The Chemical Abstracts Service Chemical Registry System. I. General Design†

P. G. DITTMAR, R. E. STOBAUGH,* and C. E. WATSON

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received December 5, 1975

The Chemical Abstracts Service (CAS) Chemical Registry System is a computer-based system that uniquely identifies chemical substances on the basis of composition and structure. Since initial operation in 1964 as a stand-alone input, storage, and retrieval system for structural representations of organic chemical compounds, the scope of the CAS Registry System has steadily increased to include all types of chemical substances and the entire system has been integrated into CAS indexing operations. The third refinement of this system, Registry III, which has been in operation for over a year, involves major changes in Registry records but no change in the basic algorithmic techniques for registering chemical substances. The previous format for listing atoms and bonds has been modified so that each ring system is now separately identified, and this ring-system identifier is used in the record for each substance that contains that ring. These modifications support CAS nomenclature derivation and also a computer-based structure output system. The general design of Registry III, which involves a structure record of cyclic and acyclic segments, is presented.

INTRODUCTION

Chemical Abstracts Service (CAS) in its mission of covering the world's primary chemical and chemical engineering literature has long been heavily involved with handling of information about chemical substances. A particular problem has been the identification of reoccurrences of substances in scientific literature. Such identification is necessary to assure consistency in the indexes to *Chemical Abstracts.* These indexes are based on a systematic chemical nomenclature, and the technique used for many years to assure consistency was to name each indexed substance every time it was cited and to file the name alphabetically in an index with its citations. This resulted in a large amount of redundant effort in repeated renamings. To remedy this situation, the CAS Chemical Registry System was developed in the early 1960's. The main objective was the establishment of a computer-based system to identify chemical substances uniquely on the basis of molecular structure. It was also realized that the CAS Registry Number would provide a machine address that would allow linking of files containing structural information, nomenclature, indexing, and bibliographic information. Moreover, this linking function was seen as a potential general purpose linkage based on substance identification that could eventually involve external files and data banks, primary literature, and other secondary services.

The initial and experimental computer-based system, referred to as Registry I, established the viability and validity of the registration concept for fully defined organic chemical substances. Registry I began operating late in 1964. In 1968, the scope of the system was increased as additional classes of substances were handled, and this system, referred to as Registry II, began to be integrated into the CAS indexing operation. The most recent version of the Registry System, Registry III, was placed in operation in 1974 and embodies major adjustments in the Registry records to make the system more effective in its support of the CAS index nomenclature function and the computer-based structure output operation through explicit identification of specific ring systems. As its use has been expanded, Registry has proved itself to be eminently reliable and consistent as a substance identification tool. It has become an essential CAS production tool supporting CA index input and final product compilation. The CAS Chemical Registry has also found widespread interest and support from the scientific and technical community.

Table I. CAS Chemical Registry Coverage (as of November 1975)

Type of substance	No. of entries
Fully defined substances	2,950,000
Substances with partly defined structures	48,800
Polymers	95,500
Coordination compounds	201,400
Alloys	46,700
Mixtures	7,000

The foundation of the system is an algorithm that generates a unique and unambiguous computer-language description of the molecular structure of a substance. This includes not only a description of atoms and bonds but also their arrangement in space, i.e., the stereochemistry of the substance.

When a new, unique structural description is added to the Registry Master File, the substance it represents is automatically assigned a serial number, or machine address, called a CAS Registry Number. The CAS Registry Number, which itself has no chemical significance, provides a brief and unique means of substance identification. It designates only one substance, and thus provides a means of linking often unrecognized synonymous names with the description of molecular structure. Whenever a substance is encountered again, the originally assigned Registry Number is retrieved automatically from the Registry System. Because its assignment and use are independent of any system of chemical nomenclature, the CAS Registry Number assures continuity of substance identification.

At the heart of the Registry System is the Registry Master Structure File which presently contains records of over three million unique substances. Records of new structures are being added at the rate of 300,000–350,000 per year. All chemical substances indexed in *Chemical Abstracts* since the Registry System began operation, as well as a large number of substances derived from a variety of reference works, are included in the Registry Master File.

An average substance in the Registry System contains 43 atoms, 22 of which are nonhydrogen atoms. It contains about 1.5 ring systems with an average of eight atoms per ring system. (For the purposes of the Registry System, "ring system" means any contiguous cyclic arrangement, such as cyclopentane, naphthalene, quinoline, etc.)

The CAS Chemical Registry System is presently capable of generating machine structure records for a variety of chemical substances including fully defined organic and inorganic substances, ions, elements, free radicals, polymers,

† Presented to the Division of Chemical Literature, 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 8, 1975.

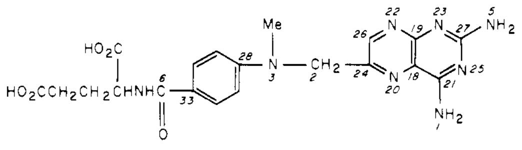
* To whom correspondence should be directed.

coordination compounds, alloys, and mixtures. It also is capable of handling certain partially defined structures—generally those types for which the location of certain chemical groups, positions of unsaturation, or the site of an esterification has not been specified. Table I provides Registry statistics for some of these classes of substances.

REGISTRY III SUBSTANCE REPRESENTATION

A chemical substance which is to be recorded by machine processing in the structure-based portion of the Chemical Registry System must be capable of being described in terms of a set of topological characteristics. This means that substances to be registered based upon structural data are ones for which the known data include a molecular formula, the basic relationship of the atoms present, and sufficient other data to allow complete and unambiguous identification of the substance. Each substance registered is represented in the system by a Unique Chemical Registry Record (UCRR). This machine record consists of four components: (a) the connection table topology, which is a detailed inventory of the atoms and bonds that comprise the basic structure of the substance; (b) a text descriptor component which defines the known stereochemical characteristics of the molecule; (c) an isotopic labeling component which identifies any labeled atoms in the structure; and (d) a derivative component, which indicates that the registered substance is a salt or complex of the substance defined by the first three components of the record.

