Proc. Estonian Acad. Sci. Chem., 1994, 43, 4, 121—136
https://doi.org/10.3176/chem.1994.4.01

Dedicated to George Vleduts and Joshua Lederberg, the pioneers of computer assistance in chemistry.

THE TWO FORMAL LANGUAGES OF CHEMISTRY — THEIR SEMANTICS AND SYNTAX

Ivar UGI, Natalie STEIN, and Bernhard GRUBER

Institut für Organische Chemie und Biochemie der Technischen Universität München (Institute of Organic Chemistry and Biochemistry, Technical University of Munich), Lichtenbergstr. 4, D-85747 Garching, BRD (Germany)

Received August 2, 1994; accepted August 25, 1994

Abstract. Chemical formulas, the basis of communicating chemical information and chemical reasoning, are terms of a formal language. The translation of molecular structures into chemical formulas can be regarded as the semantics. The syntax is given by the rules about relations between chemical systems, in particular their interconversions by chemical reactions. Mathematics is used as a formal language in chemistry through the algebra of the *be-* and *r*-matrices and the theory of chemical identity groups. This leads to the notion of chemical grammar, which serves as a basis of the computer assisted deductive solution of chemical problems.

Key words: language, grammar, syntax, semantics, computer-chemistry, automated reasoning, modelling, formal chemistry.

1. INTRODUCTION

The science of chemistry emerged in an age when abstract approaches received little attention [1]. Various chemical concepts and "theories" were developed more on the basis of *natural* languages than on mathematics. Even today, most definitions and explanations in chemistry are given in terms of some natural language.

It is noteworthy that in early chemistry a *formal* language was also developed, the language of *chemical formulas*. This was not only due to the desire of symbolic representations and visualizations of molecular structures but also to the missing expressiveness of natural languages for chemical objects and facts. Many chemical phenomena and arguments can only be efficiently expressed by the customary language of chemistry, the chemical formulas.

Formal languages are a subject of linguistics as well as of informatics. These sciences define a formal language as a *set of symbols* and a *set of syntactic rules*. The symbols may form sequences, the so-called *expressions*, according to the syntactic rules. The syntactic rules include a (possibly empty) set of transformation rules that allow the construction of other expressions from given expressions. These transformations are purely

syntactic, i. e. they only manipulate symbols, regardless of their meaning. Formal languages do *not* require the notion of an interpretation, nor does a *theory*. A theory consists of a language and a set of sentences, called *axioms*. Axioms determine the set of valid sentences of a language. Not all sentences of a language are valid.

By *interpreting* the symbols of a language, the set of valid sentences provides a meaning, namely its *semantics*. If all axioms of the theory are valid for a given interpretation, then it is called a *model*.

Although formal languages, as well as theories, do not require the notion of an interpretation, they are constructed with interpretations in mind. This applies particularly to the formal language of chemical formulas [2].

The symbols of the language of chemical formulas are the nodes of the molecular graphs that represent the constituent atoms of the chemical structure. Atoms can be classified by mapping the respective elements to the finite set of chemical elements. This is denoted by using a second symbol. Finally, two functional symbols describe the distribution of valence electrons, chemical bonds, and also their redistribution, which corresponds to chemical reactions.

The valence chemical properties of the elements and the general principles are formulated, using the syntactic means of the language. They determine the relations between molecular systems, including their interconversions by chemical reactions. These lead also to the set of axioms in the electronic theory of chemical reactivity. The ability of achieving this involves a far-reaching interpretation of chemistry. To become fluent in the formal language of chemical formulas takes prolonged training. This is a disadvantage the formal languages generally have in contrast to natural languages. However, some advantages of the formal languages prevail: there are no unintentionally incomplete statements, and there is no ambiguity. Due to their exactness, the formal languages are the one and only basis of constructing computer programs.

The algebra of *be-* and *r*-matrices is also a formal language. It has sometimes even advantages compared with the language of chemical formulas. The mathematical nature of the algebra of the *be-* and *r*-matrices simplifies the translation of the language of chemical formulas into a (formal) computer language.

At the present state of the art, formal languages in chemistry are used for automated problem-solving, such as computer assisted synthesis design [3, 4]. In the near future, automated deductive reasoning by purely syntactic transformations will probably become one of the most important fields in mathematical chemistry.

In his visionary novel "1984", George Orwell describes how a totalitarian government limits and influences the thoughts of the people by manipulating their language. This succeeds quite well, because human reasoning and communication depend on what can be expressed by the available language. In chemistry, progress in paradigms and abstract concepts coincides with new developments in its formal languages. The state of the art of chemistry is reflected by the state of its languages.

2. THE TRANSITION FROM ALCHEMY TO SCIENTIFIC CHEMISTRY, A MATTER OF LANGUAGES

Present-day chemistry is a fairly young science. The history of scientific chemistry comprises barely two centuries. The roots of chemistry, however, reach far back into the prehistoric ages.

Until the eighteenth century, the knowledge in chemistry consisted of individual observations and experiences. Between these no relations were established, and the records of chemistry were limited to the description of facts and phenomena. Chemical substances were generally classified by their consistency. Thus, olive oil and vitriolic oil belonged to the same class of oils. Ethanol (spir. vini), stannic chloride (spir. fumans Libavii), aqueous ammonia (spir. cornu cervi), and nitric acid (spir. nitri) were classified as spirits [5].

The insight that the customary qualitative observations in chemistry must be supplemented by quantitative measurements paved the way for a transition from phenomenological alchemy to scientific chemistry. The balance was the first and most important instrument for measurements in chemistry, the thermometer followed soon. The introduction of the balance into chemistry by Lavoisier [6] led to quantitative analyses. He defined chemistry as an analytical science. In the era of dominating analytical chemistry, combustion analyses of organic substances were the foundation of conceptual progress in chemistry. Thus many fundamental principles and relations were established, such as the *principle of mass conservation* and Dalton's *simple rules about the multiple stoichiometric relations* [7] from which the existence of atoms was inferred.

Berzelius [8] introduced the concept of isomerism for "distinguishable compounds that have the same molecular weight and the same element composition". The existence of isomeric chemical compounds stimulated A. v. Humboldt [9] to conclude that the molecules — these were then still called atoms, the smallest indivisible units of a given compound — must be endowed with an intrinsic structure.

