(40) Allen, F. H.; Kennard, O.; Taylor, R. Systematic Analysis of Structural Data as a Research Technique in Organic Chemistry. *Acc. Chem. Res.* **1983**, *16*, 146-153.
(41) Allen, F. H.; Kennard, O. The Cambridge Structural Database: Current Applications and Future Developments. In *Crystallographic Databases*; Allen, F. H., Bergerhoff, G., Sievers, R., Eds.; International Union of Crystallography: Chester, U.K., 1987; pp 55-76.
(42) Cremer, D.; Pople, J. A. A General Definition of Ring-Puckering Coordinates. *J. Am. Chem. Soc.* **1975**, *97*, 1354-1358.
(43) Altona, C.; Sundaralingam, M. Conformational Analysis of the Sugar Ring in Nucleosides and Nucleotides: A New Description Using the Concept of Pseudorotation. *J. Am. Chem. Soc.* **1972**, *94*, 8205-8212.

Chemical Abstracts Service Chemical Registry System. 13. Enhanced Handling of Stereochemistry†

J. E. BLACKWOOD, P. E. BLOWER, JR.,* S. W. LAYTEN, D. H. LILLIE, A. H. LIPKUS, J. P. PEER, C. QIAN, L. M. STAGGENBORG, and C. E. WATSON

Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43210

Received January 14, 1991

CAS registers stereoisomers using *text descriptors* related to the stereochemical descriptors of the corresponding chemical names. This system works well for the unique registration of stereoisomers, but it is difficult to relate the text descriptor to the atoms and bonds of the Registry connection table. This limits its usefulness for substructure search or display of stereochemistry in the structure diagram. CAS is currently augmenting the Registry connection table with atom/bond-specific stereodescriptors. This article focuses on two aspects of this work: the representation of stereochemistry in the connection table and techniques for converting the Registry structure file to the stereoaugmented format.

INTRODUCTION

The Chemical Abstracts Service (CAS) Chemical Registry System marked its 25th anniversary in 1990. The Registry System is a computer-based system that identifies substances on the basis of their molecular structure. Begun originally to support substance indexing for *Chemical Abstracts* (CA), the Registry System has influenced the work of chemists and other scientists around the world. Perhaps best known as the source of CAS Registry Numbers and of the Registry File on the STN International network, the Chemical Registry System provides a foundation for substance identification used by the scientific community worldwide, with over 10 million substances currently on file. The CAS Registry Number, which links the structure with the *CA* index name and other data, is used for chemical substance identification by many governmental agencies and industrial organizations.

The Registry System has evolved over a period of years. The initial 1965 version, Registry I, was designed to be as specific as possible. It registered only fully defined organic compounds. Other substances such as inorganic compounds, polymers, and compounds with partially known structures were manually assigned Registry Numbers on the basis of their structural diagrams, names, or molecular formulas. In 1968 the second version of the CAS Registry System, Registry II, extended machine registration to include inorganic substances, coordination compounds, polymers, mixtures, alloys, and certain incompletely defined substances. Registry III, the current version of the Registry System, became operational in late 1973. Although no changes were made in the basic algorithmic techniques for registration, Registry III made a major adjustment to the Registry records so that the system would more effectively support CAS index nomenclature generation and computer-based structure output. The design, content, functions, statistics, special features, and input structure conventions have been described in detail in previous papers.[1-11]

The boundaries of the Chemical Registry System are constantly being extended, both in content and the manner in which information is added. In the past 25 years the CAS Chemical Registry System has evolved from a production tool for CAS publications and services to a vital, useful service for the scientific community worldwide. And with this expanded role, the Registry System and related services will become even more responsive to the needs of scientists and engineers.

CAS is currently developing Registry IV. Improvements currently underway include enhanced alloy processing, improvements in structure display, faster registration due to a new online editorial input system, addition of biosequence data, and introduction of display and search of stereochemical information in the Registry File. Additional items being investigated include modifications to allow scientists to search using a variety of conventions for representing substances and better support for industrial and engineering users of information about metals and alloys, polymers, ceramics, and composites.

Stereochemical representation[3,12,13] has been an integral part of the CAS Chemical Registry System since its inception as Registry I. Stereoisomers are considered to be different, individual substances, and each is recognized as unique. The unique identification of stereoisomers permits the storage and retrieval of information collected from scientific literature about a specific isomer. Currently stereochemical data is only recorded in the chemical name and a controlled vocabulary field known as the CAS Text Descriptor. This is being enhanced by a technique for recording specific atom and bond stereochemistry in the connection table,[1] the Registry structure record maintained for each substance.

REPRESENTATION

Several different techniques will be used to represent atom/bond-specific stereochemistry in the connection table record, depending on the type of stereocenter being described. Tetrahedral stereocenters, stereogenic double bonds, and allenes are described by using a parity descriptor, similar to that described by Petrarca et al.[14] and adapted by Wipke and Dyott[15] in their Stereochemically Extended Morgan Algorithm. The relative Cahn-Ingold-Prelog[16,17] (CIP) ranks of the neighboring atoms are recorded along with the parity data. This information allows straightforward calculation of an *R*

† Dedicated to Professor Michael Lynch, whose participation in work on computer representation of stereochemistry more than 20 years ago helped provide the basis for the work reported here.

0095-2338/91/1631-0204$02.50/0

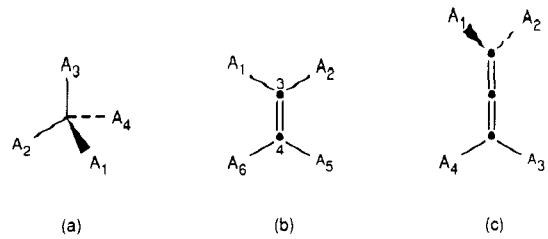

Figure 1. Illustration of *even* parity for (a) tetrahedral, (b) double bond, and (c) allene stereocenters.

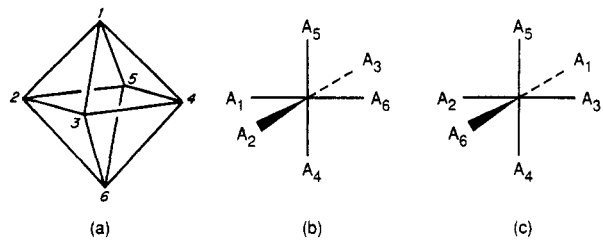

Figure 2. Representation of an octahedral complex: (a) geometric model with labeled coordination sites; (b) schematic octahedral complex, subscripts indicate the relative atom numbers; (c) another view of the same schematic complex.

or *S* descriptor for the center according to the CIP rules.

Tetrahedral Centers. The configuration of tetrahedral centers is described by a parity descriptor that depends on the atom numbers of the atoms adjacent to the center. The rule for determining the parity of tetrahedral centers can be stated as follows: The center is viewed from the lowest numbered neighbor. If the path generated by citing the remaining neighbors in increasing order is clockwise, then the parity of the center is *even*, otherwise it is *odd*; see Figure 1a.

