information for each topologic position, as shown in the following example:

Formaldehyde **54** has eight topologic positions corresponding to the binding relationships; four of them are affiliated to the carbon atom, two to the oxygen atom, and the remaining to the hydrogen atoms (Figure 13).

In a *be*-matrix each column corresponds with one atom, whereas in the *stereochemical be*-matrix each column corresponds to exactly one topologic position. Classic covalent bonds are represented by a relation between two positions. In Figure 14 these relations are marked with a "+". In the case of covalent bonds the topologic positions may be interpreted as electrons which are affected by a chemical bond and build a pair.

Analogous to the *r*-matrices there exist *stereochemical r-matrices* which are combined with the stereochemical *be*-matrices via the equation B R = E. The operation denotes the "binary exclusive or" which is a very simple operation. However, the composition of two $n \times n$-matrices will always be a polynomial complex problem, regardless how many entries are *undefined*, *nonzero*, or *nonfalse*, respectively, unless the matrices have properties or inherent structures which allow data reduction via their semantics. Such properties of stereochemical *be*- and *r*-matrices are the reason why these matrices can be compressed to vectors as described in the following.

6.3. Algebra of *s*- and *r*-Vectors. With a closer look at any *stereochemical be*-matrix one recognizes that *exactly one* position is in a symmetric relation with *exactly one* other. This is true for all topologic positions that are part of a covalent bond. A nonbonded topologic position is in relation with itself but also only in relation with one topologic position. Therefore, this relation is an automorphism (a bijective mapping of a set onto itself) on the finite set of topologic positions. In consequence, the relation can be formulated as a permutation.

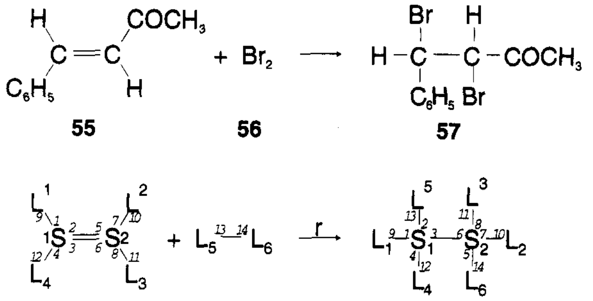


Figure 15. Addition of bromine to an ethane derivative.

For example (1 5)(2 6)(3 7)(4 8) represents **54**. Noted as a vector it is [5 6 7 8 1 2 3 4]. This vector is called the *s*-vector (*s*tereochemical *vector*) for **54**.

A chemical reaction is a redistribution of valence electrons. In terms of permutational vectors it is a permutation of topologic positions, represented by a *r*-vector (*r*eaction *vector*). A *r*-vector is combined with a *s*-vector via the functional composition $R \circ B = E$. This is a very basic operation with linear complexity to the length of the vector. A linear space complexity is obtained as well. For representing full chemical information of a chemical system this is the least complexity one may expect. The costs for composing two vectors can only be reduced by restricting the set of permuting entities which corresponds to a shortening of the vector. Usually this set is determined by the "reactivity" of the topologic positions of the chemical system, as shown in the illustrative example of the addition of bromine to an ethane derivative.[69] This example also shows the capabilities of the algebra of *s*- and *r*-vectors in representing constitutional information as well as configurative information.

The molecules **55** and **56** react to molecule **57**. The 14 interesting topologic positions within this chemical system were numbered as shown in Figure 15. Assuming the vector [1 2 3 4 5 6 7 8 9 10 11 12 13 14] to be the *s*-vector representing the set of totally unbonded topologic positions, the *s*-vector η = [9 5 6 12 2 3 10 11 1 7 8 4 14 13] represents the educt ensemble **55** + **56**. The addition of **56** corresponds to the breaking of the bond between the two bromine atoms—(13 14)—the breaking of the π-bond—(2 5)—and the taking up of the bromine atoms by the carbon atoms—(5 14) and (2 13). Composing these permutations, the *r*-vector ρ = (2 13)(5 14)(2 5)(13 14) = (2 14)(5 13) = [1 14 3 4 13 6 7 8 9 10 11 12 5 2] results. The product **57** is represented by the composition of the *r*-vector and the *s*-vector: $\pi = \rho \circ \eta$ = [1 14 3 4 13 6 7 8 9 10 11 12 5 2] $\circ$ [9 5 6 12 2 3 10 11 1 7 8 4 14 13] = [9 13 6 12 14 3 10 11 1 7 8 4 2 5].

The representation of one of the stereoisomers of **57** is obtained by applying the *r*-vector ι = (2 9)(1 13)(1 9)(2 13) = (1 2)(9 13) = [2 1 3 4 5 6 7 8 13 10 11 12 9 14] on the *s*-vector π: $\sigma = \iota \circ \pi$ = [13 9 6 12 14 3 10 11 2 7 8 4 1 5].

The algebra of *s*- and *r*-vectors unifies the theory of the chemical identity group and the algebra of *be*- and *r*-matrices. For the class of *s*-vectors which fulfill the property that each topologic position of the skeleton is assigned exactly to one ligand, the main theorems of the theory of the chemical identity group are valid as well, due to an isomorphism as proven by Gruber.[67] The substitution of the atomic symbols by the letters S and L in Figure 15 give a hint for a partition in terms of the theory of the chemical identity group.

The biggest advantage of the algebra of *s*- and *r*-vectors over the algebra of *be*- and *r*-matrices or any other existing representation is the inherent structure which is entirely based on group theory: If configurational aspects can be neglected, the automorphism group of the topologic positions may be factorized by the subgroup of all configurational permutations. The set of permutations of constitutionally equivalent atoms form another normal subgroup.[67] There also exists a subgroup which corresponds to the chemical identity group.[67] It contains all permutations which represent ligand-independent motions, such as rotations around a single bond, etc.

By the structuring effect of such relations (and intense analysis of the chemistry of the underlying chemical system), a chemist should obtain clear, consistent, unredundant, and sound information about the chemical system under investigation. And what a chemist appreciates: each mathematical object has a chemical interpretation or even a chemical counterpart.

7. CLASSIFICATION, CANONICAL NUMBERING, AND STRING NOTATION OF REACTIONS

The DU-theory can be used to create hierarchic order in the set of all chemical reactions. This is not only well-suited as a basis of documentation systems for chemical reactions, but is also useful in the algorithmic generation and the systematic computer-assisted discovery of new reactions.

Classification schemes for reactions generally follow two different aspects. One places the main interest on reactants and products ("educt/product-oriented approach"). The other one views a reaction as a transformation of molecular graphs.