Topology. The topological component of the UCRR consists of a nonredundant, lexicographically ordered connection table. This connection table is derived by computer program from the input table, which is redundant, nonunique, and unambiguous. The unique connection table describes the acyclic nodes and the interconnections or bonds between the acyclic nodes. Cyclic or ring nodes are defined in terms of the Ring Identifier(s) for the ring system(s) in which the ring nodes are contained. Acyclic connections between ring systems or between a ring system and an acyclic node are also defined. The following example in Figure 1 graphically illustrates the topological segment in which an identifying number is substituted for the portion of the structure that corresponds to a defined ring system. This identifying number, which is a composite expression functioning as a machine address, will be described later in more detail. The structure (Figure 1), which is shown with the principal nodes numbered to indicate the sequence in which they would be listed in the canonical form of its connection table, contains pteridine and benzene rings. When the pteridine ring is replaced by its unique identifier, 591U.385*.57P, and the benzene ring by its identifier, 46T.150A.182, the structure can be defined by the representation in Figure 2, where the numerals over the bonds specify the nodes in the ring system at which the acyclic groups are attached. The adjusted unique connection table then consists of (1) a listing of the ring identifiers, (2) the connection table representations of the acyclic atoms and bonds, (3) the connections between the ring and acyclic components of the structure, and (4) a cross reference for the ring nodes between the ring numbering and the substance numbering (see Figure 3). The corresponding ring systems are stored in the file of ring systems in the form of a unique connection table for each ring (see Figure 4). Any references such as abnormal masses or charges to ring nodes are made using the node numbers of the full structure.



Registry Number 59-05-2

Figure 1.

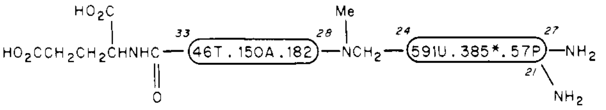


Figure 2.

```
(1) 591U . 385*. 57P
    46T.150A .182            CO2H
     1     2  3   5      6    |
(2) NH2 . CH2N . NH2 . CNHCHCH2CH2CO2H
              |          ||
              Me         O

(3) 1-21, 2-24, 3-28, 5-27, 6-33
```

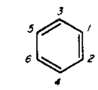
(4) 591U.385*.57P

1	2	3	4	5	6	7	8	9	10
18	19	21	20	22	23	25	24	26	27

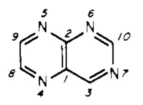
46T.150A.182

1	2	3	4	5	6
28	29	30	31	32	33

Figure 3.



Ring Identifier 46T.150A.182



Ring Identifier 591U.385*.57P

Figure 4.

Text Descriptor. The text descriptor component of the UCRR may be used to describe the stereochemistry for a given topology, or it may be used to verbalize some special nonstereochemical characteristic of the substance which cannot presently be adequately described by other methods.

Stereochemical descriptors are selected from one or more of several classes of descriptors, depending on the type of substance and stereochemistry involved. Seven classes of stereochemical descriptors are presently defined as follows:

Absolute configurational descriptors
Relative configurational descriptors
Optical rotation descriptors
Stereo parent descriptors
Carbohydrate/amino acid descriptors
Trivial name (partially specific) descriptors
Coordination descriptors

Each of these descriptor types has a controlled vocabulary which permits rigorous editing of the descriptor both in itself and as it relates to other descriptors for the same topology. Examples of these descriptors are presented in Appendix A and a comprehensive coverage of this topic will be presented in a future publication.[1]

Nonstereochemical descriptors are used thus far for five classes of substances. They distinguish between different substances which otherwise have the same topological representation e.g., incompletely described substances, radical ions, minerals, mixtures, and alloys. They will be extended to other classes should it be found necessary. The use of these descriptors is illustrated in the discussion on specialized classes of substances in Appendix B.

Isotopic Labeling. The isotopic labeling component of the UCRR is used to describe three basic types of labeling:

Specific labeling—this condition exists when a specific atom is labeled with a specific mass value (both the atom number and the mass value are recorded).

Hydrogen labeling—this condition exists when one or more hydrogen isotopes are attached to a specific atom(s). Cited are the mass value, the number of such atoms, and the atom number to which they are attached.

Nonspecific labeling—this condition exists when it is known that some atom(s) in the structure is(are) labeled but the specific location is unknown. In this case, the identifying data consist of the element symbol, a mass value, and a coefficient to express how many such atoms exist.

Derivatives. The derivative component of the UCRR is used to describe a simple salt or complex of the substance defined by the first three components of the record. In addition, this portion of the structure record contains several specific items, as described:

Charge—this item of data consists of the numerical value of the charge and the sign, + or −, associated with it. The sign and value are then linked to the appropriate atom in the structure through the atom number.

Delocalized charge—this item is similar to the charge above except that the charge value and sign are associated with more than one atom. Therefore, the data consist of the sign and a series of atom numbers (three or more) over which the charge is distributed.

Valence—this item is similar but not identical with the normal chemical valence. The valence is a value calculated from the environment of an atom. It is the sum of the bond lines, the value of the charge, and the number of attached hydrogen atoms. The valence value is cited only when it differs from a value that is normally assumed for that element.

Tautomer mobile group—the tautomer identification routines that have been defined operate on the principle that certain structural characteristics may be input in various ways without changing the identity of the substance. These characteristics may be one or more hydrogen atoms, negative charges, or, in special cases, a positive charge which may migrate or appear at either end of a sequence of atoms. The tautomer mobile group identifies the type of migrating function (hydrogen, negative charge, positive charge), the number of such functions, and the atom numbers at which these functions might be localized. More details on this aspect of the record are described later.