Double Bonds. For double bonds, the following convention is adopted: The double bond is viewed from the normal to the plane containing the double bond and its neighboring atoms. A parity is determined for each end point of the double bond by determining whether the ascending sequence numbers of neighbors proceeds in a clockwise (*even* parity) or anticlockwise (*odd* parity) manner. If the parities of the end points are the same, then the bond has *even* parity. If the parities of the end points are opposite, then the bond has *odd* parity; see Figure 1b.

Allenes. A parity descriptor is assigned to an allene system according to the following rule: The center is viewed along the allene axis from either end. If the path generated by citing the two neighbors on the near end and the lowest numbered neighbor on the far end in increasing order is clockwise, then the parity of the center is *even*, otherwise it is *odd*; see Figure 1c. In addition to the parity information, the atoms of the double-bond system and the relative CIP ranking of the atoms attached to the end points are included with the stereochemical information.

Coordination Centers. Coordination centers are described by using a *geometry descriptor* and a *configuration descriptor*. The geometry descriptor indicates which idealized model[18,19] describes the coordination geometry; the configuration descriptor is analogous to a parity descriptor and describes the geometrical arrangement of ligands around the center. Models have been defined for the 17 nontetrahedral geometries[20] recognized by CAS.

The model for an octahedral complex is illustrated in Figure 2a with the coordination sites labeled from 1 to 6. Figure 2b shows a schematic octahedral complex with ligating atoms A_1, ..., A_6, where the subscripts indicate the relative atom numbers in the connection table. Thus, A_1 represents the lowest number ligating atom in the connection table, A_2 the second lowest, etc. The configuration descriptor associated with the octahedral center in the connection table would be a list of the site numbers of the ligating atoms listed in ascending atom number order. For the schematic example in Figure 2b, the configuration descriptor would be **235614** because A_1 occupies site 2, A_2 occupies site 3, A_3 occupies site 5, etc.

Table I. Site Permutations for the Octahedral Model

1 2 3 4 5 6	3 1 2 6 4 5	5 1 4 6 2 3
1 3 4 5 2 6	3 2 6 4 1 5	5 2 1 4 6 3
1 4 5 2 3 6	3 4 1 2 6 5	5 4 6 2 1 3
1 5 2 3 4 6	3 6 4 1 2 5	5 6 2 1 4 3
2 1 5 6 3 4	4 1 3 6 5 2	6 2 5 4 3 1
2 3 1 5 6 4	4 3 6 5 1 2	6 3 2 5 4 1
2 5 6 3 1 4	4 5 1 3 6 2	6 4 3 2 5 1
2 6 3 1 5 4	4 6 5 1 3 2	6 5 4 3 2 1

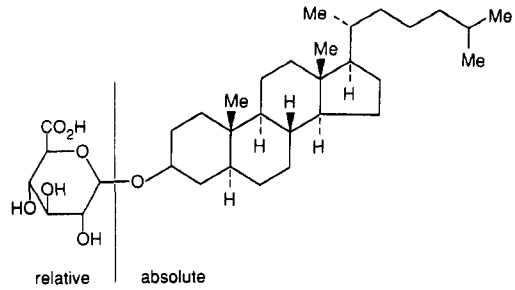

Figure 3. Substance with partial stereochemistry, cholestan-3-yl-glucopyranosiduronic acid.

Because of the inherent symmetries of the coordination models, there are generally several equivalent descriptors for the same configuration. Figure 2c shows another view of the same complex obtained by a clockwise rotation about the vertical axis. So an equivalent configuration descriptor for the complex is **524613**. In general, any view of the octahedron obtained by rotating it about one or more of its symmetry axes leads to an equivalent configuration descriptor. In all, there are 24 equivalent descriptors for each distinct configuration of an octahedral complex.

To account for equivalent representations, there will be a table of symmetry operations for constructing equivalent representations maintained for each coordination model. Thus, associated with each model is a *Site Permutation Table* giving the symmetry operations which can be used for constructing equivalent representations. Table I is the site permutation table for the octahedral complex. Each entry[21] in one of the site permutation tables is a numerical sequence, giving a permutation of the atoms occupying the sites of the model. It describes a rotational symmetry operation, and the complete table lists all such operations for the corresponding geometry. At each position *i* in a sequence is the numeric label of the destination site under the symmetry operation. Thus, if *j* is the *i*th digit in the sequence, then the atom occupying site *i* moves to site *j* under the symmetry operation. For example, the second entry in the octahedral table is **134526**. This is interpreted as follows: the atom in site 1 moves to site 1 (i.e., remains in place), the atom in site 2 moves to site 3, the atom in site 3 moves to site 4, etc., corresponding to a clockwise rotation about the vertical axis. These site permutation tables will be used for comparing the stereochemistry of two coordination centers with the same geometry.

Partial Stereochemistry. The four methods just described can be used to represent the absolute configuration of tetrahedral, double bond, allene, and coordination stereocenters. In order to use these methods in the CAS Registry System, we extended them to cover cases where stereochemistry is only partially known. This can be illustrated by the cholestane derivative in Figure 3. In this example, the absolute stereochemistry is known for all centers in the cholestane substructure, except atom 3 which is unknown. However, the configurations of four of the centers on the carbohydrate

Table II. Registry Statistics for Stereospecific Substances

type	number	percent
systematic nomenclature	1 638 000	16.2%
stereoparents	786 000	7.8%
coordination compounds	277 000	2.7%
total	2 701 000	26.7%

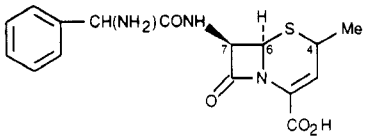

Figure 4. Substance with a systematic stereodescriptor, [6*R*-(6α,7β)]-7-[(aminophenylacetyl)amino]-4-methyl-8-oxo-5-thia-1-azabicyclo[4.2.0]oct-2-ene-2-carboxylic acid.

moiety are only known relative to each other, except the anomeric carbon which is also unknown. Thus, the four centers on the pyran ring can exist in two states: One state is given by parity descriptors derived from the diagram as shown. But an equally valid representation is obtained by inverting all parity descriptors in the pyran ring—independent of the centers in the cholestane substructure.

To handle cases like this, we define a *group descriptor*, which can be associated with a specific subset of stereogenic atoms. The group descriptor indicates which stereocenters are in the group and whether the stereochemistry for centers within the group is relative or racemic. For example, all four centers on the pyran ring in Figure 3 comprise a relative set and would be assigned a common group number. Group descriptors are not assigned to absolute or unknown stereocenters.

FILE CONVERSION

Stereochemistry is expressed in CA index names in three ways: systematic stereodescriptors, stereoparents, and coordination compound stereodescriptors. We have developed a separate procedure for converting each of these three types of stereodescriptors to atom/bond-specific descriptors in the Registry connection table. Table II gives statistics for the Registry File as of October 1990. At that time, the total file contained over 10 million machine-registered substances. Current plans call for file conversion to be done in three stages. The first stage, which addresses substances with systematic stereodescriptors, is now well underway.