In the first half of the nineteenth century many competing notions about the intrinsic structure of matter have been discussed. Then the *classical structure theory* evolved. It is the result of the collective contributions of Couper, Erlenmeyer, Frankland, Gerhardt, Kekulé, Kolbe and Wislicenus, the leading chemists of their time.

In 1860 at the Karlsruhe International Congress of Chemistry the structure theory was generally accepted as the valid description of molecular structure. Edwin Hjelt [5], the renowned Finnish historian of chemistry, described this illustrious gathering of chemists rightfully as an "event of unique splendour and importance in this world".

The term *structure* was then used for what we now call *chemical constitution*: the chemical constitution of a molecule is described by the number and kind of the constituent atoms, and by stating for each atom its immediate neighbouring atoms to which it is connected by covalent chemical bonds. The chemical constitution of molecules is represented by their constitutional formulas.

The first and the most important formal language of chemistry is the language of constitutional formulas, the final step in the transition from alchemy to scientific chemistry. The language of chemical formulas is based on the most important abstraction process in the history of chemistry, an intellectual achievement that is commensurable to the introduction of Kopernican heliocentric model of the solar system.

In essence, the early versions of this language, the *structural theory* of chemistry, consisted only of its *semantic part*, the translation of molecular "structure" into constitutional formulas, and only minor elements of its *syntax* were already used, such as some stoichiometric and valence chemical rules about the transformations of constitutional formulas in the representation of chemical reactions.

The classical structure theory enabled chemists to carry out systematically planned multistep syntheses of complex organic molecules. Today synthesis is the main objective of chemistry, while analytical chem-

istry is an auxiliary discipline, but still a very important field, since its contributions are absolutely necessary for chemistry and many other sciences. Dramatic progress in instrumentation, sometimes even an overkill in sophisticated and expensive spectroscopic equipment, is characteristic and seemingly a status symbol in analytical chemistry, while the instruments of syntheses have not changed much — in the laboratory most syntheses are still carried out in a simple three-necked glass flask.

The major parts of the syntax of the formal language of constitutional formulas emerged in this century, when the classical structural theory was confirmed by quantum chemistry, and when chemical reactions were interpreted as conversions of reactants into reaction products by redistribution of the valence electrons. Quantum chemistry, in particular the so-called valence bond theory, led to the presently accepted interpretation of chemical bonds and the rules by which the participants of chemical reactions are changed and their constitutional formulas are transformed [10]. The rules about the allowable redistributions of valence electrons are an important part of the syntax of the language of chemical formulas.

The postulate of the asymmetric carbon atom by Van't Hoff and Le Bel [11] was the beginning of stereochemistry. The distinct stereoisomers, their synthesis and identification, properties and behaviour are the subject of stereochemistry. According to their traditional definition [12], stereoisomers are distinct chemical compounds, whose molecules have the same chemical constitution but differ by the relative spatial arrangement of the constituent atoms. In classical stereochemistry the stereoisomers are represented by their rigid geometric models and the corresponding projection formulas or descriptors that refer to the latter. The language of traditional stereochemistry is confined to its semantic part.

3. THE MATHEMATISATION OF CHEMISTRY

Hilbert [1] called mathematics *die Lehre von den formalen Systemen.* Mathematics and physics are intertwined to the extent that often it is not clear where physics ends and pure mathematics begins. Mathematics has entered chemistry mainly via physical theories which are used in chemistry for computing some measurable properties of molecular systems. Many chemists believe that a theory is not a "real theory" if it does not ultimately yield numbers that can be compared with the numerical values of some experimental results.

Since long physics-based mathematical theories like quantum chemistry play an important role in the interpretation and prediction of the observable physical properties of chemical systems and thus, indirectly, their chemical behaviour. In these chemical applications, mathematics is clearly not used in the sense of a language for processing and communicating chemical information, nor can it be used for the direct solution of the traditional problems of chemistry, like the design of syntheses or the elucidation of reaction mechanisms. The solutions of traditional chemical problems are chemical objects, i. e. molecules and chemical reactions, that are characterized by some given or expected relations to other objects of chemistry.

The direct use of mathematics in chemistry without the intermediacy of physics is qualitative in nature. It corresponds more to a language than to a traditional theory. Discrete mathematics dominates in this long-neglected field. In recent decades, however, the important and fast-growing discipline of mathematical chemistry has evolved.

Shortly after the *structural formulas* had been introduced, Sir Arthur Cayley [13] treated chemical formulas as graphs of binary relations. Graph theory is now widely used as a means of representing and visualizing molecular systems [14].

With his enumeration theorem Pólya [15] initiated the group theoretical classification and enumeration of isomers and their families. This has now developed into a wide and substantial area of mathematical chemistry. Until recently most of these direct applications of mathematics in chemistry have been confined to aspects like the description and visualization, the classification and enumeration of chemical objects [16]. This use of mathematics as a language of chemistry is essentially confined to representation purposes.

3.1. Equivalence classes and chemical identity

A full translation of chemistry into mathematics is accomplished through an analysis of the *logical structure of chemistry* which is a complex, contiguous network of relations, mostly equivalence relations (reflexive $[xRx]$, symmetric $[xRy \Rightarrow yRx]$, and transitive $[xRy, yRz \Rightarrow xRz]$ relations between the elements of a set), that define equivalence classes of chemical objects and processes [17]. The hierarchic classification of the structural features of molecules on which the various types of isomerism are based [18] is the basic framework for the logical structure of chemistry. The concept of chemical identity and the non-geometric definition of stereoisomers play an important role in the abstraction process that reveals the logical structure of chemistry. The term *stereoisomer* denotes a molecule that is a stereoisomer of a reference molecule but also a chemical compound, whose molecules are the stereoisomers of the molecules of a reference compound.

Molecules that belong to the same chemical compound and interconvert spontaneously under the given observation conditions are chemically identical. Stereoisomeric molecules have in common the same chemical constitution but are not chemically identical [19, 20].

The above definition implies that chemistry has two aspects, constitutional chemistry and stereochemistry. Accordingly, the logical structure of chemistry consists of two parts; each can be represented by its own mathematical model.