Single atom fragment—this item is the record of the occurrence in a structure of a separate (termed "disconnected") single nonhydrogen atom (with or without attached hydrogen) or a hydrogen ion, which is handled by relating that occurrence directly to the substance topology. Such disconnected structural representations result from the structure conventions for the following: (1) simple metal salts resulting from metal replacement of hydrogen in the acidic function of acids, in the hydroxyl group of alcohols and phenols (and thio, seleno, and telluro analogs), and in the amino group of amines and amides; (2) hydrogen acid salts of organic nitrogen bases; (3) quaternary ammonium, sulfonium, and diazonium halides; (4) various compounds or complexes with H_2O, NH_3, H_2S, etc. The data recorded for such atoms consist of: element symbol, valence (if other than standard), charge, number of attached hydrogens, and a coefficient indicating the ratio of the disconnected atom to the main topological fragment.

Bookkeeping. A fifth component of the UCRR is a group of data recorded for bookkeeping rather than substance identification purposes. Among these items are the following:

Registry Number—this unit is the file identifier of the fragment work unit.

Molecular formula—this datum represents the Hill-ordered molecular formula of the atoms.

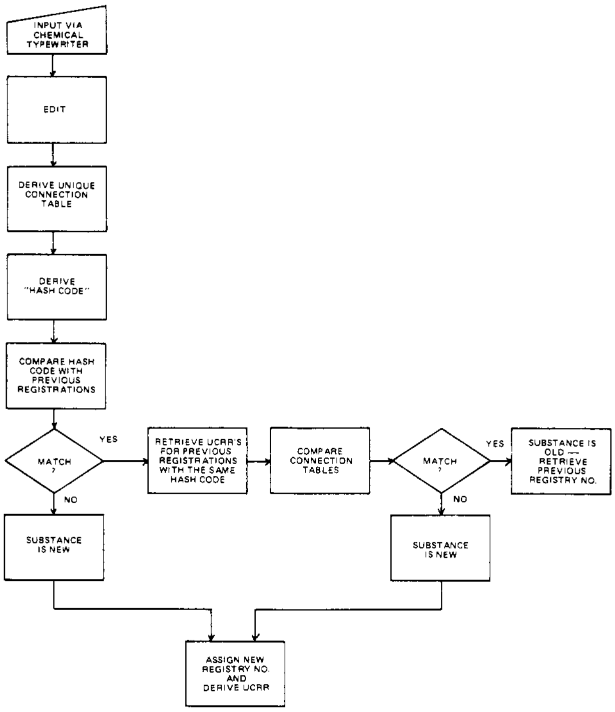

INPUT VIA CHEMICAL TYPEWRITER
EDIT
DERIVE UNIQUE CONNECTION TABLE
DERIVE "HASH CODE"
COMPARE HASH CODE WITH PREVIOUS REGISTRATIONS
MATCH ?
YES
NO
SUBSTANCE IS NEW
RETRIEVE UCRR'S FOR PREVIOUS REGISTRATIONS WITH THE SAME HASH CODE
COMPARE CONNECTION TABLES
MATCH ?
YES
SUBSTANCE IS OLD — RETRIEVE PREVIOUS REGISTRY NO.
NO
SUBSTANCE IS NEW
ASSIGN NEW REGISTRY NO. AND DERIVE UCRR

Figure 5. The registration process.

Transaction date—this is a date which records the activity against a Registry Number work unit.

THE REGISTRATION PROCESS

The registration process consists of several steps as illustrated in Figure 5, beginning with the input of a chemical substance record, going through a programmed edit of the input data, derivation of a unique connection table, and, finally, data base matching operations.

Input. Substances are input for registration using a chemical typewriter or cathode ray tube and light pen to generate the image of the substance to be registered. The data input consists of all the items necessary for complete description of the given substance—the atoms and bonds present, any ratio, charge, or mass data, and the appropriate text descriptor.

Editing. The programmed editing is such that performance of the various individual checking procedures allows the data that pass the edits to be certified as valid. Human inspection and possible correction are required only in the cases where an error or a possible error is identified. Following the edit of the input data, various manipulations of the data must be performed to prepare the data for further processing. Such manipulations include the identification of the rings within the input data, identification of alternating bonds, identification of tautomeric conditions, and the separation and tagging of components in the input record which require special processing.

The edits performed on the data input consist of the following classes:

1. Edits for control information
   a. Keyboarding batch control
   b. Temporary identification number
   c. Chemical structure class type
   d. Input device
   e. Transaction type
   f. Input document source
2. Edits for chemical content
   a. Molecular formula format and symbols
   b. Valid element symbols in structural description
   c. Valid bond values

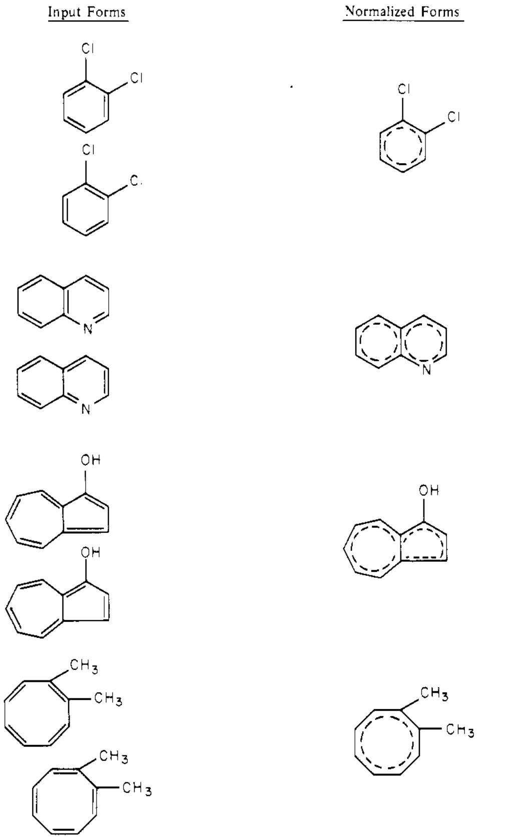


Figure 6. Examples of normalization of alternating cyclic single and double bonds.