Systematic Stereodescriptors. Stereodescriptors for systematically named substances use a combination of absolute, relative, and rotational terms. An absolute term is assigned to a *reference* center, and all other centers are related to it using relative terms. The absolute terms *R* and *S* follow the system developed by Cahn, Ingold, and Prelog. Various kinds of relative descriptors are used. The most readily interpreted as those that describe the stereochemical relationships of substituents on a ring system. In general terms,[22] *cis* and *trans* are used for small rings with only two stereogenic atoms; *exo*, *endo*, *syn*, and *anti* are used for bicyclo[X.Y.Z]anes; and α and β are used for other situations where the stereocenters are in the same ring system as the reference center. The relative descriptors *R*/S** are used for all other tetrahedral stereocenters, and *E/Z* are used for stereogenic double bonds. The use of these terms is illustrated in Figures 4 and 5.

If there is any danger of confusion, descriptor terms are associated with locants. It is important to note, however, that locants do not refer to a single numbering system for the total substance. Instead, the locants for each nomenclature fragment start with *one* (1). In Figure 5 for example, the locants in 1α,4aβ,8aα refer to the fragment *naphthalenyl*, while the locants in 2*R**,3*S** refer to the fragment *oxirane*.

The general approach for converting Registry substances with systematic stereodescriptors consists of three stages. The

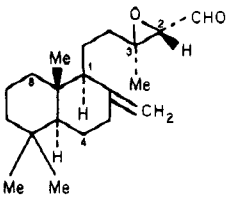

Figure 5. Substance with a systematic stereodescriptor [1*S*-[1α-(2*R**,3*S**),4aβ,8aα]]-3-[2-(decahydro-5,5,8a-trimethyl-2-methylene-1-naphthalenyl)ethyl]-3-methyl oxiranecarboxaldehyde.

first stage attempts to assign the nomenclature stereodescriptors to the appropriate atom or bond of the connection table. This is complicated by the fact that stereodescriptors use nomenclature locants which are unrelated to the atom numbers in the Registry connection table. In the early 1970s, CAS developed a *nomenclature translation* program[23] to convert systematic names of organic compounds to connection tables. This program determines the correspondence between nomenclature locants and atom numbers which is used to map stereodescriptors into specific atoms.

The second stage identifies and resolves any problems which arise because of insufficient or ambiguous descriptors. Generally, we expect a one-to-one correspondence between the stereodescriptors in the name and the stereocenters in the connection table. There are, however, legitimate cases where there are more stereocenters than stereodescriptors. First, some structural configurations are implicitly understood to be cis, such as bridges in bicyclic systems and double bonds in small rings. The other situation involves cases of valid partial stereochemistry. These include derivatives of penicillins and cephalosporins and substances with unknown configuration at a heteroatom. Figure 4 is an illustration of valid partial stereochemistry with two stereocenters of unknown configuration, one at atom 4 and the other in the phenylglycine side chain.

The third stage of conversion translates the nomenclature stereodescriptor to an atom/bond-specific parity descriptor. This process is straightforward for *R/S*, *R*/S*, and *E/Z*. Translating the relative ring descriptors (e.g., cis, α, or exo) is somewhat more complex because it requires identifying the *ring plane* and the substituents referred to by the relative descriptor.

The process of translating systematic stereodescriptors to connection table parity descriptors can be illustrated for the substance in Figure 5.

1. The stereodescriptor is parsed into three sets: the absolute descriptor 1*S* and two sets of relative descriptors 1α, 4aβ, 8aα and 2*R**,3*S**.
2. The chemical name is passed through the nomenclature translation program which generates a connection table and an atom-locant correspondence table.
3. The Cahn–Ingold–Prelog procedure is invoked for each candidate stereocenter in the connection table, and this information is used to identify stereocenters.
4. The nomenclature stereodescriptors are mapped into connection table atoms by using the atom-locant correspondence table created in step 2. The absolute descriptor is assigned to atom 1 of the naphthalene ring, which is the reference center.[24] Then the two sets of relative descriptors 1α, 4aβ, 8aα and 2*R**,3*S** are associated with the appropriate centers in the naphthalene and oxirane rings, respectively.
5. The program checks for inconsistencies between the stereodescriptors in the name and the stereocenters in the connection table, but none are detected.
6. The absolute descriptor is processed by applying Procedure A with arguments S_CTR set to the ref-

erence center and DESCRIPTOR = *S*.

7. Each term in the first relative descriptor set is processed.

7a. The term 1α refers to the reference center and arbitrarily designates the α-side of the naphthalene ring as that on which the carbon substituent lies. When 1α is encountered, REF_CTR is set to atom 1 of the naphthalene ring, the reference center; REF_SUB is set to the side-chain carbon, the cited substituent attached to REF_CTR; and PARITY is set to the parity descriptor previously assigned to REF_CTR in step 6.

7b. The processing done for the remaining terms 4aβ and 8aα in the same and will be explained for 4aβ. When this term is encountered, REL_CTR is set to atom 4a of the naphthalene ring, the relative center; REL_SUB is set to hydrogen, the only substituent attached to REL_CTR; and DESCRIPTOR is set to BETA, indicating that the cited substituents attached to REF_CTR and REL_CTR are on opposite sides of the ring. Finally, RING_PATH is constructed which lists the atoms describing a cyclic path containing REF_CTR and REL_CTR and defines the reference plane. The atoms in RING_PATH are listed in the order they are encountered as one proceeds around the ring in either direction. Then Procedure B is applied, which calculates the parity of atom 4a.

8. Finally, each term in the second relative descriptor set is processed. The interpretation of *R*/S** is analogous to *R/S*, except that *R** means the center has the same configuration as the reference center, and *S** means it has the opposite configuration. In this case, the reference center has the *S* configuration. Thus, the descriptor *2R*,3S** means that atom 2 of the oxirane ring has the *S* configuration and atom 3, the *R* configuration. This is processed by applying Procedure A twice, first with S_CTR set to atom 2 of the oxirane ring and DESCRIPTOR = *S* and second with S_CTR set to oxirane atom 3 and DESCRIPTOR = *R*.

Limitations on Algorithmic Conversion. Development of a software system to convert the systematic stereodescriptors to parity descriptors is now well underway. Statistics we have gathered during implementation indicate that we can expect to convert approximately 65% of these substances algorithmically.

There are several different reasons why algorithmic conversion will fail for the remaining substances with systematic stereodescriptors. About 8% of these substances are coordination compounds, incompletely defined substances, and polymers that the algorithm will not attempt to convert at this time. Another 10% of the substances do not have the complete systematic stereodescriptor required for conversion. Some of these contain stereoparent name fragments—typically carbohydrates—that we may eventually be able to convert by the algorithm for stereoparents. Others have a descriptor of *stereoisomer* that cannot be converted algorithmically.