3.2. Constitutional chemistry and its traditional representation

Molecules are aggregates of atomic cores (atomic nuclei and inner shell electrons) that are connected by valence electrons. In general the valence electrons form covalent bonds, or they belong to the individual atomic cores as so-called free valence electrons. A covalent bond (of formal order one) is a pair of valence electrons that is shared by two neighbouring atomic cores. The latter are thereby held together. During chemical reactions the valence electrons are redistributed, while the atomic cores remain unchanged.

This simple picture of molecules and chemical reactions is the foundation of a mathematical model of the logical structure of constitutional chemistry.

The chemical constitution of a molecule or *ensemble of molecules* (EM) is customarily described by a constitutional formula in which the atomic cores are represented by chemical element symbols, the covalent bonds by lines, and the free valence electrons by dots.

In modern chemical documentation the chemical constitution of molecules is generally recorded in terms of *connectivity matrices* [21].

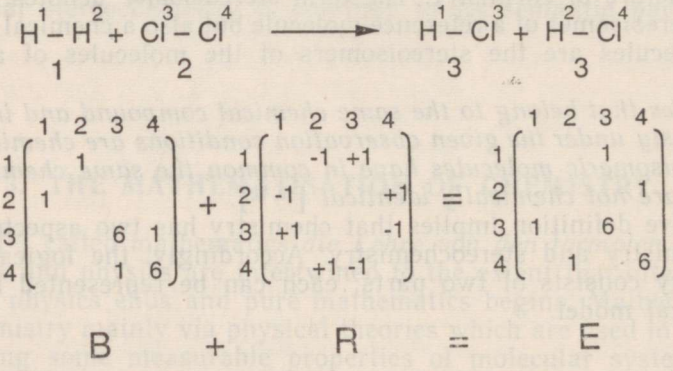
These are quadratic tables that specify covalently bound neighbouring atoms for each atomic core by their off-diagonal entries. The diagonal entries are chemical element symbols that label the rows/columns. No mathematical operations have yet been defined for the connectivity matrices. They are just tabular representations of constitutional formulas.

3.3. The theory of the *be*- and *r*-matrices

When the diagonal entries of the connectivity matrices are replaced by the numbers of free electrons at the atomic cores, they are converted into ***be***-matrices (***b***ond and ***e***lectron matrices) [22], which represent the chemical constitution of molecules as well as polymolecular EMs, and are genuine mathematical objects with well-defined mathematical properties.

The rows/columns of a *be*-matrix **B** of an EM are assigned to the atomic cores. The off-diagonal entries $b_{i,j}$ $(=b_{j,i};\ i \neq j)$ are the (integer) formal bond orders between the atomic cores A_i and A_j. The diagonal entries $b_{i,i}$ are the numbers of free valence electrons at the atomic cores A_i.

Thus an EM_B which consists of hydrogen **1** and chlorine **2**, is represented by the matrix **B**, and the constitution of the EM_E of two molecules of hydrogen chloride **3** is expressed by **E** (the zero-entries of the matrices have been omitted for the sake of simplicity).



In **B** the entry $b_{1,2}=1$ in the first row and second column indicates that the hydrogen atoms number one and number two are connected by a covalent bond, whose formal bond order is 1. The fourth diagonal entry $b_{4,4}=6$ tells, that the chlorine atom number four carries 6 free valence electrons.

A chemical reaction $EM_B \longrightarrow EM_E$ is represented by the transformation of the *be*-matrix **B** of EM_B into the *be*-matrix **E** of EM_E by the addition of a so-called *r*-matrix **R** (***r***eaction matrix) to **B** according to the matrix equation

$$\mathbf{B}+\mathbf{R}=\mathbf{E},$$

the fundamental equation of the algebra of the *be*- and *r*-matrices (matrices are added entry by entry: $\mathbf{B}+\mathbf{R}=\mathbf{E} \Rightarrow b_{i,j}+r_{i,j}=e_{i,j}$; the above matrix representation of the reaction $1+2 \rightarrow 3+3$ illustrates this formalism).

The off-diagonal entries $r_{i,j}$ $(=r_{j,i})$ of **R** indicate the alterations in bond order of the covalent bonds between A_i and A_j, while the diagonal

entries $r_{i,i}$ are changes in the number of free valence electrons at the atomic cores A_i.

The *be*- and *r*-matrices are not only a replacement of constitutional formulas and patterns of "electron pushing arrows" by matrices, but a full translation of constitutional chemistry into the language of mathematics. An essential feature of the translation of chemistry into mathematics is that besides the *semantics* of constitutional chemistry, that is given by the correspondence of constitutional formulas and *be*-matrices, we have also a chemistry-oriented *syntax* for the interconvertibility and other *relations* between the molecular systems. This syntax is contained in the algebraic rules for combining the *be*- and *r*-matrices, that are codified in the 18 theorems of the algebra of the *be*- and *r*-matrices [22]. In this algebraic representation of chemistry the new aspect is the direct mathematical representation of chemical processes and relations by matrix transformations, that can be generated according to strictly formal algorithms [23]. *Nota bene*: The scope and validity of formal algorithms must be proven like a theorem.

Chemical problems can be solved by solving the matrix equation $\mathbf{B}+\mathbf{R}=\mathbf{E}$. Its mathematical results can answer chemical questions.

The global algebraic model of the logical structure of chemistry can also be expressed and visualized by a geometrical model. The $n\times n$ *be*-matrices **B** and **E** of *n*-atomic isomeric EMs correspond to *be*-points $P(\mathbf{B})$ and $P(\mathbf{E})$ in an n^2-dimensional space, and the respective *r*-matrices to *r*-vectors that connect these points. Thus, the topological nature of chemistry as a whole and the vectorial nature of chemical reactions are visualized [22, 24].

Chemical reactions can be manipulated according to the laws of vector algebra. This is of great interest in its own right, but of particular importance for the computer-assisted treatment and the prediction of chemical reactions [3].

The "l_1-distance" between $P(\mathbf{B})$ and $P(\mathbf{E})$, their "taxidriver distance",

$$d(\mathbf{B}, \mathbf{E}) = \sum |b_{i,j} - e_{i,j}|$$

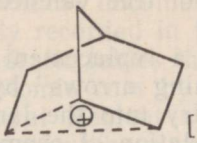
corresponds to the "length" of the *r*-vector that connects the *be*-points of the respective EMs. It is the so-called *chemical distance* (CD) between EM_B and EM_E [22, 24–26]. The CD has a clear chemical meaning. It is twice the number of valence electrons that are redistributed during the interconversion of EM_B and EM_E.