The educt/product approach can be quite successful for practical purposes, but it leads to a rather casuistic treatment of the problem and is thus less amenable to a generally applicable and hierarchically structured classification. As a corollary, applications of this approach usually distinguish between more or less "important" reaction partners; i.e., they distinguish between reactants and reagents, main products and byproducts, etc. This leads to the frequently observed omission of some structures from the description of reactions.[70–75]

Under the aspect of a reaction as a transformation of a molecular graph, the reaction is represented by an operator that performs this transformation. In accordance with chemical reality, that operator effects some redistribution of the valence atoms, i.e., a change of constitution; it possibly also describes a spatial rearrangement, i.e., a change of configuration or even conformation. (Note that the set of atoms of the molecular graphs will remain unchanged). We will—in the present context—deal only with the changes of constitution, i.e., with that part of the operator that effects the valence electron shift. The description of the reaction thus is a record of the redistribution of the electrons.

7.1. Reactions and Transformations. The distinction of the two approaches to descriptions of reactions has only been recently introduced into the chemical nomenclature.[76] "*A transformation must be distinguished from a reaction. The full description of a reaction would state or imply all the reactants used and all the products formed. In a transformation one is concerned only with changes in one particular species designated as the 'substrate'.*"—"*A transformation is distinguished from a reaction in that it describes only those changes that are involved in converting the structure of a substrate into that of a product.*"

The considerations in this section refer only to reactions. We have developed some programs that complement data collections of transformations into reactions, either by calculating stoichiometric factors for cases where all reactant and product structures are present[77,78] or by inferring missing structures by means of chemical heuristics.[79,80]

7.2. Valence Electron Redistribution Schemes. One of the earliest descriptions of electron shifts was presented by Vladutz. Neglecting topological aspects, he arrived at a description that summarizes the changes of bond orders on either side of the reaction equation.[7,81,82] Reaction-invariant bonds and atoms of the reaction core were part of the code which thus represented the RA-class (see section 7.3).

An early topological code was introduced by Arens.[83–86] It was limited to strictly linear or monocyclic SACOBOs (see section 7.4); carbon atoms were implied. Only three cases could be represented: even-numbered cyclic shifts, odd-numbered cyclic shifts, and even-numbered linear shifts. Reaction-invariant bonds were not coded. The codes thus represented certain subsets of R-categories.

A coding scheme proposed by Hendrickson[87–89] is applicable to a restricted subset of reaction types.

Zefirov proposes a formal-logic-based code[39,90,91] which seems to be widely applicable. Thus far no canonicalization has been described.

Fujita, in numerous papers,[73,92–100] proposes a coding scheme FORTUNITS, where bond rearrangement patterns for transformations are coded together with reaction invariant bonds and atoms of the (possibly incomplete) reaction core. The lack of a clear distinction of the hierarchical levels results in a highly complex set of rules and a difficult canonicalization. FORTUNITS thus is an (incomplete) code on the level of the RA-class.

None of the aforementioned coding schemes provide for a clear hierarchic structure that is based on a mathematical model and on a clear distinction of reaction-variant and -invariant bonds as one basis of the hierarchy.

7.3. Classification of Reactions by the DU-Model. A hierarchic classification of chemical reactions follows directly from the mathematical model of constitutional chemistry. It is based on a hierarchy of characteristic features by which equivalence classes of reactions can be established.[55,101]

By means of the irreducible *r*-matrix $\mathbf{R}^I$ we define the reaction core.[102] The atom vector $\mathbf{a}^I$ of the reaction core consists of those elements of the atom vector **a** that correspond to the rows/columns of the irreducible *r*-matrix. The reaction-invariant bonds between atoms *i* and *j* of $\mathbf{a}^I$ are represented by the elements b^I_{ij} of a matrix $\mathbf{B}^I$, the intact *be*-matrix.

A first level of classification is based on the chemical distance. Thereafter, all reactions that follow the same electron redistribution pattern are represented by the same irreducible *r*-matrix. They form a reaction category. A reaction category can be subdivided into RB-classes. An RB-class comprises all reactions of an R-category whose reactants and products contain the same scheme of reaction invariant bonds in the reaction core. The RA-classes can be further hierarchically subdivided into equivalence classes R1, R2, ..., R*n* according to the environments of the members of $\mathbf{a}^I$ in the 1st, 2nd, ..., *n*th sphere of the a^I_i in the molecular graphs of either side of the reaction equation. Presupposing a canonical numbering of $\mathbf{a}^I$, we thus give the following definitions:

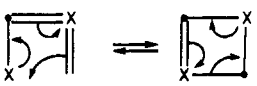
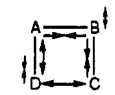
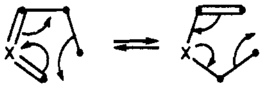
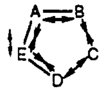
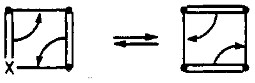
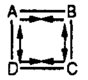
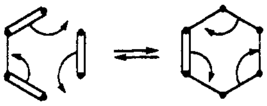
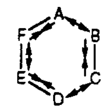
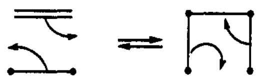
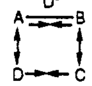
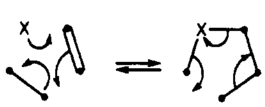
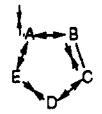
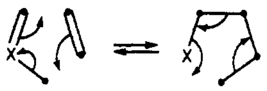
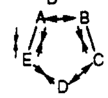
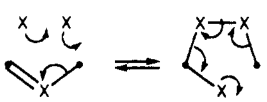
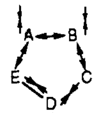
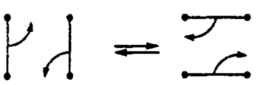
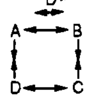
(a) Reactions with the same chemical distance *D* form an equivalence class of reactions, the RD-category.

(b) Reactions whose *r*-matrix leads to the same $\mathbf{R}^I$ form an equivalence class of reactions, the R-category.

(c) Members of an R-category which have the same intact *be*-matrix $\mathbf{B}^I$ form an RB-class.

(d) Members of an RB-class which have the same $\mathbf{a}^I$ form an equivalence class, the RA-class.

Table 3. Some Examples of Reaction Schemes and String Notations[a]

no.	reactn scheme[b]	scheme of the reactn core	direcn	c	canonical R-string (full notation)	short notation	Zefirov's SEQ-notation
1			→ ←	 C	11.1.1/20.2(1101)BADC 11.1.1/20.2(1110)DABC		[4$\alpha\gamma$]
2			→ ←	C 	11111/2(11001)EABCD 1111.1/0000.2(10011)ABCDE	Z5A –Z5A	[1,4γ] – [1,4β]
3			→ ←	 C	1111(1111)ABCD 1111(1111)ADCB	Z4 Z4	
4			→ ←	 C	111111(111010)AFEDCB 111111(110101)EFABCD	Z6 Z6	[2+4]
5			→ ←	C 	1111(1000)ABCD 1111(0100)DABC	Z4 Z4	[2+(1,1)]
6			→ ←	 C	1111.1/0000.2(10000)BCDEA 11111/2(01000)ABCDE	–Z5A Z5A	[1 + 2 + (1,1)]
7			→ ←	C 	11111/2(10100)EABCD 1111.1/0000.2(01010)DCBAE	Z5A –Z5A	[2 + (1,2α)]
8			→ ←	C 	11010;00.1.10/20.20.2(10000)DEACB 110.10;00.1.10/20.22(01000)AEDBC		[1 + 1 + (1,2α)]
9			→ ←	 C	1111( )ABCD 1111( )ADCB	Z4 Z4	[(1,1) + (1,1)]

[a] The schemes of the reaction cores are drawn with the symbols for bonds broken or disappearing free electron pairs (→←) and for bonds made or newly appearing free electron pairs (←→) as proposed by Vladutz.[81,82] Zefirov's[91] SEQ notation is given for comparison. [b] These are the schemes shown in Table 1 of ref 92. [c] The canonical *reference direction* is marked by a "C" in this column.