For about 11% of the substances, the conversion algorithm will fail because the nomenclature translation program cannot generate a connection table. This number includes substances named prior to the Ninth Collective Index Period (1972–1976). So some improvement in the success rate of algorithmic conversion could be obtained by making extensions to the nomenclature translation program. Finally, we estimate another 6% of the substances will fail conversion because of micellaneous limitations in the algorithm. Substances that fail the algorithmic processing will be converted manually by entering

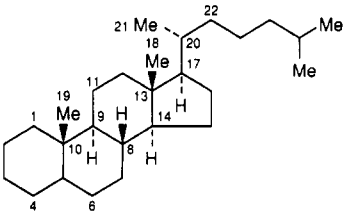

Figure 6. Stereoparent diagram for cholestane with nomenclature atom numbers.

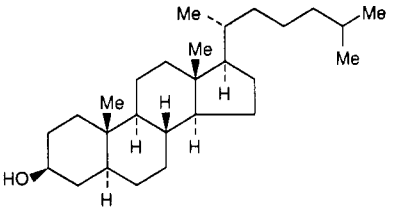

Figure 7. Stereoparent with modified stereochemistry, text descriptor 3B,5A.CHOLEST.

the structure diagram with stereochemistry. Thus, we expect a significant manual effort will be necessary to add connection table level stereochemistry to the Registry File.

Stereoparents. The second phase of file conversion addresses stereoparents; we have just started implementing these procedures. Stereoparents are derived from the common names of natural products. The following discussion focuses on stereoparents from the steroid, alkaloid, or terpene classes. The stereoparent is the simplest member of a family of related, cyclic structures with complex stereochemistry. The stereochemistry is given by a stereospecific structure diagram which appears in the *CA Chemical Substance Index*; this is illustrated by cholestane (Figure 6). Stereochemistry is indicated in the diagram by dotted lines (α bonds), if the substituent is below the plane of the ring, and wedged lines (β bonds), if it is above the plane. The diagram also gives the numbering system of the stereoparent so that locants can be used to indicate modifications of the parent structure.

The stereochemical descriptor for a stereoparent expresses the additions and differences between the structure reported and the stereoparent diagram. The naming conventions allow for specification of abnormal or additional stereochemistry and modifications in the molecular topology of the parent structure. Additional or abnormal stereochemistry on ring atoms of the stereoparent is designated by citing the atom number and direction (α or β) of the preferred substituent as illustrated in Figure 7. The structure has an additional 3-hydroxy group on the β-side of the ring. The stereochemistry of the hydrogen substituent at the 5 position must also be described because no stereochemistry is implied at that position in the cholestane stereoparent. Since normal stereochemistry is retained at all other stereocenters, the stereodescriptor is 3β,5α.

Four types of changes to the topology of the ring system can be indicated by keywords: *nor*, ring contraction; *hom*, ring expansion; *sec*, ring cleavage, and *cyc*, cyclization. A structure derived from ring cleavage of the 6–7 bond in the cholestane system is illustrated in Figure 8. Again the stereochemistry at the 5 position must also be described; in this case, the aldehyde group is the preferred substituent. The name (5α)-6,7-secocholestane-6,7-dial expresses variations in stereochemistry from that of cholestane.

The procedure for converting stereoparents uses the Text Descriptor. This conveys the same stereochemical information as the index name but in a more compact format. The text descriptor is divided into four fields separated by periods and has the format:

⟨stereo modifications⟩.⟨stereoparent⟩.⟨substituents⟩.⟨ring modifications⟩

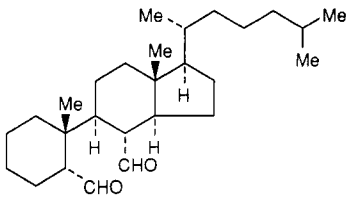

Figure 8. Stereoparent with modified topology, text descriptor 5B.CHOLEST..2,6,7-SEC.

Table III. Stereochemical Information for the Cholestane Template

atom	parity	substituent	orientation	stereocenter
1	odd	D	β	no
2	even	D	β	no
3	even	D	β	no
4	even	D	β	no
5	even	H	β	unspec
8	odd	H	β	yes
9	even	H	α	yes
10	odd	19	β	yes
13	even	H	β	yes
14	even	H	α	yes
17	odd	20	β	yes
20	odd			yes

The ⟨stereo modifications⟩ field describes abnormal stereochemistry of the stereoparent or the stereochemistry of additional substituents. The ⟨stereoparent⟩ field is an abbreviation of the stereoparent name; e.g., CHOLEST. The ⟨substituents⟩ field describes amino acid or carbohydrate substituents. The ⟨ring modifications⟩ field describes changes in topology or node constitution of the stereoparent ring system.

The procedure for converting substances described as stereoparents will be based on the use of stereospecific templates—one for each stereoparent. Each template will be a stereospecific connection table constructed algorithmically from the stereospecific structure diagram, illustrated by cholestane in Figure 6. The template connection tables will use the nomenclature system of atom numbering, and the parity descriptor for each stereocenter will be based on this numbering system. Parity descriptors will be calculated directly from the structure diagram by the procedure described by Wipke and Dyott.[25] The relative α/β descriptors that are used to cite modified stereochemistry on the ring system depend on the *view* of the stereoparent. Thus, in addition to parity descriptors, we will also record *view information* for each cyclic stereocenter. This gives the orientation of the stereobond and attached substituent[26] and will allow us to interpret any abnormal stereochemistry cited in the text descriptor. For stereocenters with unspecified configuration such as atom 5 in cholestane, we calculate parity assuming a β-hydrogen but mark the center as unspecified. The stereochemical information that will be recorded for the cyclic stereocenters of the cholestane template is shown in Table III. Only parity descriptors are calculated for acyclic stereocenters. View information is not needed because the relative α/β descriptors are not used for such centers.

We will also need view information for ring atoms in the template that are not chiral centers to interpret any additional stereochemistry at ring atoms cited in the text descriptor. For such centers, we will calculate a *pro-parity* descriptor. Pro-parity gives the parity that the nonchiral center would have if the β-hydrogen were replaced by a more preferred substituent. The stereochemical information recorded for nonchiral centers is illustrated in Table III for atoms in the A-ring of cholestane; the *D* listed under substituent denotes the distinguished β-hydrogen.

The procedure for adding atom/bond-specific stereodescriptors to the connection table, in simplified form, consists of the following steps: (1) Retrieve the Registry File substance and parse its stereodescriptor into components that identify the stereoparent and describe the topological and stereochemical modifications. (2) Modify the topology of the template as described. (3) Perform a match by using the modified template to locate the stereoparent portion of the file structure. (4) Transfer parity from the template to the file substance. This processing can be illustrated for (3β,5α)-cholestan-3-ol, Figure 7.