The interconversions of isomeric EMs by chemical reactions proceed preferentially along pathways of minimum CD. This is called the ***principle of minimum CD*** (PMCD) [25, 26]. The PMCD is a quantitative version of the somewhat vaguely formulated classical principle of minimal structure change [27]. The PMCD is useful in many ways. For instance, it serves as the basis of a hierarchic classification system for chemical reactions, and of selection procedures for computer generated solutions to chemical problems [24, 28, 29]. The PMCD is an important element of the syntax in the mathematical language of chemistry.

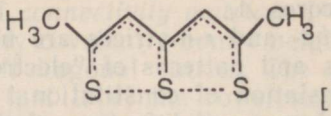
3.4. Delocalized electrons and extended *be*- and *r*-matrices

The chemical constitution of many molecules with delocalized and "non-classical" bond-systems cannot adequately be described by conventional chemical formulas, nor by *be*-matrices with integer formal bond orders [4].

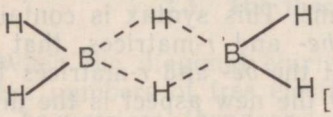
Examples are:



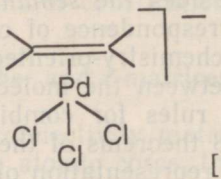
[30],



[31],



[32],



[33].

Such molecules could be represented by formulas with fractional bond orders [22], or by families of resonance structures with integer bond orders and the corresponding *be*-matrices [34]. Neither approach is satisfactory, in particular with regard to computer assistance in chemistry.

Recently we introduced the e*x*tended *be*-matrices (*xbe*-matrices) as a mathematical representation for EMs with systems of *d*elocalized *e*lectrons (DE-systems) [3, 4]. An *xbe*-matrix **XB** or **XE** of an EM accounts for the localized covalent bonds of the EM by its *be*-matrix-part. Each DE-system is represented by an extra row/column, whose off-diagonal non-zero entries $b_{i,\,n+k}=b_{n+k,\,i}=1$ $(1\leqslant i\leqslant n,\ 1\leqslant k\leqslant m$; n: number of atoms; m: number of DE-systems) indicate the participation of A_i in the k-th DE-system, and whose diagonal entry $b_{n+k,\,n+k}$ is the number of delocalized electrons belonging to the k-th DE-system.

In analogy to the theory of the *be*- and *r*-matrices, chemical reactions are represented by the additive transformation of the *xbe*-matrices by *xr*-matrices according to

$$\mathbf{XB}+\mathbf{XR}=\mathbf{XE}.$$

When the off-diagonal DE-entries $b_{i,\,n+k}=b_{n+k,\,i}$ are neglected, the algebra of the *xbe*- and *xr*-matrices follows the same theorems that govern the algebra of the *be*- and *r*-matrices [22].

Thus, in the transition from the language of the *be*- and *r*-matrices to the language of the *xbe*- and *xr*-matrices, the semantics changes substantially. However, the syntax is essentially preserved, with the exception of the valence chemical boundary condition [3].

For computer assistance in chemistry it is advantageous to represent EMs with DE-systems and their reactions by data structures which correspond to the *xbe*- and *r*-matrices [4]. This may be considered as a special version of the present mathematical formal language of chemistry. It is novel as a representation, but as a language of chemistry it is an exact image of the algebra of the *xbe*- and *xr*- matrices in both aspects, semantics and syntax.

3.5. Stereochemistry

Stereoisomers, the differences in their observable properties and their behaviour are the topic of stereochemistry.

In stereochemistry a higher degree of abstraction was hampered by the overemphasis of rigid geometric molecular models and their point group symmetries. For a long time no significant progress was made.

Due to the non-rigid nature of many types of molecules, a non-geometric, more abstract definition of stereoisomers on the basis of the notion of *chemical identity* is more appropriate (see above) [19]. This definition has led to a unified theoretical treatment of stereochemistry, profound

changes in the interpretation of stereochemistry [20], an improved definition of chirality and new powerful tools for the solution of stereochemical problems, especially when non-rigid molecules and relations between molecules are involved.

Based on the concept of permutational isomerism [17], the theory of the chemical identity groups (CIG-theory) was formulated [19].

The concept of permutational isomerism was introduced in 1970 [17], but permutation isomers were already enumerated and classified as unspecified "isomers" more than 30 years earlier by Pólya [15], and later by many others [16].

Permutation isomers with a monocentric skeleton are always stereoisomers, whereas stereoisomers are not necessarily permutation isomers, e. g. if they may have different molecular skeletons. Permutation isomers with a polycentric skeleton may also differ constitutionally [17].

With the use of ligand permutations in the description, classification and enumeration of permutation isomers [16], the semantic part of the theory of permutation groups was introduced as part of a mathematical formal language of stereochemistry. In 1984 the syntax was added by the introduction of the *set-valued mappings* (SVM) of the coverings and partitions of $\mathrm{Sym}(L)$ [19]. $\mathrm{Sym}(L)$ is the symmetric permutation group on n ligands.

We now enter a brief outline of the CIG-theory by considering a reference model E of a molecule. It consists of n skeletal sites that form a molecular skeleton which is not necessarily contiguous, and a set of n distinct ligands L whose placement at the skeletal sites is characteristic for this reference model. The model E is assumed to have a fixed orientation in space. Any model is chemically identical to the reference model and belongs to the reference isomer X. It must be obtained from the latter by a rotation of the whole model or by an intramolecular movement. This takes place spontaneously under the given observation conditions.

The permutations of the distinct ligands produce chemically identical models from the reference model and form a group, the CIG of the reference isomer. This CIG $S(X)$ with $|S(X)|$ elements is a subgroup of $\mathrm{Sym}(L)$, the symmetric permutation group on n ligands. When all ligand permutations $\lambda \in \mathrm{Sym}(L)$ are applied, $n!$ distinct models result. These form a family of $n!/|S(X)|$ distinct permutation isomers, which are represented by the left cosets $\lambda S(X)$ of $S(X)$ in $\mathrm{Sym}(L)$. If some of the ligands are chemically indistinguishable, the family of the permutation isomers of X is represented by the double cosets $\Sigma\lambda S(X)$ in $\mathrm{Sym}(L)$; Σ is the group of all permutations of indistinguishable ligands.