d. Valence
e. Oxidation state and Stock number
f. Mass indication
g. Charges
h. Coordination substance structuring conventions
i. Agreement between molecular formula and structure input
j. Multipliers and other coefficients
k. Proper use of groups, fragments, and expressions
l. Expansions of shortcut symbols
m. Consecutive multiple bonds
n. Valid text descriptors

Validity tables, used in various edits, are listed in Appendix C. More comprehensive coverage of the specific edits will be presented in future publications.

Derivation of the Unique Connection Table. Following the editing of the input structure, two processes must be performed prior to the derivation of the Unique Connection Table. These consist of, first, an algorithmic identification of all cyclic pathways within the input structure, and, second, a "normalization" of bonds within the input substance to account for varying forms of input for the same substance (see Figure 6). The normalization routines detect two types of variation in substance input as follows:

1. Variation in location of alternating cyclic single and double bonds. The technique used has remained basically unchanged in the development of registration processes. A search is made for the presence of an even-numbered, completely cyclic path in which each node is attached by one single and one double bond to its immediate neighbors in the path and in which no nodes are cited more than once. Identification of such a path results in the conversion of all the bonds along the path from either single or double to a different type, termed "alternating". All cyclic paths within a substance are exhaustively analyzed for this condition. The procedure is iterative in that, once converted, a bond is used as either a single or double bond when tracing another path.

2. Variations resulting from migrating functions. Techniques have been developed to handle the phenomenon of tautomerism, by which a substance reacts as if it were comprised of more than one structure. Tautomeric conditions as defined below are identified in the input structural record, and the bonds involved, input as single and double, are converted into a different type, or *tautomeric*. The specific type of tautomerism handled is prototropy, in which hydrogen is the mobile entity. The same technique is also used in situations which are not tautomeric, but which bear a formal analogy, i.e., when the mobile entity is a charge instead of the hydrogen atom.

In Registry III, the generalized tautomer expression is as follows:

$$\mathrm{M{=}Q{-}({-}M'{=}Q'{-})_{n}ZH(D,T,^{-})} \rightleftarrows \mathrm{(^{-},T,D)HM{-}Q{=}({=}M'{-}Q'{=})_{n}Z}$$

where M (and M') and Z are any combinations of N (trivalent), O, S, Se, and Te (the last four bivalent); and Q (and Q') may be C, N, P, As, Sb, S, Se, Te, Cl, Br, or I. As is indicated, the entity may be H, the isotopes D and T, or the negative charge, and this item may be used repeatedly in concatenated three-atom sequences. Any combination of acyclic and cyclic bonds is allowed.

In addition, another type of tautomeric condition is defined as follows:

$$\mathrm{R}\overbrace{\mathrm{N{-}Q{=}N^{+}}}^{\mathrm{X}}\mathrm{R'} \rightleftarrows \mathrm{R^{+}}\overbrace{\mathrm{N{=}Q{-}N}}^{\mathrm{X}}\mathrm{R'}$$

where Q = C or N, X represents a cyclic situation, and R and R' are acyclic, nonhydrogen attachments. Figure 7 provides illustrations of tautomer normalization.

An "override" of the tautomer normalization, which may be indicated at input time, has been provided for the rare instances when one specific tautomeric form is desired for machine storage.

Data Base Matching. In order to register a substance it is necessary to derive the new form of the Registry structure record. The system derives a modified form from the unique connection table used in earlier versions of the Registry System.[2] This form is modified so that all acyclic nodes will be brought together in the connection table. This preliminary form of the connection table is then used to match against the master file of previously encountered substances.

The matching procedure is accomplished by first deriving a hash code from the connection table.[3] All previously encountered substances with the same hash code (usually only one, very rarely two or three) are then retrieved and compared with the input substance one at a time until either an exact match is found or all of the substances with this hash code have been tried. If an exact match is found, the Registry Number of the previously encountered substance is retrieved. Otherwise, the substance is new, in which case a new Registry Number is assigned and the ring systems are extracted from the input substance and identified.

In the comparison of the input substance to the file substance, the connection table for the file substance is retrieved, any ring system records for the file substance are retrieved, and the full form of the connection table is reconstituted. The matching then proceeds by comparing the topology segments. If a match is found on the basis of this comparison, the text descriptor segments are compared. If a match is found on the basis of the text descriptor segments, the labeling segments are compared. If a match is again found, then the derivative segments are compared. If a match is found at the derivative segments, the substance is old and the Registry Number of the file substance is assigned to the input substance. If no match is found in any one of the above comparisons, the input