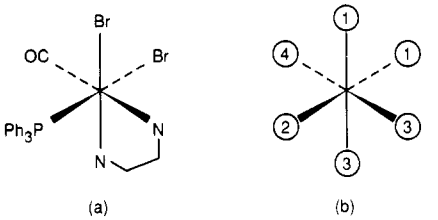

Figure 9. Example of coordination stereodescriptors: (a) stereodiagram with text descriptor OC-6-32-C (bidentate ligand is $NH_2CH_2CH_2NH_2$), (b) schematic model with CIP rank assignments.

1. The substance is retrieved from the Registry File, and its text descriptor is parsed into components that identify the stereoparent and describe the topological and stereochemical modifications. The stereoparent is CHOLEST, and the stereochemical modifications are 3B, 5A.
2. The CHOLEST template is retrieved. In this example, no topological modifications were described in the text descriptor. If there were, the template connection table would be changed appropriately.
3. A match is performed between the modified template and the substance to locate the stereoparent portion of the substance.
4. Parity is transferred from the template to the file substance by applying the following logic for each template node, T_NODE:

4a. Let S_NODE be the matching substance node determined by step 3 and let S_DESC be the stereodescriptor for S_NODE. Note that for most nodes, S_DESC will be *nil*.

4b. If (S_DESC ≠ *nil*) then
If (S_DESC = *R* or *S*) then apply Procedure A.
Else if (S_DESC = X) then set the descriptor of S_NODE = unknown.
Else apply Procedure C.
Else if (T_NODE is a stereocenter) then apply Procedure C.
Else if (S_NODE is a stereocenter) OR (T_NODE is an unspecified stereocenter) then set the descriptor of S_NODE = unknown.

Coordination Compounds. The third phase of file conversion addresses coordination compounds; we expect to begin implementing these procedures sometime in 1991. Stereodescriptors for coordination centers[27] consist of four parts: (1) a symmetry site term, (2) a configuration number, (3) a chirality symbol, and (4) a ligand stereochemistry segment. The symmetry site term identifies the molecular geometry at the coordination center; e.g., square planar (SP-4), trigonal bipyramid (TB-5), and octahedral (OC-6). The configuration number is a sequence of digits describing the spatial arrangement of ligands about the center in terms of their CIP rank. The details of determining the configuration number vary somewhat from one geometry to another, but they can be illustrated for the OC-6 case (Figure 9) in which the configuration number has two digits. The first digit gives the CIP rank of the atom *trans* to the most preferred atom. The second digit gives the CIP rank of the atom *trans* to the most preferred of the remaining atoms—those in the plane perpendicular to the principal axis. In cases where there are sets

of constitutionally equivalent atoms, a choice can arise in selecting the preferred atom. In such cases, preference is given to the atom *trans* to the least preferred atom; this is called the *trans maximum* rule. The chirality symbols *C* (clockwise) and *A* (anticlockwise) are used to distinguish enantiomeric configurations. This is determined by viewing the complex along the principal axis from the highest ranking atom and tracing a clockwise or anticlockwise path from the highest ranking atom in the perpendicular plane to the higher ranking of the two atoms adjacent to it. These rules are illustrated in Figure 9. For octahedral complexes with two or three bidentate ligands oriented in a skew configuration, the helicity symbols Δ and Λ are used instead of chirality symbols.

Algorithms for converting the stereochemistry of coordination compounds from text descriptors to configuration descriptors have been developed for the following geometries: square planar, square pyramidal, trigonal bipyramidal, and octahedral. Tetrahedral coordination geometry has been excluded since it can be handled by using the methods developed for tetrahedral carbon. Algorithmic conversion of trigonal prismatic geometry, as well as all geometries with coordination numbers higher than six, has been deemed unnecessary since such compounds are relatively uncommon and can be converted manually.

The stereochemistry of a coordination compound can be reconstructed from its topological structure if the associated stereochemical text descriptor is properly interpreted. The first three parts of the text descriptor (i.e., symmetry site term, configuration number, and chirality symbol, if present) are needed. The procedure can be outlined as follows: (1) Rank the ligand atoms attached to the coordination center according to the CIP rules. (2) Select the geometric model, as indicated by the symmetry site term, and assign the CIP rank numbers to the sites on the model so that they are consistent with the configuration number and any chirality symbol. (3) Find all possible ways to map the ligand atoms to the model sites so that the CIP rank for each atom is the same as that of its site and any chirality in the resulting structure agrees with the text descriptor. (4) Store each structure produced if it has not been produced and stored previously. The conversion is considered successful if one and only one structure results; manual review is needed otherwise.

An example of this procedure applied to a coordination compound taken from the Registry File will be described in detail. The topological structure of this compound is given in Figure 10a. The CIP rank numbers for the ligand atoms attached to the metal are also shown. In this example, a four-way tie among the CIP ranks is found due to the presence of topologically equivalent atoms. The occurrence of such ties is expected to be quite common. The stereochemical text descriptor for this compound is OC-6-33, which can be parsed into symmetry site term (OC-6), configuration number (33), and chirality symbol (*nil*, in this case). The symmetry site term of OC-6 indicates that the octahedral model for coordination geometry is to be used. The assignment of the CIP rank numbers to the sites of this model can be made in a stepwise manner by interpreting each digit of the configuration number 33 according to the rules used for generating the number. The first digit means that a rank 3 is trans to the highest priority rank, which is always rank 1. The second digit means that a rank 3 is trans to the highest ranking atom in the perpendicular plane, which in this example is rank 2. The remaining two ranks, both 3 in this example, form the third trans pair. The way in which these ranks are assigned to specific sites on the model is arbitrary as long as the three trans relationships deduced from the configuration number are preserved. However, if there is a chirality symbol in the text descriptor (there is none in this example), the chirality of the

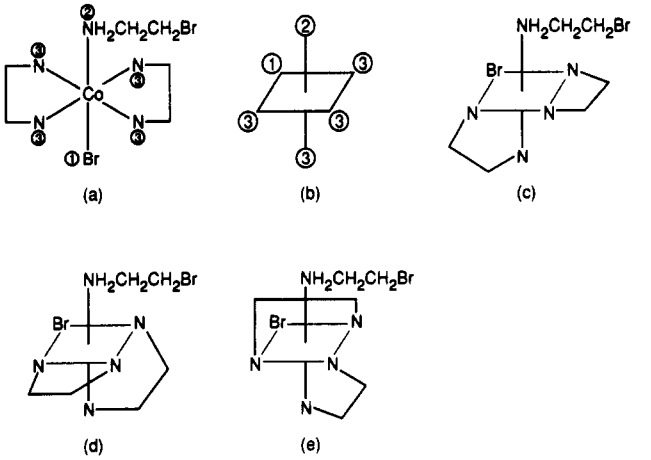

Figure 10. Reconstructing the stereochemistry of a coordination compound from the text descriptor OC-6-33 (bidentate ligands are $NH_2CH_2CH_2NH_2$): (a) topological diagram of the compound; (b) CIP ranks placed on octahedral model according to text descriptor; (c)–(e) diagrams illustrating possible stereochemical structures for the compound.

assigned model is determined, and if it is opposite to that of the given chirality symbol, the site assignments of one of the trans pairs must be reversed. Figure 10b shows an octahedral model with the assignments of CIP rank numbers for this example.