(e) Members of an RA-class for which the first neighbor sphere of each member of $\mathbf{a}^{\mathbf{I}}$ is the same, form an equivalence class, the R1-class.

(f) (Recursively while the number of members of subsequent classes decreases:) members of an R(n–1)-class for which the nth neighbor sphere of each member of $\mathbf{a}^{\mathbf{I}}$ is the same on either side of the reaction equation, form an equivalence class, the Rn-class.

This classification can be carried forward to the extended form of the *r*- and *be*-matrices, i.e., the *xr*- and *xbe*-matrix formalism.[103,104]

7.4. Canonicalization and String-Notation. We have defined a canonicalization of the reaction description on the basis of a lexicographical ordering of a suitably defined concatenation of elements of $\mathbf{R}^{\mathbf{I}}$, $\mathbf{B}^{\mathbf{I}}$, $\mathbf{a}^{\mathbf{I}}$, and the spheres around each $a^{I}{}_{i}$.

We observe that in many (*but not all!*) cases, a bond rearrangement follows a *SACOBO* (sequence of alternating changes of bond orders). We decided to employ this heuristic in order to create both a more natural looking numbering and a particularly facile computation of that numbering.[105,106] It should be emphasized, however, that the SACOBO is *not* the mathematical basis of the canonicalization as, e.g., in Arens' coding scheme.[84–86]

The canonicalization rule, in essence, emphasizes maximum SACOBOs. Numerically, this corresponds to maximizing the concatenation of subsequent off-diagonals of R^{I}, read with alternating signs.[55,101,107] In order to do this, we need a rule for the ordering of the values of the entries of the matrices. In the case of *r*- and *be*-matrices, this ordering is defined by the natural ordering of the integers. For the *xr*- and *xbe*-matrices, an ordering of the constitution-changing operators has to be defined explicitly by chemical considerations.[108]

From this concatenation we have derived a mathematically defined ("unprejudiced"), universally applicable string notation for chemical reactions (within the boundaries of the DU-model). The sections of these R-strings are concatenated according to their hierarchic level. Right truncation of R-strings leads to increasingly more general notations.

The R-strings, thus provide for a nested lexicographical ordering of reactions by classes within classes and categories. This generating of new reactions by path finding algorithms[109,110] again facilitates discovery of new similarities within sets of reactions that hitherto were considered unrelated, and finally leads to the discovery of new reaction types.[111]

As earlier observed by Arens,[83–86] the overwhelming majority of organic reactions is representable by simple linear or circular bond and electron shifts. For these cases, we have proposed a concise notation,[55] which we meanwhile have slightly modified, particularly with regard to delocalized systems and shifts of free electrons.[104] Linear shifts are denoted as K-types, circular shifts as Z-types. Metathesis reactions A—B + C—D → A—D + B—C, are formally—irrespective of mechanism—regarded as a circular shift of bonds around a four-membered cycle; the notation for this is Z4. Diels–Alder reactions, Cope rearrangement, and similar electrocyclic reaction types are Z6. The canonicalization makes free electrons either appear (or disappear) at

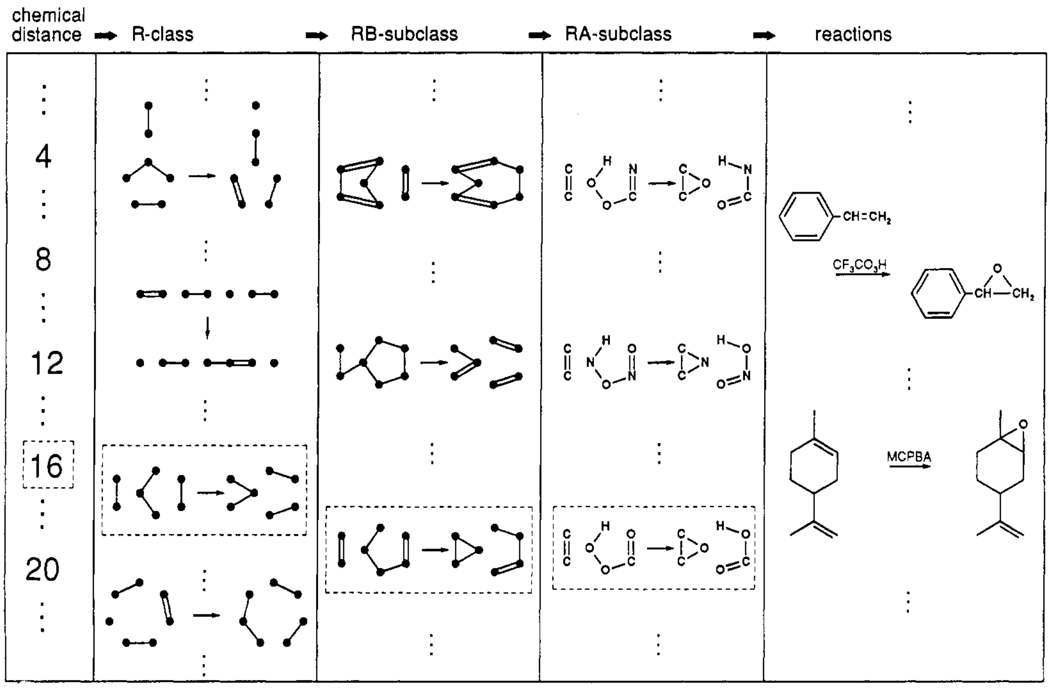


Figure 16. Hierarchic classification of chemical reactions.

the end of a chain or in the 1-position of a ring. They are indicated by appending an "a" (for a single electron) or an "A" (electron pair). Or else, they migrate from somewhere down the chain or ring to the 1-position; then the alphabetical character corresponding to the origin of that migration is appended. The allyl rearrangement of a radical, A═B—C• → •A—B═C, is denoted by K3c. Table 3 gives examples of such string notations (together with Zefirov's notations from Table 1 in ref 91). Integers representing the reaction-invariant bonds may then be appended and, after these, the list of element symbols of the reaction core.