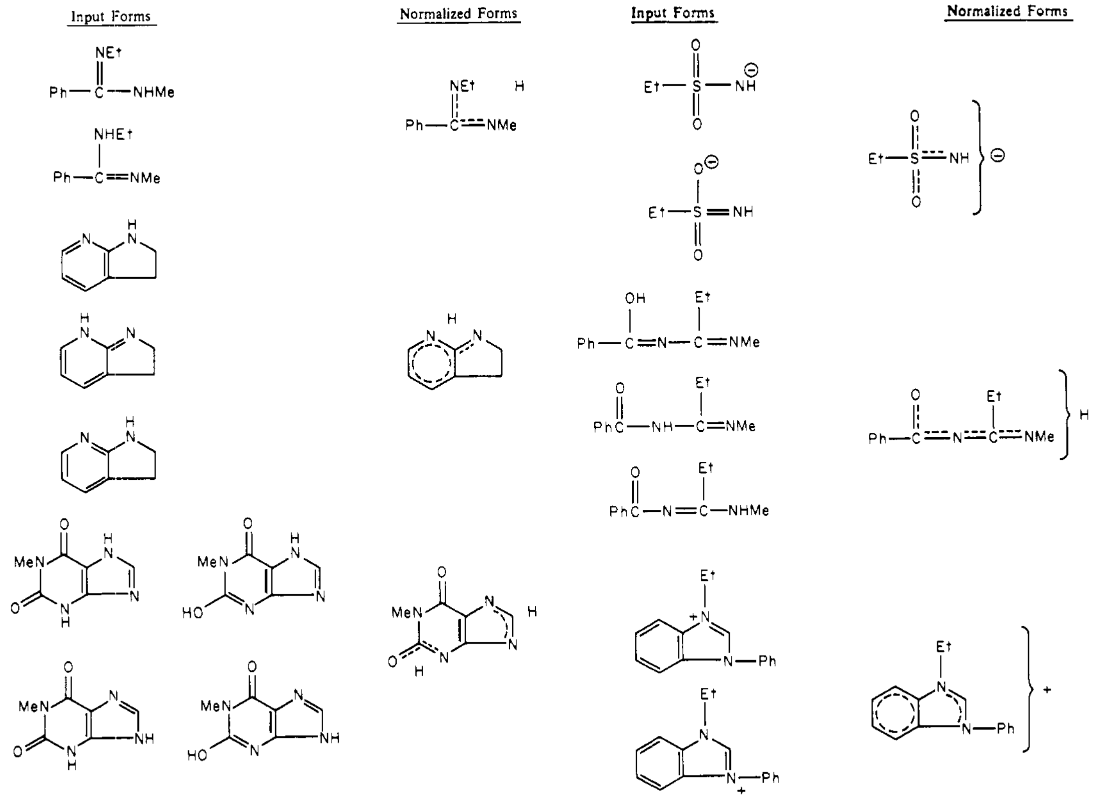
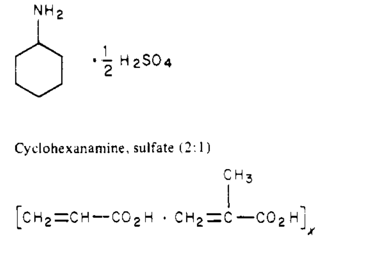


Figure 7. Examples of tautomer nomalization.

substance does not match the file substance. The next file substance with the same hash code is retrieved and the above procedure is repeated.

Some types of chemical substances are represented in a "disconnected" format consisting of two or more separate components. Two examples of such substances are: salts of acids or bases, in which the components are all "multi-atom fragments", i.e., multiple atoms with or without attached hydrogen; and polymers expressed in terms of their monomers. The examples of Figure 8 illustrate the "component" treatment. In Registry III, each component is separately and uniquely identified by a Registry Number. The total substance is then given a unique Registry Number. The UCRR for such a substance is simply an identification by Registry Number of the components and an indication of their relationship to one another within the substance. Graphically, this is represented as shown below.


Cyclohexanamine, sulfate (2:1)

Propenoic acid, polymer with 2-methylpropenoic acid

Figure 8.

Input substance	Component substances	Input substance registration
A·*n*B	A B	Reg No. A·*n*Reg No. B

RING SYSTEM IDENTIFICATION

During the derivation of the UCRR, the registration process in Registry III uniquely identifies ring systems contained within substances being registered. In so doing, support is provided to the subsequent naming of substances and graphic display of substances.

The method used physically partitions the substance into acyclic and ring system groups and stores the ring system groups in a ring file. The structure record for the substance contains the ring identifier(s) for the ring system(s). The procedure of establishing the identity of a ring system is similar to the procedure for substances.

The identifier assigned to each ring system, the Ring ID, is a composite expression. Each unique ring graph is assigned an identifier, each node variation of a ring graph is assigned an identifier, and each bond variation for a graph/node combination is assigned an identifier. Thus the Ring ID can be represented as follows:

Ring ID = Ring Graph ID/Node Variant ID/Bond Variant ID

The structure record for a ring system is graph/node/bond specific and is composed of the graph, nodes, and bonds with the fully defined Ring ID.

Ring system identification is, then, the process of comparing the representation of the ring systems of those incoming substances for which the topology segment is new to the representations of the ring systems previously encountered. This is done to establish whether the ring system is a new ring graph and hence contains new node variation and bond variation or is identical at the graph level, graph/node level, or graph/node/bond level with a previously encountered ring system.

SUMMARY

The registration techniques described herein have substantially met the basic objective of CAS substance handling, that of chemical substance identification. The Chemical Registry System has become an integral part of the index processing at CAS. Some 25,000 substance identifications are handled each week, providing rigorous controls for the systematic nomenclature published in the CA chemical substance indexes. Only for newly registered substances are systematic names generated; names for all others are retrieved automatically.

The CAS Registry Number has become an invariant identification tag for each chemical substance that, to an increasing extent, links the substance with its citations in the secondary and primary literature and with references in specialized file collections. Registry Numbers are included in the substance entries in CA volume and collective indexes and in certain CA sections. They appear in a growing number of primary journals, handbooks, and files, such as *The Journal of Organic Chemistry*, *Inorganic Chemistry*, *Angewandte Chemie*, "U.S. Adopted Names", the "U.S. Pharmacopeia", and the "National Formulary".