It is next necessary to map each of the ligand atoms to a model site having the same CIP rank. It is not sufficient to perform this mapping in only one way because in some cases there may be two or more stereochemical structures consistent with the stereochemical text descriptor (as will be shown below, this situation arises when there are unrealistically strained structures consistent with the text descriptor). To ensure that no realistic structures are missed, all the possible mappings between ligand atoms and model sites are found with the use of standard combinatorial computing techniques, e.g., backtrack searching. In addition, a test is performed to reject those mappings in which the CIP ranks of atoms and sites do not agree.

The generation of all mappings in the present example can produce only three different structures, panels c–e in Figure 10. But structures that can be obtained from more than one mapping will be produced more than once. This is the case for each of the structures shown. The conversion algorithm must therefore avoid storing structures each time they are produced by a different mapping. Redundant storage of the same structure is avoided by comparing structures, as they are produced, with any previously stored structure. The comparison is carried out by searching for topological matches between the current structure and the stored structure. When a topological match is found, the site permutations for the corresponding geometric model are examined to determine whether one structure can be rotated so that it coincides with the other. The search continues until the current structure is shown to be redundant or until no more topological matches are found, in which case the current structure is stored.

When there is a chirality symbol in the text descriptor, the algorithm must determine that a structure has the given chirality before storing it. For instance, if the text descriptor for the example were OC-6-33-Λ (rather than OC-6-33) then only structure c, which is the Λ stereoisomer, would be stored. When there is no chirality symbol in the text descriptor, as in the example, it does not necessarily mean there is no chirality in the original structure; it may mean only that the chirality is unknown. Therefore, when no chirality symbol is given, the algorithm does not store a structure if its enantiomer is already stored, but instead flags the stored enantiomer as

relative, which indicates that the specific chirality shown is arbitrary since the actual chirality was not given in the text descriptor. To determine whether the enantiomer of a structure has been stored, the algorithm simply reverses the sites of one of the trans pairs of atoms and performs the same structure comparison described previously. For this example, the output of the algorithm is structure c or d, whichever is produced first, along with the *relative* flag. This output describes the stereochemistry of the compound in Figure 10a as reconstructed from its text descriptor.

Structure e is achiral, but cannot be rejected by the algorithm on this basis since the given text descriptor has no chirality symbol. However, this structure is unrealistically strained due to the fact that the bridge between two of the nitrogen atoms is clearly too small to allow a trans relationship. In order to detect this type of unrealistic structure, the algorithm includes a test that examines each trans pair of atoms within the same ligand, compares the length of the smallest (topological) path connecting them within the ligand to a preset minimum allowed pathlength, and rejects the structure if the former is smaller than the latter. This test will immediately reject structure e when it is produced (thus ensuring that there is only one output structure). The minimum allowed pathlength must be set to a value small enough that it will not reject "borderline" structures in which the strain is considerable but not impossible; otherwise, some legitimate structures might be rejected.

IMPLICATIONS

Chemists have long been aware of the importance of stereochemistry. In the past few years state-of-the-art synthesis, separation, and analysis techniques have promoted greater practical application of stereochemistry, especially in the drug development process. Approximately half of the drugs now in the marketplace are chiral compounds.[28] Access to information about specific stereoisomers is becoming increasingly important to users of *Chemical Abstracts*.

Although a nomenclature approach to uniquely identifying stereoisomers serves CAS's internal requirements for support of CA indexing, this approach does not completely support the needs of scientists accessing the Registry system via CAS ONLINE. In particular, the Registry File does not currently support stereo display and search. CAS's efforts to include stereo information in the connection table are an excellent example of the new philosophy of Registry IV. Registry is no longer viewed as only supporting CAS internal operations but rather all users of the Registry System.

Addition of stereochemistry to Registry connection tables makes this information available in a form more accessible for computer handling and manipulation. This will permit stereochemical information to be used to enhance other services which will benefit both internal and external users of the CAS Registry File. The first of these enhancements will be to generate stereospecific images for structure display. In the Registry File, structure images are constructed from the connection table via an Algorithmic Structure Display (ASD) program. ASD is now being completely revised both to improve the general display quality and to provide stereospecific structure images. Display of stereochemistry will follow normal conventions with dashed and wedge bonds used to show the orientation of substituents. Display of the CIP *R*/*S* descriptors will also be available on request. One important requirement for the display of a stereospecific structure is that the structure be recognizable and familiar both in terms of shape and overall orientation. Thus, a key feature of the revised ASD program is the use of highly specific templates—especially for carbohydrates, steroids, and other stereoparents.

Another benefit of the stereoaugmented connection table is that it permits stereochemical enhancement of substructure search. The incorporation of stereochemical information into substructure search should be, for the most part, a straightforward extension of topological search techniques. However, when that information involves relative, as opposed to absolute, stereochemistry, the search problem can be somewhat more complicated.[29] The addition of stereochemical capabilities implies that the user could specify a query with stereochemistry and be given an answer set of structures each of which contains the query with the desired stereochemistry. Thus, a user will be able to find a specific stereoisomer without sifting through hundreds of "false drops". This should be especially important when dealing with natural products.

CONCLUSION

CAS has undertaken a multi-year effort to renovate the Registry System. As part of this effort, we are currently engaged in adding atom/bond-specific stereodescriptors to the connection tables of stereospecific substances already on the Registry File. These extensions will, in turn, permit us to enhance structure display and substructure search to accurately use the stereochemical information available. We have defined a representation that covers the types of stereochemistry cited in CA index names and localized to an atom or bond, including cases where stereochemistry is only partially known. In addition, we have developed algorithms for converting the stereodescriptor in the CA index name to atom/bond-specific descriptors in the connection table. We expect that 60–75% of the stereospecific substances on the Registry File can be converted algorithmically. Substances that fail the algorithmic processing will be manually converted by entering the structure diagram with stereochemistry.

PROCEDURES

Procedure A. Calculation of Parity for Atom Stereocenters. This procedure calculates a parity descriptor for atom stereocenters (S_CTR) given an *R*/*S* descriptor (DESCRIPTOR). For the manipulation of neighbor lists, we use placeholders for H-isotopes (H, D, T) and phantom substituents (e.g., a lone pair) on trivalent stereocenters; suitable values are T = 996, D = 997, H = 998, phantom = 999.

A1. Set NBRS to the neighbors of S_CTR listed in the order determined by the CIP procedure.

A2. Calculate *T*, the total number of transpositions needed to arrange NBRS in ascending order.

A3. If DESCRIPTOR is *S*, $T = T + 1$.

A4. If $T \pmod 2 = 0$ then parity = even.
Else parity = odd.