Any member of a coset or double coset that corresponds to a given permutation isomer may be used as its nomenclatural descriptor [17].

A SVM is a total subjective mapping of an element in a set onto another set. Within the CIG-theory the SVMs are used to determine equivalences of molecular models and the permutation isomers to which they belong. Thus a variety of equivalence relations and the respective equivalence classes are established between models of permutation isomers. These may belong to the same family or to different families. Interconvertibility by chemical reactions is treated as an equivalence relation, and also the formation of distinct stereoisomers by a common reaction, that affects the chemical constitution of the reactants. Thus, the SVMs have a very wide scope of applications in stereochemistry [3, 19]. They reveal those equivalence relations that make up the logical structure of stereochemistry. The SVMs are the backbone of the syntax of the permutational language of stereochemistry. Substantial progress has been made in the permutational treatment of stereochemistry by the computer-oriented so-called *accumulations* [2, 3] of equivalence classes and the introduction

of *filters* [3, 35]. Through the filters it is possible to "filter" from the formally possible, conceivable stereochemical information those parts that belong to chemical reality. Thus, like in any grammar, the exceptions to the general rules are accounted for.

4. COMPUTER ASSISTANCE IN CHEMISTRY

4.1. Empirically-based retrosynthetic analysis

The late George Vleduts * [36], one of the pioneers of modern computer-assisted documentation of chemical reactions, proposed in 1963 to use so-called reaction libraries for generating retrograde synthetic pathways from a target molecule to the available starting materials. In 1967 began the implementation of the computer programs LHASA [37] and its offspring SECS [38] for the design of syntheses by so-called retrosynthetic analysis according to the ideas published by Vleduts. In a joint effort the major Swiss and German chemical companies implemented CASP [39], a retrosynthetic analysis program. CASP resulted from SECS by a massive extension of its reaction library and improvements of its overall structure. With CASP a pinnacle was reached. Beyond CASP no further significant progress seems to be possible in the field of retrosynthetic analysis.

For more than a decade empirically-oriented computer-assisted synthesis design by retrosynthetic analysis on the basis of reaction libraries — phrase books in the language of chemical formulas — has been the main topic of computer assistance in chemistry. The respective computer programs are generally very large and must be operated by full-time specialists. The syntheses are generated by retrieval and manipulation of stored chemical information. The resulting syntheses are derived from the published literature in a straightforward manner. The development of computer programs for retrosynthetic analysis had started with excellent ideas, great enthusiasm and high hopes. However, despite great efforts and expenditures no widely accepted major information-based general computer program for the design of multistep-syntheses has resulted. The principal reasons for the stagnation of retrosynthesis are that no new ideas have fertilized this field, and the lacklustre acceptance by the chemists as the potential users. The users prefer to participate directly in the design of syntheses rather than to use computer programs through the intermediacy of a specialist. The repertoire of empirically-based retrosynthetic analysis is less than a thinking chemist can accomplish when he uses data banks of chemical reactions directly.

Most of the prominent authors in this discipline, like Corey and Wipke [37, 38], have left the field, and the major chemical companies that were active in data-oriented retrosynthetic analysis have discontinued their endeavours. When asked by one of us (*I. U.*), why he had not taken the initiative in the implementation of synthesis design programs according to his early proposal, Vleduts answered that on second thought he had understood how much effort and expenditures such a program would require, and what its limitations would be. Computers are not very efficient in directly solving chemical problems by straightforward methods that simulate simple reasoning and information processing by human beings.

---

* Until 1961 he lived and worked in Moscow; subsequently he emigrated to the USA. He died in 1990.

4.2. The DENDRAL project

About the time, when the development of the empirical synthesis design programs began, also a different type of attempt was made to introduce computer assistance into chemistry. It was the DENDRAL project of Lederberg et al. [40]. Its goal was to create a system of computer programs for the elucidation of molecular structures from spectroscopic data. Here the basic approach was to first determine the molecular substructures of a given compound from its spectroscopic data, and then to "assemble" the complete molecular structure from its substructures by graph theory based methods.

The DENDRAL project was terminated, presumably for financial reasons, before it could be completed. Nevertheless, it is a most important milestone, and not only in the field of mathematically-based computer assistance in chemistry. DENDRAL is the archetype of a whole generation of computer programs for elucidating and generating molecular structures [41]. The program GENOA [42], part of DENDRAL, belongs to the most efficient computer programs for the latter purpose.

4.3. Mathematics-based problem-solving in chemistry

As a formal language of chemistry, mathematics can substantially improve the scope of computer assistance in the direct solution of chemical problems. Not only does it lead to new methods in problem-solving, but it can also contribute new non-arbitrary ways for obtaining the desirable solutions without the need to screen, evaluate and select them from the tremendous amount of computer-generated conceivable but mostly irrelevant results [29].

The feasibility of computer programs for the solution of chemical problems on the basis of the algebra of the *be*- and *r*-matrices was ascertained by a research project completed in 1974 [43]. Subsequently the Munich Project [3] was devised, a master plan for the exploitation of mathematics as a basis of computer programs for the deductive solution of chemical problems, without the need to use detailed empirical chemical information. The Munich Project comprises the continued analysis of the logical structure of chemistry and its representation by mathematical formalisms, the development of algorithms and computer programs for the solution of chemical problems, including the required software infrastructure, and finally, test runs of the implemented computer programs, as well as the experimental verification of the computer-generated solutions of chemical problems. Although the details of the Munich Project have changed over the years, its essence persisted.

In the meantime the theory of the *be*- and *r*-matrices has been extended to include the EMs with delocalized electron systems [4] and has been supplemented by the CIG theory [19] and its recent amendments [2, 3, 35]. Thus a mathematical edifice has been erected, which is a full and specific second language of chemistry.