Registry III has expanded the system's capabilities to the extent that machine support is now provided for manual generation of CA index names for new substances. The Registry III feature of separate identification of ring systems has made possible machine generation of structure diagrams from the structural record.[4,5] This same feature provides the basis for the present development effort leading to machine generation of most of the CA systematic index names in 1976–77.[6]

ACKNOWLEDGMENT

The development of the CAS Chemical Registry System was substantially supported by the National Science Foundation (Contract C656). Chemical Abstracts Service, a division of the American Chemical Society, gratefully acknowledges this support.

APPENDIX A: STEREOCHEMICAL DESCRIPTORS

1. Absolute Configurational Descriptors. These descriptors indicate the spatial disposition of groups about a dissymmetric portion of a structure. This indication is made by comparing a preferred orientation of the structure with an external standard. The orientation used is according to the Cahn–Ingold–Prelog Sequence Rule.[7]

Valid Terms: R S

2. Relative Configurational Descriptors. Relative descriptors indicate the stereochemical interrelationships of groups of a compound. When it is necessary to select one of two groups in order to assign a descriptor, the Sequence Rule[7] preferred group is selected. The valid relative configurational terms are

CIS
TRANS
ENDO
EXO
SYN
ANTI
A (for α)
B (for β)
R*
S*
E
Z

3. Optical Rotation Descriptors. These descriptors indicate the sign of rotation of plane polarized light. The valid terms are

(+) (−) (±)

4. Stereoparent Descriptors. These descriptors are names or shortened versions of names which in themselves imply stereochemistry involving fused ring systems and two or more chiral centers. Some examples of the 129 valid terms (as of June 1975) are

(11) CYTOCHALASAN
ACONITANE
ANDROST
CEVANE
CHOL
CRINAN
DAMMARANE
GON
LANOST
MORPHINAN
OLEANANE
PANAMINE
PREGN
PROST
SOLANIDANE
THALMAN
URSANE
YOHIMBAN

5. Amino Acid and Carbohydrate Descriptors. The standard configurational prefixes for amino acids serve as descriptor terms, singly or in combination. The valid terms are

D L DL

The descriptors for carbohydrates are derived from the anomeric prefixes, configurational prefixes, and basic word roots. Various combinations are used. Valid terms are

A (for α)
B (for β)
D
L
DL
ALLO
XYLO
LYXO
GLUCO
ARABINO
MANNO
ALTRO
IDO
TALO
GALACTO
RIBO
GULO

6. Trivial Name Descriptors. Like stereoparent descriptors, trivial name descriptors are names which carry some stereochemical implications, but unlike the case of stereoparents, the stereochemistry is either incompletely known or difficult to systematize. Examples of valid terms are

CHETOCIN
CLERODENDRIN-A
EVONINE
13B-COBIN
VESCALAGIN
RETIN
JULICHROME-Q
MONENSIN

7. Coordination Compound Descriptors. For coordination numbers greater than 3, descriptors for nuclear stereochemistry are made up of four parts: (1) a system indicator for molecular geometry, (2) a configuration number, (3) a chirality symbol, and (4) a ligand stereochemical indicator. Examples of combinations are

OC-6-55-DELTA
OC-6-24-A
SP-4-2
TB-5-11
OC-6-12-(S), (S)

APPENDIX B: SPECIALIZED EXAMPLES OF SUBSTANCES

Mixtures. Machine registration of mixtures is essentially the registration of the components of the mixture as accomplished by input of the topology of each of the component(s) or, alternatively, by input of a predetermined Registry Number(s) in place of the topology for the component(s).

Registration is based upon the Registry Numbers of the components of the mixture as linked in expressions. Thus, it is necessary to register the components prior to the registration of the mixture.

Mixture input consists of the following items of data:

Molecular formula—a composite formula made up of the formulas of each component, kept as a separate moiety.

Components—topological representation or Registry Number.

Composition—a ratio of 1:1 is assumed between components; each component may be made up of fragments having ratios other than 1:1.

Text descriptor—the nonsteric descriptor MX is associated with the mixture itself; each component may have steric and/or nonsteric descriptors.

Labeling—isotopic labeling is associated with the mixture only if it cannot be associated with one of the components; thus, only nonspecific labeling is input at the mixture level.

Derivatives—derivative data are associated with the mixture components (see Figure 9).

Alloys. Alloys are considered a type of mixture, i.e., a mixture of metals or a metallic based substance. Unlike mixtures as defined for registration by CAS, alloys are often differentiated on the basis of a composition factor. Thus, the amount of a metal present in an alloy is significant.

Alloy input consists of the following items of data:

Molecular formula—a disconnected formula of the fragments.

Alloy components—a disconnected formula in which each component of the alloy is cited as an input fragment.

Terracoat (mixture of 5-ethoxy-3-(trichloromethyl)-1,2,4,thiadiazole with pentachloronitrobenzene)
Molecular formula $C_5H_5Cl_3N_2OS \cdot C_6Cl_5NO_2$
Text Descriptor MX

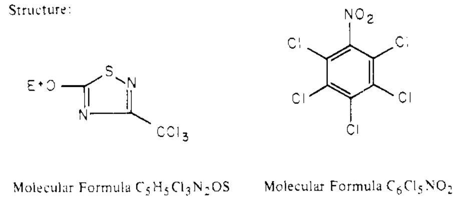


Figure 9.

Composition—an item of data associated with each fragment which states whether the fragment is the base metal or a component, and a numeric value showing the percentage composition by weight of the whole for each fragment. The percentage composition may be specific or may be given in ranges for some or all components.

Text descriptor—a nonsteric descriptor associated with the substance designating the substance as an alloy and containing an alloy code name or number if appropriate.

Labeling—no labeling exists for the alloy itself; however, components may be labeled.