Procedure B. Translation of Relative Descriptors for Ring Substituents. This procedure is used for translating the relative descriptors for ring substituents (viz., α, β) to parity descriptors. It can also handle the other relative ring descriptors (viz., *cis*, *trans*, *exo*, *endo*, *syn*, *anti*) provided these descriptors are first translated to ALPHA or BETA before the procedure is called.

B1. Let REF_NBR_1 and REF_NBR_2 be the left and right neighbors of REF_CTR on RING_PATH, considered as a circular list. Let REL_NBR_1 and REL_NBR_2 be the left and right neighbors of REL_CTR on RING_PATH.

B2. List the neighbors of REF_CTR in ascending order. If PARITY is odd, transpose neighbors 1 and 2. (H-isotopes and phantom substituents are treated in the same way as Procedure A.)

B3. For $i = 1, 2$, if atom REF_NBR_*i* is not in position *i*, then move it to position *i* by a transposition and also transpose neighbors 3 and 4.

B4. List the neighbors of REL_CTR such that REL_NBR_1 is first and REL_NBR_2 is second.

B5. If DESCRIPTOR is ALPHA, place REL_SUB in the same relative position (third or fourth) on the neighbor list as REF_SUB. Otherwise, place it in the opposite relative position.

B6. Place the remaining neighbor of REL_CTR in the vacant position.

B7. Count the number of transpositions needed to arrange the resulting neighbor list in ascending order and assign parity of even or odd to REL_CTR.

Procedure C. Transfer of Parity for Stereoparent Atoms. This procedure transfers a parity descriptor from a stereoparent template node to the matching node of a substance from the Registry File.

Arguments:

T_NODE	template node
T_ORIENT	substituent orientation (α/β) on template node
T_PARITY	template node parity
S_NODE	substance node matching T_NODE
S_DESC	stereodescriptor for S_NODE
ATOM_MAP	template to substance atom map

Note 1. On entry, ATOM_MAP is the atom mapping of the neighbors of T_NODE to the neighbors of S_NODE determined by substructure search. However, S_NODE may have one or two neighbors that do not appear on ATOM_MAP. In order to translate stereochemistry correctly in steps C3–C8, these unmatched neighbors must be paired with the hydrogen neighbors of T_NODE. Recall that if T_NODE is a ring atom with two hydrogen neighbors, the β-hydrogen is distinguished and denoted as D.

Note 2. For template stereocenters with unspecified stereochemistry—such as atom 5 of cholestane—the parity is calculated for a β-hydrogen. However, if S_NODE is a member of a *sec*-ring, the descriptor refers to the substituent with the highest CIP rank. For example, the descriptor 5A in Figure 8 refers to the aldehyde substituent at atom 5, not the hydrogen. Thus, we may have to adjust T_ORIENT so that it refers to the S_NODE substituent that is paired with the T_NODE hydrogen.

C1. Complete the atom map; see Note 1.

If there is an unmatched neighbor S_NBR attached to S_NODE such that S_NBR is a carbon and the bond to S_NODE is a ring bond, then

Pair S_NBR with the H atom on T_NODE.

If T_NODE has two hydrogen neighbors, then pair the remaining unmatched neighbor of S_NODE with D.

Else, if there are two unmatched neighbors attached to S_NODE, then order them according to their CIP ranks and pair them with the D and H neighbors of T_NODE, respectively.

Else, pair S_NBR with the D neighbor of T_NODE.

C2. Processing for *sec*-rings; see Note 2.

If T_NODE is an unspecified stereocenter AND T_NBR is a member of a *sec*-ring then

Let S_NBR1 be the neighbor of S_NODE paired with T_NBR and S_NBR2 the neighbor paired with the hydrogen.

If RANKS(S_NBR1) > RANK(S_NBR2) then T_ORIENT = α.

C3. Construct T_NBRS by arranging the neighbors around the T_NODE in ascending order.

C4. If T_PARITY is odd interchange any two nodes on T_NBRS.

C5. Construct S_NBRS by replacing each node on T_NBRS by its image on ATOM_MAP.

C6. Calculate T, the total number of transpositions needed to arrange S_NBRS in ascending order.

C7. If T_ORIENT $\neq$ S_DESC, $T = T + 1$.

C8. If T (mod 2) = 0 then parity = even.
Else parity = odd.

REFERENCES AND NOTES

(1) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. Chemical Abstracts Service Chemical Registry System. 1. General Design. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111–21.

(2) Freeland, R. G.; Funk, S. A.; O'Korn, L. J.; Wilson, G. A. Chemical Abstracts Service Chemical Registry System. 2. Augmented Connectivity Molecular Formula. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 94–8.

(3) Blackwood, J. E.; Elliott, P. M.; Stobaugh, R. E.; Watson, C. E. Chemical Abstracts Service Chemical Registry System. 3. Stereochemistry. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 3–8.

(4) Vander Stouw, G. G.; Gustafson, C.; Rule, J. D.; Watson, C. E. Chemical Abstracts Service Chemical Registry System. 4. Use of the Registry System To Support the Preparation of Index Nomenclature. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 213–8.

(5) Zamora, A.; Dayton, D. L. Chemical Abstracts Service Chemical Registry System. 5. Structure Input and Editing. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 219–22.

(6) Stobaugh, R. E. Chemical Abstracts Service Chemical Registry System. 6. Substance-Related Statistics. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 76–82.

(7) Mockus, J.; Stobaugh, R. E. Chemical Abstracts Service Chemical Registry System. 7. Tautomerism and Alternating Bonds. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 18–22.

(8) Moosemiller, J. P.; Ryan, A. W.; Stobaugh, R. E. Chemical Abstracts Service Chemistry Registry System. 8. Manual Registration. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 83–8.

(9) Ryan, A. W.; Stobaugh, R. E. Chemical Abstracts Service Chemical Registry System. 9. Input Structure Conventions. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 22–8.

(10) Hamill, K. A.; Nelson, R. D.; Vander Stouw, G. G.; Stobaugh, R. E. Chemical Abstracts Service Chemical Registry System. 10. Registration of Substances from Pre-1965 Indexes of *Chemical Abstracts*. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 175–9.

(11) Stobaugh, R. E. Chemical Abstracts Service Chemical Registry System. 11. Substance-Related Statistics: Update and Additions. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 180–7.

(12) Blackwood, J. E.; Giles, P. M., Jr. Chemical Abstracts Stereochemical Nomenclature of Organic Substances in the Ninth Collective period (1972–1976). *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 67–72.

(13) *Chemical Abstracts Index Guide*, Appendix IV, 1989, pp 180I–99I.

(14) Petrarca, A. E.; Lynch, M. F.; Rush, J. E. A method for generating unique computer structural representations of stereoisomers. *J. Chem. Doc.* **1967**, *7*, 154–65.

(15) Wipke, W. T.; Dyott, T. M. Stereochemically unique naming algorithm. *J. Am. Chem. Soc.* **1974**, *96*, 4834–42.

(16) Cahn, R. S.; Ingold, C.; Prelog, V. *Angew. Chem., Int. Ed.* **1966**, *5*, 385–551.