7.5. Documentation of Reactions in the CASTOR System. The linear notation is well suited to construction of a reaction database by using a conventional database management system. The first implementation of CASTOR was installed on a CDC mainframe using CDC's EDMS (Evolutionary Database Management System) as the DBMS and FORTRAN programs for I/O, canonicalization, and interfacing to the DBMS.[79,105,112,113]

The database CASTOR is organized into three sections. Reaction types are stored as segmented R-strings, i.e., as concatenations of the substrings of the R-category, the RB-class, and the RA-class. Molecules are stored in separate files. For each individual reaction, a reaction document is stored, where pointers connect the code of the reaction type with the reaction cores of molecules that are reactants and products of the reaction. Other keys point to files that contain associated, e.g., bibliographic or numerical information.

Searching can be done on the various levels of the hierarchical classification. It is possible to find all reactions which belong to one R-category or to retrieve all reactions which belong to the same RB-class or to the same RA-class. The query for an RB-class is subdivided into two steps: invariant *bonds* and invariant *free electrons*. Specifying the invariant free valence electrons implies a restriction of the choice of chemical elements of the $\mathbf{a}^{\mathrm{I}}$ core. Therefore the user may specify the invariant bonds alone or the invariant bonds together with the invariant free electrons. Since the latter appear at the very end of the linear notation of the RB-class, their omission corresponds to a right truncation of the RB-part of the R-string.

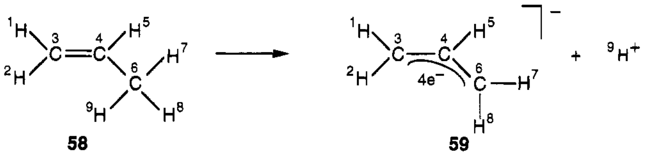


Figure 17. Formation of the allylic system.

A new version of the documentation system is now under development. It is a portable implementation for IBM-compatible PCs, using a commercially available relational DBMS.

7.6. IGOR and the Hierarchic Classification of Chemical Reactions. When used for the generation of unprecedented chemical reactions, the program IGOR proceeds down this hierarchy. Starting from a specific R-category (*r*-matrix) all RB-classes that belong to this R-category are generated. The user can select from these RB-classes those that should be expanded into the different RA-subclasses. Augmentation with peripheral substructures then leads to "real" chemical reactions. This last step is left to the user's chemical experience and intuition.

Figure 16 illustrates this generation process. Starting at the highest hierarchic level of chemical distance in each column, one representative selected member is emphasized by a dashed rectangle. For this selection some resulting subclasses are shown next to the right of the corresponding column.

8. TREATMENT OF DELOCALIZED ELECTRONS: EXTENDED *BE*- AND *R*-MATRICES

The representation of molecules with sets of delocalized electrons like multicenter bonds or organometallic compounds can be accomplished by extending the *be*- and *r*-matrices.

Conventional *be*- and *r*-matrices are perfectly suited for the representation of organic molecules and their reactions. The only limitation is that simple types of chemical bonds like single, double, or triple bonds are required. Difficulties arise if the considered EMs contain more complicated structural elements with electrons that are delocalized among several atomic cores. Aromatic structures may still be represented

$$
\begin{array}{l|ccccccccc|c|}
 & \mathrm{H} & \mathrm{H} & \mathrm{C} & \mathrm{C} & \mathrm{H} & \mathrm{C} & \mathrm{H} & \mathrm{H} & \mathrm{H} & \mathrm{D1} \\
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & \\
\mathrm{H}(1) & & & 1 & & & & & & & \\
\mathrm{H}(2) & & & 1 & & & & & & & \\
\mathrm{C}(3) & 1 & 1 & & 2 & & & & & & \\
\mathrm{C}(4) & & & 2 & & 1 & 1 & & & & \\
\mathrm{H}(5) & & & & 1 & & & & & & \\
\mathrm{C}(6) & & & & 1 & & & 1 & 1 & 1 & \\
\mathrm{H}(7) & & & & & & 1 & & & & \\
\mathrm{H}(8) & & & & & & 1 & & & & \\
\mathrm{H}(9) & & & & & & 1 & & & & \\
\hline
\mathrm{D1} & & & & & & & & & &
\end{array}
+
\begin{array}{l|ccccccccc|c|}
 & \mathrm{H} & \mathrm{H} & \mathrm{C} & \mathrm{C} & \mathrm{H} & \mathrm{C} & \mathrm{H} & \mathrm{H} & \mathrm{H} & \mathrm{D1} \\
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & \\
\mathrm{H}(1) & & & & & & & & & & \\
\mathrm{H}(2) & & & & & & & & & & \\
\mathrm{C}(3) & & & & -1 & & & & & & +1 \\
\mathrm{C}(4) & & & -1 & & & & & & & +1 \\
\mathrm{H}(5) & & & & & & & & & & \\
\mathrm{C}(6) & & & & & & & & & -1 & +1 \\
\mathrm{H}(7) & & & & & & & & & & \\
\mathrm{H}(8) & & & & & & & & & & \\
\mathrm{H}(9) & & & & & & -1 & & & & \\
\hline
\mathrm{D1} & & & +1 & +1 & & +1 & & & & +4
\end{array}
\longrightarrow
\begin{array}{l|ccccccccc|c|}
 & \mathrm{H} & \mathrm{H} & \mathrm{C} & \mathrm{C} & \mathrm{H} & \mathrm{C} & \mathrm{H} & \mathrm{H} & \mathrm{H} & \mathrm{D1} \\
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & \\
\mathrm{H}(1) & & & 1 & & & & & & & \\
\mathrm{H}(2) & & & 1 & & & & & & & \\
\mathrm{C}(3) & 1 & 1 & & 1 & & & & & & 1 \\
\mathrm{C}(4) & & & 1 & & 1 & 1 & & & & 1 \\
\mathrm{H}(5) & & & & 1 & & & & & & \\
\mathrm{C}(6) & & & & 1 & & & 1 & 1 & & 1 \\
\mathrm{H}(7) & & & & & & 1 & & & & \\
\mathrm{H}(8) & & & & & & 1 & & & & \\
\mathrm{H}(9) & & & & & & & & & & \\
\hline
\mathrm{D1} & & & 1 & 1 & & 1 & & & & 4
\end{array}
$$

Figure 18. *xbe*- and Matrix and *xr*-matrix for the formation of the allylic system.

by sets of resonance structures, according to VB-theory. Of course, this is not very elegant. For multicenter bonds and organometallic compounds, this approach is certainly not adequate.

Therefore we developed an extension of the *be*- and *r*-matrices.[114] The resulting *xbe*-and *xr*-matrices represent localized bonds and electrons in the same way as conventional *be*- and *r*-matrices. Systems of delocalized electrons (*de*-systems) are accounted for in extra rows/columns of the *xbe*-matrix.