Derivatives—a derivative type of data other than percentage composition is not associated with the alloy itself.

Alloy composition—the data when present consist of a numeric value for an individual component expressed in the form of three characters, two numerics, and a decimal point in which the decimal is either the second or third character, i.e., 10 (decimal implied); 1.0, 0.1; 12, 1.2, 0.1; 75, 7.6, 0.8; etc. (see Figure 10).

Example:

Alnico V

Molecular Formula Fe.Co.Ni.Al.Cu.Ti.Si

Text Descriptor AY.ALNICO V

Structure:		
	Fe	49
	Co	24
	Ni	15
	A	8
	Cu	3
	Ti	0.5
	Si	0.1

Figure 10.

Polymers. Polymers are expressed either in terms of a repeating unit without end groups $(\text{-A-B-C-})_n$ or with end groups $X(\text{-A-B-C-})_nY$ or in terms of a homopolymer or copolymer expression $(A)_X$ or $(A \cdot B)_X$. A wide number of variations may take place in such a representation. However, the fundamental concepts of component registration apply to polymers in such a way that special techniques are unnecessary except for the handling of repeating unit end groups. The following example serves as a guide in the handling of polymers expressed in terms of the monomers (see Figure 11). Input for this class of polymers is analogous to that for mixtures.

Isotactic methyl methacrylate-styrene polymer

Molecular formula $(C_5H_8O_2 \cdot C_8H_8)_X$

Text descriptor PM. ISOTACTIC

Structure:

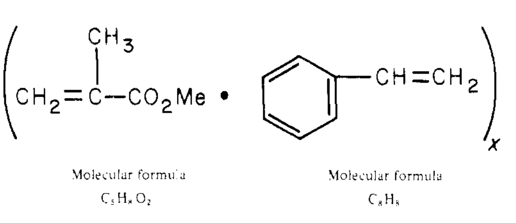


Figure 11.

Incompletely Described (ID) Substances. The topological segment of the structure record of these substances may not reflect the degree of specificity designated in a primary document. The nonsteric descriptor provides the additional specificity.

ID descriptors are divided into subclasses, examples of which are listed below with illustrative structures and descriptors. All the subclasses include the ID identification code. The compounds shown in Figure 12 are those for which the designations RING or CHAIN totally indicate the degree of specificity.

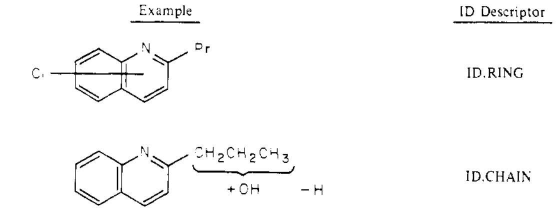


Figure 12.

The compounds shown in Figure 13 are those for which the chain or ring size and atom content must be indicated along with the designation CHAIN or RING. Atom content is expressed in CA preferred order (Hill) without hydrogens. Size and atom content is placed after the designator as shown.

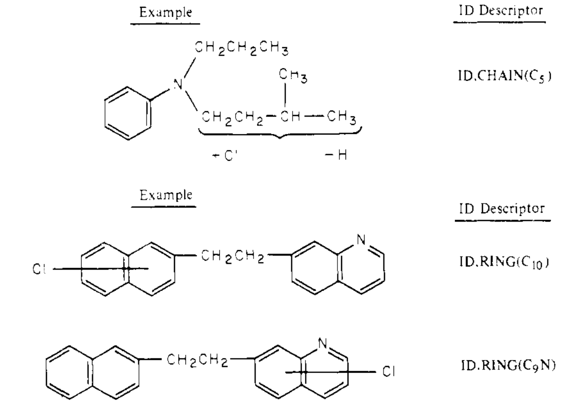


Figure 13.

The compounds listed in Figure 14 are those for which the use of branching prefixes, such as TERT, ISO, or SEC, totally indicate the degree of specificity given by an author and are to be treated as the examples indicate.

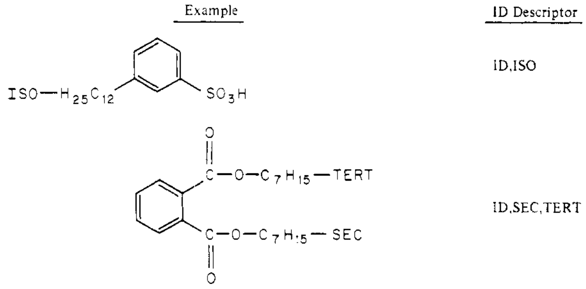
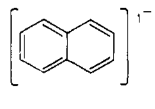


Figure 14.

Radical Ions. This class of compounds is handled quite simply by using the radical ion identification code (RI) and the charge of the radical ion as shown below.



RI.(1-)

APPENDIX C: REGISTRY III VALIDITY TABLES

Certain standard values associated with the chemical elements that are used in Registry III editing procedures are