(17) Prelog, V.; Helmchen, H. Basic principles of the CIP-system and proposals for a revision. *Angew. Chem., Int. Ed.* **1982**, *21*, 567–83.

(18) Petrarca, A. E.; Rush, J. E. Computer generation of unique configurational descriptors for stereoisomeric square-planar and octahedral complexes. *J. Chem. Doc.* **1969**, *9*, 32–7.

(19) Choplin, F.; Marc, R.; Kaufmann, G.; Wipke, W. T. Computer design of synthesis in phosphorus chemistry: automatic treatment of stereochemistry. *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 110–8.

(20) Ref 13, pp 184I–187I.

(21) Even though they are similar in appearance, the sequences listed in a site permutation table should not be confused with the sequences describing a configuration as they are quite different. The sequences listed in a site permutation table are *not* configuration descriptors, rather each is a symmetry operation which, when applied to a configuration descriptor, generates an equivalent configuration descriptor.

(22) Detailed restrictions on the use of the various types of relative descriptors are given in Ref 13, pp 181I–183I.

(23) Vander Stouw, G. G.; Elliott, P. M.; Isenberg, A. C. Automated conversion of chemical substance names to atom-bond connection tables. *J. Chem. Doc.* **1974**, *14*, 185–93.

(24) The reference center is the lowest numbered stereogenic position in the parent ring. When chiral elements are present in two ring systems, the parent is the system containing the greater number of stereogenic centers.

(25) Wipke, W. T.; Dyott, T. M. Simulation and evaluation of chemical synthesis. Computer representation and manipulation of stereochemistry. *J. Am. Chem. Soc.* **1974**, *96*, 4825–34.

(26) Occasionally, structure diagrams show a stereobond to a less preferred substituent. In such cases, the view information is adjusted to reflect the preferred substituent. Compare the entry for atom 17 in Table III with Figure 6.

(27) Brown, M. F.; Cook, B. R.; Sloan, T. E. Stereochemical notation in coordination chemistry. Mononuclear complexes. *Inorg. Chem.* **1975**, *14*, 1273-8.
(28) *C&E News*, July 9, **1990**, 9.
(29) Lipkus, A. H.; Blower, Jr., P. E. Stereochemical substructure searching: Handling of relative configurations. *Abstracts of Papers*, The Second International Meeting on Chemical Structures, Noordwijkerhout, The Netherlands, June 3-7, 1990.

Classification of Chemical Reactions: Potential, Possibilities, and Continuing Relevance†

DAVID BAWDEN
Department of Information Science, The City University, Northampton Square, London, EC1V 0HB England

Received March 18, 1991

An overview is given of the nature of, and methods for, classification of chemical reactions. Particular emphasis is placed on classifications based on formal structural change and on reaction mechanism, and on the role of classifications within computerized reaction retrieval systems. It is argued that classification is of great and continuing value in information retrieval, information discovery, and education.

INTRODUCTION

"The problem of classifying chemical reactions is intrinsically a complex one. None of the methods used to date has satisfied all the requirements of chemists. The reason for this is that there are so many attributes of a chemical reaction in which chemists might be interested." M. F. Lynch[1]

"Many proposals exist for the classification of chemical reactions. Some classifications have been devised simply to provide an author with a convenient framework for some limited purpose; others have been intended to be more rigorous and inclusive. Most have features in common, but the different schemes almost never agree when examined in detail: in short, there exists no consensus about the proper classification of reactions." D. P. N. Satchell[2]

In this paper, we shall consider the nature and meaning of the classification of chemical reactions, its value, and ways in which it may be achieved. We shall also consider how the utility and applicability of reaction classification can be extended in the future, with particular reference to computerized reaction databases. The choice of references is deliberately selective rather than comprehensive, with the aim of covering, and setting into context, all major strands in research into and operational use of reaction classification.

By "classification of reactions", I mean schemes for displaying and understanding the variety of chemical reactions and their interrelations. It is worthwhile at the outset stating what reaction classification is *not*.

First, it is not simply a scheme of nomenclature, notation, coding, or keywording for reaction description, although any classification must necessarily have its own notation. Reaction nomenclature has lagged considerably behind that for chemical structure, although some proposals for a systematic nomenclature, allowing access to *types* of reaction have been made.[3,4] Nomenclatures and similar descriptors, however, do not show the interrelation between reactions, which we shall see to be the essence of any reaction classification.

Second, it is not a computer-searchable form of reaction description. The computerized reaction retrieval procedures pioneered by Vladutz, Lynch, and Willett[1,5-9] and incorporated into operational reaction retrieval systems such as ORAC and REACCS are not in themselves classificatory, although, as we shall see, they may be used to formulate classifications. Nor are the similarity searching routines being incorporated into reaction systems classifications in themselves, though they also may be a tool for classification building.

† This paper is dedicated to Professor Michael Lynch, who first stimulated my interest in reaction retrieval.

VALUE OF REACTION CLASSIFICATION

There are a number of clearly understood benefits of classification, within information systems in general. They are involved in four distinct, though interrelated, processes. The first is the "straightforward" retrieval of information. The second is the less-well-understood process of information discovery, through analysis, correlation, and reasoning by analogy. The third is the teaching and exposition of the variety and scope of chemical reactions. The fourth is the systematization of chemical reaction information for use by other algorithmic systems.

For retrieval, classification confers four particular advantages. First, it can make retrieval from computerized databases more efficient. Second, it makes retrieval simpler and more accessible by encouraging a browsing approach. Third, it gives access to information at precisely the required level of specificity or generality and particularly enables easy access to "generic" types of information. Fourth, it makes it much more convenient to select subsets of related information for subsequent detailed search or analysis.

For information discovery and analysis, the creative and innovative use of the information resource classification allows the identification of new types of entity and the estimation of the qualities and properties of such entities. It is a powerful source of analogical reasoning and a tool for the understanding of entity relationships. In chemistry, the best known example is Mendeleev's use of the periodic table in classifying the elements and making rationalizations and predictions in analogical reasoning.

For chemical reactions, classification may be seen both as an aid to retrieval and as a discovery and analysis tool. In the former application, a classification will be a complement to structure-based retrieval techniques, to mechanistic nomenclatures, etc. In the latter, it may be used, for example, to identify new types of reactions, whose feasibility may then be investigated in the laboratory, or to show the presence of similar reaction mechanisms in apparently diverse structural environments. As Arens puts it:[10] "generalization can uncover the formal similarity of a wide range of reactions and can lead to their classification and to a strategy for the design of new ones".

For teaching and exposition, the value of classifications as an aid to understanding is well-understood in many subject areas. For chemistry, where much of the skill of a teacher lies in showing the order and interrelation hidden by a mass of facts, classificatory principles are invariably used, explicitly or implicitly, in dealing with concepts of chemical structure. This is less so in the reaction domain, and a major reason for

0095-2338/91/1631-0212$02.50/0