If an EM contains n atoms and m *de*-systems, its *xbe*-matrix consists of a $n \times n$ *be*-part and m additional *de*-rows/-columns. The *be*-part reflects the covalent bonds between pairs of atoms and the distribution of free electrons. Each *de*-row/-column represents one *de*-system: the diagonal *de*-entry ($xb_{n+k,n+k}$; $1 \leq k \leq m$) is the number of delocalized electrons and the off-diagonal *de*-entries ($xb_{i,n+k} = xb_{n+k,i}$; $1 \leq k \leq m$, $1 \leq i \leq n$) symbolize whether atom i of the EM participates in the kth *de*-system ($xb_{i,n+k} = 1$) or not ($xb_{i,n+k} = 0$). The off-diagonal *de*-entries do not count toward the electron balance.

The conversion of one EM into another EM by a chemical reaction is represented by an extended reaction matrix. This symmetric $(n + m) \times (n + m)$ *xr*-matrix is elementwise added to the *xbe*-matrix of the educt-EM. Its nonzero entries are positive or negative integers. The entries $xr_{i,j}$ ($1 \leq i, j \leq n$) correspond to those of the *r*-matrices. They reflect the changes of the integral bond orders and changes of the numbers of free valence electrons at the atomic cores. Each diagonal *de*-entry $xr_{n+k,n+k}$ ($1 \leq k \leq m$) indicates the change of the number of delocalized electrons of the kth *de*-system. The off-diagonal *de*-entries $xr_{i,n+k} = xr_{n+k,i}$ may be +1, -1, or 0. This depends on whether i gains (+1) or loses (-1) its membership in the kth DE-system.

9. NEW DATA STRUCTURE FOR THE REPRESENTATION OF CHEMICAL CONSTITUTION

After the introduction of the *xbe*- and *xr*-matrices, the implementation of new data structures for molecules and their reactions followed suit. These data structures are based on dynamic lists. This allows a rapid processing of the chemical information.

'Each atom and each bond of the EM is represented by its own descriptor. It contains the essential information. An atom is characterized by its atomic number, its number of free valence electrons, and its localized charge. A bond descriptor always refers to a connection between a pair of atoms. Simple chemical bonds are therefore representable by a single descriptor. The representation of *de*-systems always requires sets of descriptors. The descriptors for each pairwise connection that is part of the *de*-system are concatenated by special references. Each bond descriptor contains the type of the bond. This type may be one of the following symbolic values: ***single***, ***double***, ***triple***, ***quadruple*** for the simple bond types, *edsys* for *de*-systems with electron deficiency (multicenter bonds), *pisys* for aromatic *de*-systems or conjugated π-chains, and *coord* for the connections between a metal atom and the atoms of an organic ligand. In our article[114] more types of *de*-systems were introduced. However, this concept proved to be too complicated, because the distinction of these types required too much chemical knowledge and could not be done formally. Now the three *de*-types are defined as follows:

edsys: the number of electrons of the *de*-system is less than two per bond (bond in the sense of connection between two atoms); all electrons are considered to be delocalized; example, 2e3c-bonds in boron compounds

pisys: the total number of (σ- and π-) electrons is more than two and less than four per bond; only the π-electrons are considered to be delocalized; example, aromatic compounds

coord: indication of connections between a metal atom and the atoms of an organic ligand; this ligand must contain a ***pisys-de***-system or ***double/triple*** bonds; no electrons are considered as delocalized, the metal atom and the ligand retain their electrons; example, ferrocene.

If a descriptor is of a *de*-type, it must also specify the total number of delocalized electrons of the *de*-system, the number of atoms among which the electrons are distributed, and the delocalized charge of the *de*-system.

The lists of atomic descriptors and bond descriptors contain several cross-references, e.g., from an atom to its bonds, from a bond descriptor to the pair of atoms it connects, etc. Therefore there is no need to process the lists sequentially every time a specific piece of information is required.

10. PERSPECTIVES

The amount of collected chemical data, structures, and reactions continues to grow. Computer programs for the direct solution of chemical problems are making inroads into the routine of chemical research. As a consequence of their intrinsic logic, such programs generate formal solutions. In order to be of practical value, such solutions have to be evaluated against data that represent the body of empirical knowledge. Progress in storage, retrieval, and processing of data relies on a sound mathematical basis, as represented by our algebraic model of chemistry.

REFERENCES AND NOTES

(1) Wipke, W. T.; Heller, S. R.; Feldmann, R. J.; Hyde, E. *Computer Representation and Manipulation of Chemical Information*; Wiley: New York, 1974.

(2) Ash, J. E., Hyde, E., Eds. *Chemical Information Systems*; Ellis Horwood: Chichester, U.K., 1975.

(3) Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J. *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project*; McGraw-Hill: New York, 1980.

(4) Gray, N. A. B. *Computer-assisted Structure Elucidation 2*; Wiley-Interscience: New York, 1980.

(5) Lederberg, J. Topological Mapping of Organic Molecules. *Proc. Natl. Acad. Sci. USA* **1965**, *53*, 134–139.

(6) Vladutz, G.; Finn, K. A. *Proceedings of the Department of Mechanization and Automation of Information Work*: Academy of Sciences USSR: Moscow, 1960.

(7) Vladutz, G. E. Concerning one System of Classification and Codification of Organic Reactions. *Inf. Storage Retr.* **1963**, *1*, 117–146.

(8) Corey, E. J. General Methods for the Construction of Complex Molecules. *Pure Appl. Chem.* **1967**, *14*, 19–37.

(9) Corey, E. J.; Petersson, G. An Algorithm for Machine Perception of Synthetically Significant Rings in Complex Cyclic Organic Structures. *J. Am. Chem. Soc.* **1972**, *94*, 460–465.

(10) Corey, E. J.; Howe, W. J.; Orf, H. W.; Pensak, D. A.; Petersen, G. General Methods of Synthetic Analysis. Strategic Bond Disconnections for Bridged Polycyclic Structures. *J. Am. Chem. Soc.* **1975**, *97*, 6116–6124.

(11) Wipke, W. T.; Rogers, D. Artificial Intelligence in Organic Synthesis. SST: Starting Material Selection Strategies. An Application of Superstructure Search. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 71–81.

(12) Gelernter, H.; Sridharan, N. S.; Hart, A. J.; Yen, S. C.; Fowler, F. W.; Shue, H. J. The Discovery of Organic Synthetic Routes by Computer. *Top. Curr. Chem.* **1973**, *41*, 113–150.

(13) Gelernter, H. Realization of a Geometry Theorem Proving Machine. *Information Processing, Proceedings of the First International Conference on Information Processing UNESCO*; Oldenbourg: Münich, Germany, 1960.

(14) Gelernter, H.; Rose, J. R. *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.

(15) Hendrickson, J. B. A Systematic Characterization of Structures and Reactions for Use in Organic Synthesis. *J. Am. Chem. Soc.* **1971**, *93*, 6847–6862.