After a long multistage development [44] of the formal reaction generators the transition table guided systems have evolved that produce the chemically meaningful solutions of the equation $\mathbf{B}+\mathbf{R}=\mathbf{E}$ very efficiently [23]. The formal reaction generators serve as the "engines" of the multipurpose computer programs IGOR [28, 45] and RAIN [46, 47]. These solve a great variety of chemical problems such as mono- and bilateral synthesis design [48], the documentation and prediction of molecular structures [49, 50] and chemical reactions [3, 28] as well as the elucidation of reaction mechanisms [51], including prebiotic and biochemical processes. The

results that are generated by IGOR and RAIN are subjected to automated formal selection procedures. These can be interactively guided by the user who can thus also emphasize his intentions. The combination of formal approaches and the capabilities of computers with the creative intelligence and the experience of the user works very well in computer-assisted solution of chemical problems. Up to now IGOR and RAIN are the only computer programs that have invented many completely new real chemical reactions [3, 52] and an extremely complex reaction mechanism [51].

In the meantime more than a dozen of unprecedented reactions have been proposed with the aid of IGOR; these have been successfully verified in the laboratory [3]. A very complex reaction mechanism has been elucidated by a combination of assistance by RAIN and experiments, that were partly suggested by RAIN [51].

5. PERSPECTIVES

The development of mathematics as a formal language of chemistry had a beneficial effect on the framework of concepts in chemistry. Before translating chemistry into mathematics it was necessary to check whether or not the existent essential notions and models suffice as a prerequisite for the formulation of a mathematical formal language of chemistry. It was found that the conventional abstract foundations of chemistry do not suffice for the given purpose. Accordingly the required revisions were made before the direct translation of chemistry into mathematics was undertaken. These endeavours resulted, for example, in modified notions of constitutional isomerism [22], chemical identity, stereoisomerism [19] and chirality [20] and the introduction of diverse new concepts like permutational isomerism [17]. On the other hand, the direct mathematical representation of chemistry yielded many novel ideas and general insights in chemistry, like the global topological picture of chemistry [53] and the vectorial nature of chemical reactions [4]. It also yielded various practically applicable formal rules like the PMCD [3, 24, 25]. Also, new dimensions in computer-assisted chemistry have emerged [54]. It is becoming increasingly clear that a mathematical formal language of chemistry is very useful in the direct computer-assisted solution of inherent problems of chemistry like the design of syntheses, the prediction of chemical reactions and the elucidation of reaction mechanisms [3]. The solutions of such problems are generally molecular systems and chemical reactions that are representable by chemical formulas or equivalent symbols.

ACKNOWLEDGEMENT

We gratefully acknowledge the financial support of our work by Volkswagen-Stiftung e. V.

REFERENCES

1. Hilbert, D., Bernays, P. Grundlagen der Mathematik. Teubner, Berlin, 1934.
2. Gruber, B. Algebraische Modellierung der Stereochemie. Doctoral Thesis, Technical University Munich, 1992.
3. Ugi, I., Bauer, J., Bley, K., Dengler, A., Dietz, A., Fontain, E., Gruber, B., Herges, R., Knauer, M., Reitsam, K., Stein, N. Computerunterstützte direkte Lösung chemi-

scher Probleme — die Entstehungsgeschichte und der gegenwärtige Status einer neuen 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; Ugi, I., Bauer, J., Blomberger, C., Brandt, J., Dietz, A., Fontain, E., Gruber, B., v. Scholley-Pfab, A., Senff, A., Stein, N. Models, concepts, theories, and formal languages in chemistry and their use as a basis for computer assistance in chemistry. — J. Chem. Inf. Comput. Sci., 1994, **34**, 3—16.

4. 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.
5. Hjelt, E. Geschichte der organischen Chemie. Vieweg & Söhne, Braunschweig, 1916.
6. Lavoisier, A. L. Traité élémentaire de chimie. Paris, 1789.
7. Dalton, J. Mémoirs of the Literary and Phil. Soc. Manchester, 1803, **1**, 271—282; Szabadváry, F. History of Analytical Chemistry. Pergamon, Oxford, 1966.
8. Prandtl, W. Humphrey Davy, Jöns Jakob Berzelius. Wiss. Verlagsgesellschaft, Stuttgart, 1948.
9. Humboldt, A. von. Versuche über die gereizte Muskel- und Nervenfaser, nebst Vermutungen über den chemischen Prozeß in der Tier- und Pflanzenwelt. Rottmann, Leipzig, 1797; Lippmann, E. O. von. Alexander v. Humboldt als Vorläufer der Lehre von der Isomerie. — Chemiker-Ztg., 1909, **1**, 1—7; Kerber, A., Laue, R., Moser, D. Ein Strukturgenerator für molekulare Graphen. — Anal. Chim. Acta, 1990, **235**, 221—228.
10. Dewar, M. J. S. The Electron Theory of Organic Chemistry. Clarendon Press, Oxford, 1949; Pauling, L. The Nature of the Chemical Bond. Oxford University Press, Oxford, 1950; Cornell University Press, Ithaka, 1960.
11. Weyer, J. Hundert Jahre Stereochemie — Ein Rückblick auf die wichtigsten Entwicklungsphasen. — Angew. Chem., 1974, **86**, 604—611; Angew. Chem. Int. Ed. Engl., 1974, **13**, 591—598.
12. Eliel, E. L. Stereochemistry of Carbon Compounds. McGraw-Hill, New York, 1962.
13. Cayley, A. Über 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; Herrmann, F. Über das Problem, die Anzahl der isomeren Paraffine der Formel C_nH_{2n+2} zu bestimmen. — Ber. Dtsch. Chem. Ges., 1880, **13**, 791—792.
14. Balaban, A. T. (ed.). Chemical Applications of Graph Theory. Academic Press, London, 1976; King, R. B. (ed.). Chemical Applications of Topology and Graph Theory. Elsevier, Amsterdam, 1983; Merrifield, R. E., Simmons, H. E. Topological Methods in Chemistry. Wiley, New York, 1989.
15. Pólya, G. Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. — Acta Sci. Math., 1937, **68**, 145—254; Pólya, G. Algebraische Berechnung der Anzahl der Isomeren einiger organischer Verbindungen. — Z. Krystallogr. (A), 1937, **93**, 415—443.
16. Bruijn, N. G. de. — In: Beckenbach, E. F. (ed.). Applied Combinatorial Mathematics. Wiley, New York, 1964; Ruch, E., Hässelbarth, W., Richter, B. Doppelnebenklassen als Klassenbegriff und Nomenklaturprinzip für Isomere und ihre Abzählung. — Theor. Chim. Acta, 1970, **19**, 288—300; Hässelbarth, W., Ruch, E. Classifications of rearrangement mechanisms by means of double cosets and counting formulas for the numbers of classes. — Theor. Chim. Acta, 1973, **29**, 259—268; Hässelbarth, W., Ruch, E., Klein, D. J., Seligman, T. H. — In: Sharp, R. T., Kolman, B. (eds.). Group Theoretical Methods in Physics. Academic Press, New York, 1977, 617—625; Hinze, J. (ed.). The Permutation Group in Physics and Chemistry. Springer, Heidelberg, 1979; Kerber, A., Thürlings, K. J. Symmetrieklassen von Funktionen und ihre Abzähltheorie. Bayreuther Math. Schriften, Bayreuth, 1983.