(16) Moreau, G. Masso, un programme d'aide à la synthèse organique, utilisant des demi-rèactions. *Nouv. J. Chim.* **1978**, *2*, 187–193.

(17) Hendrickson, J. B.; Toczko, A. G. Synthesis Design Logic and SYNGEN (Synthesis Generation) Program. *Pure Appl. Chem.* **1988**, *60*, 1563–1572.

(18) Hendrickson, J. B. The SYNGEN Approach to Synthesis Design. *Anal. Chim. Acta* **1990**, *235*, 103–113.

(19) Ugi, I.; Kaufhold, G. Stereoselektive Synthesen IV. Der Reaktionsmechanismus stereoselektiver Vierkomponenten-Kondensationen. *Liebigs Ann. Chem.* **1967**, *709*, 11–28.

(20) Goebel, M.; Ugi, I. O-Alkyl-1-aminoglucose-Derivate als chirale Amin-Komponenten von Peptidsynthesen mittels stereoselektiver Vierkomponenten-Kondensationen. *Synthesis* **1991**, 1095–1098.

(21) Klein, M.; Ugi, I. 2-(4-Toluenesulphonyl)-3-aryl-oxaziridines as Oxidizing Reagents for P(III) Compounds. *Z. Naturforsch.* **1992**, *47B*, 887–890.

(22) Ugi, I. A Novel Synthetic Approach to Peptides by Computer Planned Stereoselective Four Component Condensations of α-Ferrocenyl Alkylamines and Related Reactions. *Rec. Chem. Prog.* **1969**, *30*, 389–311.

(23) Ugi, I. The Potential of Four Component Condensations for Peptide Syntheses–A Study in Isonitrile and Ferrocene Chemistry as well as Stereochemistry and Logics of Syntheses. *Intra-Sci. Chem. Rep.* **1971**, *5*, 229–261.

(24) Corey, E. J.; Wipke, W. T. Computer-assisted Design of Complex Organic Syntheses. *Science* **1969**, *166*, 178–192.

(25) Corey, E. J.; Cheng, X.-M. *The Logic of Computer Synthesis*; Wiley: New York, 1989.

(26) Corey, E. J.; Long, A. K. Rubenstein, S. D. Computer-assisted Analysis in Organic Synthesis. *Science* **1985**, *228*, 408–418.

(27) Brownscombe, T. F. Computer-assisted derivation of possible unimolecular pericyclic reactions. Intramolecular allene/diene rearrangements of α-allenic alcohols. Ph.D. Thesis, Rice University, Houston, TX 1972; *Diss. Abstr. Int.* **1973**, *B34* (3), 1035. Avail. through University Microfilms, Ann Arbor, MI, Order No. 73-21, 537.

(28) Gasteiger, J.; Jochum, C. EROS, A Computer Program for Generating Sequences of Reactions. *Top. Curr. Chem.* **1978**, *74*, 93–126.

(29) 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.

(30) Ruch, E.; Hässelbarth, W.; Richter, B. Doppelnebenklassen als Klassenbegriff und Nomenklaturprinzip für Isomere und ihre Abzählung. *Theor. Chim. Acta* **1970**, *19*, 288–300.

(31) Ruch, E.; Hässelbarth, W. Classification of Rearrangement Mechanisms by Means of Double Cosets and Counting Formulas for the Numbers of Classes. *Theor. Chim. Acta* **1973**, *29*, 259–268.

(32) Ugi, I.; Stein, N.; Gruber, B. The Two Formal Languages of Chemistry—the Semantic and Syntax. *Eesti Tead. Akad.*, in press.

(33) Le Bel, J.-A. *Bull. Soc. Chim. Par.* [*N. S.*] ("*Sér. 3*") **1874**, *22*, 337.

(34) van't Hoff, J. H. *Voorstel tot uitbreiding der tegenwoordig in de scheikunde gebruikte struktuur-formules in de ruimte*: Greven: Utrecht, The Netherlands, 1874.

(35) Cayley, A. Ueber die analytischen Figuren, welche in der Mathematik Bäume genannt werden, und ihre Anwendung auf die Theorie chemischer Verbindungen. *Ber. Dtsch. Chem. Ges.* **1875**, *8*, 1056–1059.

(36) Balaban, A. T., ed. *Chemical Applications of Graph Theory*; Academic Press: London, 1976.

(37) Kvasnička, V.; Pospichal, J. Graph-theoretical Interpretation of Ugi's Concept of the Reaction Network. *J. Math. Chem.* **1990**, *5*, 309–322.

(38) Koča, J.; Kratochvil, M.; Kvasnička, V.; Matyska, L.; Pospichal, J. *Synthon Model of Organic Chemistry and Synthesis Design*; Lecture Notes in Chemistry 51; Springer-Verlag: Heidelberg, 1989.

(39) Zefirov, N. S.; Tratch, S. S. Symbolic equations and their applications to reaction design. *Anal. Chim. Acta* **1990**, *235*, 115–134.

(40) Pólya, G. Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. *Acta Math.* **1937**, *68*, 145.

(41) Kerber, A.; Thürlings, K.-J. Symmetrieklassen von Funktionen und ihre Abzähltheorie. *Bayreuther Math. Schr.* **1983**, *12*, 1–235.

(42) Ruch, E.; Ugi, I. Das stereochemische Strukturmodell, ein mathematisches Modell zur gruppentheoretischen Behandlung der dynamischen Stereochemie. *Theor. Chim. Acta* **1966**, *4*, 287–304.

(43) Ruch, E.; Ugi, I. The Stereochemical Analogy Model—A Mathematical Theory of Dynamic Stereochemistry. *Top. Stereochem.* **1969**, *4*, 99–125.

(44) Dugundji, J.; Ugi, I. An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. *Top. Curr. Chem.* **1973**, *39*, 19–64.

(45) Meyer, E. Vielseitige maschinelle Suchmöglichkeiten nach Strukturformeln, Teilstrukturen und Stoffklassen. *Angew. Chem.* **1970**, *82*, 605–611; Versatile Computer Techniques for Searching Structural Formulas, Partial Structures and Classes of Compounds. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 201–227.

(46) Spialter, L. The Atom Connectivity Matrix (ACM) and Its Characteristic Polynomial (ACMCP). *J. Chem. Doc.* **1964**, *4*, 261–269.

(47) Ugi, I.; Bauer, J.; Bley, K.; Dengler, A.; Dietz, A.; Fontain, E.; Gruber, B.; Herges, R.; Knauer, M.; Reitsam, K.; Stein, N. Die Computerunterstützte Lösung chemischer Probleme—eine neue Disziplin der Chemie. *Angew. Chem.* **1993**, *105*, 210–239; Computer-Assisted Solution of Chemical Problems—The Historical Development and the Present State of the Art of a New Discipline of Chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 201–227.

(48) Ugi, I.; Fontain, E.; Bauer, J. Transparent Formal Methods for Reducing the Combinatorial Abundance of Conceivable Solutions to a Chemical Problem–Computer-assisted Elucidation of Complex Reaction Mechanisms. *Anal. Chim. Acta* **1990**, *235*, 155–161.

(49) Fontain, E.; Bauer, J.; Ugi I. Computer Assisted Bilateral Generation of Reaction Networks from Educts and Products. *Chem. Lett.* **1987**, 37–40.

(50) Fontain, E.; Reitsam, K. The Generation of Reaction Networks with RAIN. 1. The Reaction Generator. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 97–101.

(51) Fontain, E. The Generation of Reaction Networks with RAIN. 2. Resonance Structures and Tautomerism. *Tetrahedron Comput. Methodol.* **1990**, *3*, 469–477.

(52) Bauer, J.; Herges, R.; Fontain, E.; Ugi, I. IGOR and Computer Assisted Innovation in Chemistry. *Chimia* **1985**, *39*, 43–53.

(53) Bauer, J. IGOR2: A PC-Program for Generating New Reactions and Molecular Structures. *Tetrahedron Comput. Methodol.* **1989**, *2*, 269–280.

(54) Ugi, I.; Brandt, J.; Friedrich, J.; Gasteiger, J.; Jochum, C.; Lemmen, P.; Schubert, W. The Deductive Solution of Chemical Problems by Computer. Programs on the Basis of a Mathematical Model of Chemistry. *Pure Appl. Chem.* **1978**, *50*, 1303–1318.

(55) Brandt, J. Ein mathematisch begründetes hierarchisches Ordnungssystem chemischer Reaktionen und dessen theoretische und praktische Anwendung. Habilitationsschrift, Technische Universität München, 1981.

(56) Jochum, C.; Gasteiger, J.; Ugi, I. Das Prinzip der minimalen chemischen Distanz (PMCD). *Angew. Chem.* **1980**, *92*, 503–513; The Principle of Minimum Chemical Distance. *Angew Chem., Int. Ed. Engl.* **1980**, *19*, 495–505.

(57) Wochner, M.; Brandt, J.; v. Scholley, A.; Ugi, I. Chemical Similarity, Chemical Distance, and Its Exact Determination. *Chimia* **1988**, *42*, 217–225.

(58) Fontain, E. The Problem of Atom-to-Atom Mapping. An Application of Genetic Algorithms. *Anal. Chim. Acta* **1992**, *265*, 227–232.

(59) Fontain, E. Application of Genetic Algorithms in the Field of Constitutional Similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 748–752.

(60) Uma, R.; Swaminathan, S.; Rajagopalan, K. Base-Catalyzed Rearrangement of Oxy-Cope Systems. *Tetrahedron Lett.* **1984**, *25*, 5825–5828.

(61) Schubert, W.; Ugi, I. Constitutional Symmetry and Unique Descriptors of Molecules. *J. Am. Chem. Soc.* **1978**, *100*, 37–41.

(62) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.

(63) Cahn, R. S.; Ingold, C. K.; Prelog, V. Spezifikation der molekularen Chiralität. *Angew. Chem.* **1966**, *78*, 413–447; *Angew. Chem., Int. Ed. Engl.* **1966**, *5*, 385–415.

(64) Ugi, I.; Dugundji, J.; Kopp, R.; Marquarding, D. *Perspectives in Theoretical Stereochemistry*; Lecture Notes in Chemistry 36, Springer-Verlag: Berlin, 1984.

(65) Ugi, I.; Marquarding, D.; Klusacek, H.; Gokel, G.; Gillespie, P. Chemie und logische Strukturen. *Angew. Chem.* **1970**, *82*, 741–771; *Angew Chem., Int. Ed. Engl.* **1970**, *9*, 703–730.

(66) Dietz, A.; Gruber, B.; Ugi, I. "Algorithmische Behandlung stereochemischer Problemstellungen" match. Submitted for publication.

(67) Gruber, B. Algebraische Modellierung der Stereochemie. Doctoral thesis, Technische Universität München, 1992.

(68) Dietz, A. Ein Relationenkonzept zur Modellierung molekularer Systeme. Doctoral thesis, Technische Universität München, 1993.

(69) Streitwieser, A.; Heathcock, C. H. *Organische Chemie*; VCH: Weinheim, New York, 1986.

(70) Krauch, H.; Kunz, W. *Organic Name Reactions*; Wiley: London, 1964.

(71) Krauch, H.; Kunz, W. Reaktionen der organischen Chemie. Ein Beitrag zur Terminologie der organischen Chemie, 5th ed.; Dr. A. Hüthig Verlag: Heidelberg, 1976.

(72) French, S. E. Our Reaction Access System. *CHEMTECH* **1987**, 106–111.

(73) Fujita, S. Description of Organic Reactions Based on Imaginary Transition Structures. 1. Introduction of New Concepts. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 205–212.

(74) CAS-Online—CASREACT User Guide (CAS1266-0488), Chemical Abstracts Service: Columbus, OH, 1988.

(75) Chodosh, D. V. SYNLIB. In *Modern Approaches to Chemical Reaction Searching*, Proceedings of a Conference Organised By the Chemical Structure Association at the University of York, England, July 8–11, 1985; Willett, P., ed.; Gower: Aldershot, U.K., 1986; pp 118–145.

(76) Jones, F. A. Y.; Bunnett, J. F. Nomenclature for Organic Chemical Transformations. *Pure Appl. Chem.* **1989**, *61*, 725–768.

(77) Dengler, A. Algorithmen und Softwaremodule für die computergestützte Lösung chemischer Probleme. Doctoral thesis, Technische Universität München, 1992.

(78) Ugi, I.; Dengler, A. The algebraic and graph theoretical completion of truncated reaction equations. *J. Math. Chem.* **1992**, *9*, 1–10.

(79) Ugi, I. K.; Brandt, J.; v. Scholley, A.; Minker, S.; Wochner, M.; Schönmann, H.; Straupe, B. *Hierarchisch strukturiete Speicherung und Ermittlung von chemischen Reaktionen*; Forschungsbericht, BMFT-FB-ID85-005; Fachinformationszentrum Energie, Physik, Mathematik: Karlsruhe, Germany, 1985.

(80) Brandt, J.; v. Scholley, A. Schönmann, H. From Transformations to Reactions: Algorithmic Balancing of Incomplete Reaction Data. Manuscript in preparation.

(81) Vladutz, G. E.; Geivandov, E. A. *Abtomatizirovannye Informatsionnye Sistemy dlya Khimii*; Nauka: Moscow, 1974.

(82) Vladutz, G. Do We Still Need A Classification of Reactions? In *Modern Approaches to Chemical Reaction Searching*, Proceedings of a Conference Organised by the Chemical Structure Association at the University of York, England, July 8–11, 1985; Willett, P., Ed.; Gower: Aldershot, U.K., 1986; pp 202–220.