17. 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.
18. Gasteiger, J., Gillespie, P. D., Marquarding, D., Ugi, I. From van't Hoff to unified perspectives in molecular structure and computer oriented representation. — Top. Curr. Chem., 1974, **48**, 1—37.
19. Ugi, I., Dugundji, J., Kopp, R., Marquarding, D. Perspectives in Theoretical Stereochemistry. Lecture Notes Series, **36**. Springer, Heidelberg, 1984; Ugi, I. Logic and order in stereochemistry. — Chimia, 1986, **40**, 340—350.
20. Dugundji, J., Kopp, R., Marquarding, D., Ugi, I. A quantitative measure of chemical chirality and its application to asymmetric synthesis. — Top. Curr. Chem., 1978, **75**, 165—180.
21. Spialter, L. The atom connectivity matrix (ACM) and its characteristic polynom (ACMCP). — J. Chem. Soc., 1964, **4**, 261—269.
22. Dugundji, J., Ugi, I. An algebraic model of constitutional chemistry as a basis for chemical computer programs. — Top. Curr. Chem., 1973, **39**, 19—64.
23. Bauer, J. Die Erzeugung präzedenzloser Reaktionen auf der Grundlage der Algebra der BE- und R-Matrizen. Doctoral Thesis, Technical University Munich, 1981; Fontain, E., Reitsam, K. The generation of reaction networks with RAIN. 1. The reaction generator. — J. Chem. Inform. Comput. Sci., 1991, **31**, 96—101.
24. Ugi, I., Wochner, M., Fontain, E., Bauer, J., Gruber, B., Karl, R. Chemical similarity, chemical distance, and computer-assisted formalized reasoning by analogy. — In: Johnson, M. A., Maggiora, G. M. (eds.). Concepts and Applications of Chemical Similarity. Wiley, New York, 1990, 239—288.
25. Wochner, M., Brandt, J., Scholley, A. von, Ugi, I. Chemical similarity, chemical distance and its exact determination. — Chimia, 1988, **42**, 217—225.
26. Jochum, C., Gasteiger, J., Ugi, I., Dugundji, J. The principle of minimum chemical distance and the principle of minimum structure change. — Z. Naturforsch., 1982, **37b**, 1205—1215.
27. Kolbe, H. Das Prinzip der minimalen Strukturveränderung. — Liebigs Ann. Chem., 1850, **75**, 211—218; 1850, **76**, 1—9.
28. Bauer, J., Herges, R., Fontain, E., Ugi, I. IGOR and computer-assisted innovation in chemistry. — Chimia, 1985, **39**, 43—53.
29. Ugi, I., Bauer, J., Fontain, E. Transparent formal methods for reducing the combinatorial wealth of conceivable solutions to a chemical problem — computer-assisted elucidation of complex reaction mechanisms. — Anal. Chim. Acta, 1990, **235**, 155—161.
30. Olah, G., Schleyer, P. R. von. Carbonium Ions, **3**. Wiley, New York, 1972; Bartlett, P. D. Nonclassical Ions. Benjamin, New York, 1965.
31. Bezzi, S., Mammi, M., Carbuglio, C. Thio-thiophten: An unusual new type of aromatic system. — Nature (London), 1958, **182**, 247—248; Behringer, H., Reimann, H., Ruff, M. Thio-thiophtene und Sauerstoff-Isologe. — Angew. Chem., 1960, **72**, 415.
32. Wong, L. H. — In: Mellor's Comprehensive Treatise on Inorganic and Theoretical Chemistry, Vol. 5, Suppl. 2, Part 3. Longmans, London, 1981; Muetterties, E. L. (ed.). Boron Hydride Chemistry. Academic Press, New York, 1985.
33. Baeckvall, J. E., Åkermark, B., Ljunggren, S. O. Stereochemistry and mechanism for the palladium(II)-catalyzed oxidation of ethene in water (the WACKER-Process). — J. Am. Chem. Soc., 1979, **101**, 2411—2416.
34. Fontain, E. The generation of reaction networks with RAIN. 2. Resonance structures and tautomerism. — Tetrahedron Comput. Methodol., 1990, **3**, 469—477.
35. Dietz, A. Doctoral Thesis. Technical University Munich, 1992.
36. Vleduts, G. Storage and retrieval of chemical reactions. — Inform. Storage Retr., 1963, **1**, 117—146.
37. Corey, E. J. General methods for the construction of complex molecules. — Pure Appl. Chem., 1967, **14**, 19—37; Corey, E. J. Die Logik der chemischen Synthese: Vielstufige Synthesen komplexer „carbogener" Moleküle (Nobel-Vortrag). —

Angew. Chem., 1991, **103**, 469—479; Angew. Chem. Int. Ed. Engl., 1991, **30**, 455—465; Corey, E. J., Wipke, W. T. Computer-assisted design of complex organic syntheses. — Science, 1969, **166**, 178—192; Corey, E. J., Cheng, X.-M. The Logic of Chemical Synthesis. Wiley, New York, 1989; Corey, E. J., Long, A. K., Rubenstein, S. D. Computer-assisted analysis in organic synthesis. — Science, 1985, **228**, 408—418.