(83) Arens, J. F. Eenheid in Verscheidenheid Voordracht 25.Okt.1975, Versl. Gewone Vergad. Afd. Natuurk. *K. Ned. Akad. Wet.* **1975**, *84*, 155–166.

(84) Arens, J. F. A formalism for the classification and design of organic reactions.-I. The class of $(-+)_n$ reactions. *Recl. Trav. Chim. Pays-Bas* **1979**, *98*, 155–161.

(85) Arens, J. F. A formalism for the classification and design of organic reactions.-II. The classes of $(+-)_n+$ and $(-+)_n-$ reactions. *Recl. Trav. Chim. Pays-Bas* **1979**, *98*, 395–399.

(86) Arens, J. F. A formalism for the classification and design of organic reactions-III. The class of $(+-)_n$C reactions. *Recl. Trav. Chim. Pays-Bas* **1079**, *98*, 471–500.

(87) Hendrickson, J. B. Die Vielfalt thermisch pericyclischer Reaktionen. *Angew. Chem.* **1974**, *86*, 71–100; *Angew. Chem., Int. Ed. Engl.* **1974**, *13*, 47.

(88) Hendrickson, J. B. A Systematic Organization of Synthetic Reactions. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 129–136.

(89) Hendrickson, J. B.; Miller, T. M. Reaction Classification and Retrieval. A Linkage between Synthesis Generation and Reaction Databases. *J. Am. Chem. Soc.* **1991**, *113*, 902–910.

(90) Zefirov, N. S.; Tratch, S. S. Systematization of Tautomeric Processes and Formal-logical Approach to the Search for New Topological and Reaction Types of Tautomerism. *Chem. Scr.* **1980**, 4–12.

(91) Zefirov, N. S. An Approach to Systematization and Design of Organic Reactions. *Acc. Chem. Res.* **1987**, *20*, 237–243.

(92) Fujita, S. Description of organic reactions based on imaginary transition structures. 2. Classification of one-string reactions having an even-membered cyclic reaction graph. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 212–223.

(93) Fujita, S. Description of organic reactions based on imaginary transition structures. 3. Classification of one-string reactions having an odd-membered cyclic reaction graph. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 224–230.

(94) Fujita, S. Description of organic reactions based on imaginary transition structures. 4. Three-nodal and four-nodal subgraphs for a systematic characterization of reactions. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 231–237.

(95) Fujita, S. Description of organic reactions based on imaginary transition structures. 5. Recombination of reaction strings in a synthesis space and its application to the description of synthetic pathways. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 238–242.

(96) Fujita, S. Canonical Numbering and Coding of Imaginary Transition Structures. A Novel Approach to the Linear Coding of Individual Organic Reactions. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 128–137.

(97) Fujita, S. Canonical Numbering and Coding of Reaction Center Graphs and Reduced Reaction Center Graphs Abstracted from Imaginary Transition Structures. A Novel Approach to the Linear Coding of Reaction Types. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 137–142.

(98) Fujita, S. A novel approach to systematic classification of organic reactions. Hierarchical subgraphs of imaginary transition structures. *J. Chem. Soc., Perkin Trans.* **1988**, *2*, 597–616.

(99) Fujita, S. A novel approach to the enumeration of reaction types by counting reaction-center graphs which appear as the substructures of imaginary transition structures. *Bull. Chem. Soc. Jpn.* **1988**, *61*, 4189–4206.

(100) Fujita, S. Method for Processing Information on Chemical Reactions. Eur. Pat. Appl 86111173.0, Aug 12, 1986.

(101) Brandt, J.; Bauer, J.; Frank, R. M.; v. Scholley, A. Classification of Reactions by Electron Shift Patterns. *Chem. Scr.* **1981**, *18*, 53–60.

(102) Ugi, I. A Qualitative Global Mathematical View of Chemistry—James Dugundji's Contribution to Computer Assistance in Chemistry. In *Computer Applications in Chemical Research and Education*; Brandt, J., Ugi, I., Eds.; Hüthig Verlag: Heidelberg, 1989; pp 345–366.

(103) Weidinger, R. Ein stereochemischer Netzwerkgenerator. Master's thesis, Universität Passau, 1991.

(104) Brandt, J.; v. Scholley, A. A short linear notation for common reaction types. Manuscript in preparation.

(105) v. Scholley, A. Hierarchische Klassifizierung und Dokumentation von chemischen Reaktionen. Doctoral thesis, Technische Universität München, 1981.

(106) Brandt, J.; v. Scholley, A. An Efficient Algorithm for the Computation of the Canonical Numbering of Reaction Matrices. *Comput. Chem.* **1983**, *7*, 51–59.

(107) Brandt, J.; v. Scholley, A. A Systematic Classification of Reactions by Electron Shift Patterns. CNA Conference on The Future of Chemical Documentation. Session 6: Reaction Indexing, Sep 9, 1982, University of Exeter, England. See: Ash, J. E.; Chubb, P. A.; Ward, S. E.; Welford, S. M.; Willett, P. *Communication, Storage and Retrieval of Chemical Information*. Ellis Horwood: Chichester, U.K., 1985.

(108) Brandt, J.; v. Scholley, A. Definition of a linear notation for reactions based on extended reaction matrices. Manuscript in preparation.

(109) Brandt, J.; Stadler, K. A Recursive Reaction Generator. CSA Conference, July 8–11, 1985, University of York, England. In *Modern Approaches to Chemical Reaction Searching*, Proceedings of a Conference Organised by the Chemical Structure Association at the University of York, England, July 8–11, 1985; Willett, P., Ed.; Gower: Aldershot, U.K., 1986, pp 221–239.

(110) Stadler, K. P. Heuristische Reaktions-Generierung. Doctoral thesis, Technische Universität München, 1986.

(111) Herges, R. Discovery of a new reaction category using reaction databases. Presented at The Third International Conference on Chemical Structures, The International Language of Chemistry, Leeuwenhorst Congress Center, Noordwijkerhout, The Netherlands, June 6–10, 1993. See also: Herges, R. Ordnungsprinzip und Theorie complexer Reaktionen. *Angew. Chem.*, in press.

(112) Brandt, J.; v. Scholley, A.; Wochner, M.; Stadler, K. *A Documentation System for Chemical Reactions*, 188th ACS National Meeting, Aug 26–31, 1984; Section on Chemical Information, Symposium on Chemical Reactions Databases 32; American Chemical Society: Washington, D.C., 1984.

(113) Brandt, J.; v. Scholley, A.; Wochner, M. Making Chemical Reaction Data Accessible to the Non-Chemist. *Comput. Phys. Commun.* **1984**, *33*, 197–203.

(114) Ugi, I.; Stein, N.; Knauer, M.; Gruber, B.; Bley, K.; Weidinger, R. New Elements in the Representation of the Logical Structure of Chemistry by Qualitative Mathematical Models and Corresponding Data Structures. *Top. Curr. Chem.* **1993**, *166*, 199–233.