38. Wipke, W. T., Rogers, D. Artificial intelligence in organic synthesis. SST: Starting material selection strategies. An application of superstructure search. — J. Chem. Inform. Comput. Sci., 1984, **24**, 71—81.
39. Ziegler, E. (ed.). Computer in der Chemie. Springer, Berlin, 1984.
40. Lederberg, J. Topological mapping of organic molecules. — Proc. Nat. Acad. Sci. (USA), 1965, **53**, 134—139; 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.
41. Gray, N. A. B. Computer-Assisted Structure Elucidation, **2**. Wiley, New York, 1986.
42. Carhart, R. E., Smith, D. H., Brown, H., Djerassi, C. Application of artificial intelligence for chemical inference. XVII. An approach to computer-assisted elucidation of molecular structure. — J. Am. Chem. Soc., 1975, **97**, 5755—5762; Carhart, R. E., Smith, D. H., Gray, N. A. B., Nourse, J. G., Djerassi, C. GENOA: A computer program for structure elucidation utilizing overlapping and alternative substructures. — J. Org. Chem., 1981, **46**, 1708—1718.
43. Blair, J., Gasteiger, J., Gillespie, C., Gillespie, P. D., Ugi, I. CICLOPS — a computer program for the design of syntheses on the basis of a mathematical model. — In: Wipke, W. T., Heller, S. R., Feldmann, R. J., Hyde, E. (eds.). Computer Representation and Manipulation of Chemical Information. Wiley, New York, 1974, 129—146.
44. Bauer, J., Fontain, E., Ugi, I. Computer-assisted bilateral solution of chemical problems and generation of reaction networks. — Anal. Chim. Acta, 1988, **210**, 123—134.
45. Bauer, J. IGOR2: A PC-program for generating new reactions and molecular structures. — Tetrahedron Comput. Methodol., 1989, **2**, 269—280; Bauer, J., Ugi, I. Chemical reactions and structures without precedent generated by computer programs. — J. Chem. Res., 1982, (S) 298, (M) 3101.
46. Fontain, E., Bauer, J., Ugi, I. Computer-assisted bilateral generation of reaction networks from educts and products. — Chem. Lett., 1987, 37—40; Fontain, E., Bauer, J., Ugi, I. Computergestützte mechanistische Analyse der Streith-Reaktion mit dem Programm RAIN. — Z. Naturforsch., 1987, **42B**, 889—891; Ugi, I., Bauer, J., Fontain, E. Reaction pathways on a PC. — In: Zupan, J. (ed.). Personal Computers for Chemists. Elsevier, Amsterdam, 1990, 135—154.
47. Fontain, E. The problem of atom-to-atom mapping. An application of genetic algorithms. — Anal. Chim. Acta, 1992, **265**, 227—232.
48. Ugi, I., Bauer, J., Brandt, J., Friedrich, J., Gasteiger, J., Jochum, C., Schubert, W., Dugundji, J. Computer programs for the deductive solution of chemical problems on the basis of a mathematical model — a systematic bilateral approach to reaction pathways. — In: Bargon, J. (ed.). Computational Methods in Chemistry. Plenum Press, New York, 1980, 275—300; Dengler, A., Fontain, E., Knauer, M., Stein, N., Ugi, I. Competing concepts in CAOS. — Recueil, 1992, **111**, 262—269.
49. Ugi, I., Bauer, J., Fontain, E., Götz, J., Hering, G., Jacob, P., Landgraf, B., Karl, R., Lemmen P., Schneiderwind-Stöcklein, R., Schwarz, R., Sluka, P., Balgobin, N., Chattopadhyaya, J., Pathak, T., Zhou, X. X. New phosphorylating reagents and protective group techniques for oligonucleotide synthesis, as well as computer assistance in the design of reagents. — Chem. Scripta, 1986, **26**, 205—215.
50. Bauer, J., Fontain, E., Ugi, I. IGOR und RAIN — the first mathematically based general purpose computer programs for the solution of constitutional problems in chemistry and their use as generators of constitutional formulas. — Informal Commun. Math. Chem. (MATCH), 1992, **27**, 31—47.
51. Lohberger, S., Fontain, E., Ugi, I., Müller, G., Lachmann, J. Malonamide derivatives

as by-products of four component condensations. The computer-assisted investigation of a reaction mechanism. — New J. Chem., 1991, **15**, 913—917.
52. Fisher, G. B., Juarez-Brambila, J. J., Goralski, C. T., Wipke, W. T., Singaram, B. Novel conversion of aldehydes to boronic esters. Simultaneous IGOR2 computer generation and experimental observation of an unusual rearrangement of α-aminoboranes. — J. Am. Chem. Soc., 1993, **115**, 440—444.
53. Senff, A., Reichelt, S., Müller, N., Ugi, I. Topological specification of ensembles of molecules as a basis of stereochemical regards. — Theo. Chem., in press.
54. Stein, N. Das sXBE- und sXR-Modell der konstitutionellen Chemie. Doctoral Thesis. Technische Universität München, 1993; Stein, N. New perspectives in computer-assisted formal synthesis design — treatment of delocalized electrons. — J. Chem. Inf. Comput. Sci., in press.

KEEMIA KAKS FORMAALKEELT: NENDE SEMANTIKA JA SÜNTAKS

Ivar UGI, Natalie STEIN, Bernhard GRUBER

Keemilised valemid kui keemilise informatsiooni ja mõtlemise alused on formaalkeele terminid. Molekulaarstruktuuride transleerimist keemilisteks valemiteks saab vaadelda selle keele semantikana. Süntaksi moodustavad keemilistes reaktsioonides toimivad ainete interkonversiooni seaduspärasused. Matemaatika on kasutatav keemia formaalkeelena *be-* ja *r-*maatriksite algebra ja gruppide keemilise identsuse teooria vahendusel. Üldul põhinev keemilise grammatika kontseptsioon võimaldab deduktiivselt lahendada keemia probleeme arvuti abil.

ДВА ФОРМАЛЬНЫХ ЯЗЫКА ХИМИИ — ИХ СЕМАНТИКА И СИНТАКСИС

Ивар УГИ, Натали ШТЕЙН, Бернард ГРУБЕР

Химические формулы — базис химической информации и мышления — это термины формального языка. Семантику этого языка составляет перевод молекулярных структур на язык химических формул. Синтаксис образуется по правилам взаимодействия химических соединений в реакциях. Использование математики в качестве формального языка химии осуществляется через алгебру *be-* и *r-*матриц и теорию идентичности групп. Созданная концепция химической грамматики служит основой для дедуктивного решения проблем химии с помощью компьютеров.