Table II. Registry III Element Tables

Element	Valence	Mass Range	Stock No.	Charge/Bond Values			
Ac	3	216-231		+3/6			
Ag	1	100-117	1, 2	+1/2	+1/3	+1/4	+2/4
Al	3	23-30		+1/4	+3/4	+3/5	+3/6
				+3/7	+3/8		
Am	3	237-246	2, 3, 4, 5	+3/6	+3/8		
			6				
Ar	0	33-42					
As	3	68-85		−2/3	−1/3	−1/4	+1/4
				+3/4	+3/6	+5/6	
At	1	198-219					
Au	1	177-203	1, 3	+0/2	+1/2	+3/4	+3/6
B	3	8-13		−1/3	−1/4	+3/4	
Ba	2	123-144		+2/4	+2/6		
Be	2	6-12		+2/3	+2/4		
Bi	3	196-215	3, 5	+3/4	+3/5	+3/6	
Bk	3	243-251	3, 4	+3/6			
Br	1	74-90					
C	4	9-16					
Ca	2	37-50		+2/4	+2/6		
Cd	2	101-121		+2/1	+2/3	+2/4	+2/6
Ce	3	129-148	3, 4	+3/4	+3/6	+3/7	+3/8
				+3/9	+3/10	+4/6	+4/8
Cf	3	244-254		+3/6			
Cl	1	32-40					
Cm	3	238-252		+3/6			
Co	2	54-64	2, 3	−1/4	+0/4	+0/5	+0/6
				+1/4	+1/5	+1/6	+2/2
				+2/3	+2/4	+2/5	+2/6
				+2/7	+3/5	+3/6	+3/7
				+4/6			
Cr	6	46-56	2, 3, 6	+0/6	+1/6	+2/4	+2/6
				+2/7	+3/4	+3/6	+4/7
				+6/8			
Cs	1	123-144		+1/4	+1/6		
Cu	2	58-68	1, 2	+1/2	+1/3	+1/4	+2/1
				+2/2	+2/3	+2/4	+2/5
				+2/6			
Dy	3	149-167		+3/4	+3/6	+3/7	+3/8
				+3/9	+3/10		
Er	3	152-172		+3/4	+3/6	+3/7	+3/8
				+3/9	+3/10		
Es	3	245-256		+3/6			
Eu	3	143-160	2, 3	+2/4	+3/4	+3/6	+3/7
				+3/8	+3/9	+3/10	
F	1	16-22					
Fe	2	52-61	2, 3	−2/4	+0/5	+0/6	+1/5
				+1/6	+2/4	+2/5	+2/6
				+2/7	+3/4	+3/5	+3/6
				+3/7	+3/8	+4/6	
Fm	3	248-258		+3/6			
Fr	1	204-224		+1/4	+1/6		
Ga	3	63-76		+1/4	+3/4	+3/6	
Gd	3	145-162		+3/4	+3/6	+3/7	+3/8
				+3/9	+3/10		
Ge	4	65-78	2, 4	+2/4	+4/6		
H	0	1-5					
He	0	3-8					
Hf	4	157-183		+4/6	+4/7	+4/8	
Hg	2	185-206	1, 2	+1/2	+1/4	+2/2	+2/3
				+2/4	+2/6		
Ho	3	150-170		+3/4	+3/6	+3/7	+3/8
				+3/9	+3/10		
I	1	117-139					
In	3	106-124	1, 2, 3	+1/4	+3/4	+3/5	+3/6

Table II (*Continued*)

Element	Valence	Mass Range	Stock No.	Charge/Bond Values			
Ir	2	182-198	1, 2, 3, 4	+0/5	+0/6	+1/4	+1/5
			6	+2/4	+2/5	+3/5	+3/6
				+4/6			
K	1	37-47		+1/4	+1/6		
Kr	0	74-95					
La	3	124-144		+3/4	+3/6	+3/7	+3/8
				+3/9			
Li	1	5-9		+1/2	+1/4		
Lr	3	256-257					
Lu	3	155-180		+3/4	+3/6	+3/7	+3/8
				+3/9	+3/10		
Md	3	255-257		+3/6			
Mg	2	21-28		+2/3	+2/4	+2/5	+2/6
Mn	7	50-58	2, 3, 4, 6	−1/5	−1/8	+0/6	+1/5
			7	+1/6	+1/7	+2/3	+2/4
				+2/5	+2/6	+2/8	+3/5
				+3/6	+3/7	+4/6	+6/8
Mo	6	88-105	2, 3, 4, 5	+0/6	+0/7	+0/8	+1/7
			6	+2/4	+2/7	+2/8	+2/9
				+3/6	+3/7	+3/8	+4/5
				+4/6	+4/8	+4/10	+5/6
				+5/8	+6/6	+6/8	
N	3	12-18					
Na	1	20-26		+1/4	+1/6		
Nb	5	88-101	2, 3, 4, 5	−1/6	+2/6	+3/6	+3/7
				+3/8	+4/6	+5/6	+5/7
				+5/8			
Nd	3	137-151		+3/4	+3/6	+3/7	+3/8
				+3/9	+3/10		
Ne	0	17-24					
Ni	2	56-67	2, 3	+0/4	+2/3	+2/4	+2/5
				+2/6	+3/4	+3/5	+3/6
				+4/4	+4/6		
No	3	254-256		+3/6			
Np	5	227-241	3, 4, 5, 6	+3/6	+4/6	+4/8	+5/6
				+5/8	+6/8	+6/10	
O	2	13-20					
Os	2	180-195	2, 3, 4, 6	−2/9	+0/5	+0/6	+1/6
			8	+2/6	+2/8	+3/6	+4/6
				+5/6	+5/7	+6/8	
P	3	28-34		−2/3	−1/3	−1/4	+1/4
				+3/4	+5/6		
Pa	5	225-237	4, 5	+3/6	+3/8	+4/8	+5/6
				+5/7	+5/8	+5/9	+5/10
Pb	4	194-214	2, 4	+0/4	+2/3	+2/4	+2/6
				+4/6			
Pd	2	98-115	2, 4	+0/4	+2/4	+2/5	+4/6
Pm	3	140-154		+3/4	+3/6	+3/7	+3/8
				+3/9	+3/10		
Po	2	192-218	2, 4	+4/6			
Pr	3	134-149	3, 4	+3/4	+3/6	+3/7	+3/8
				+3/9	+3/10	+4/6	
Pt	2	173-201	2, 4	+0/4	+1/4	+1/6	+2/4
				+2/5	+2/6	+4/6	+5/6
Pu	4	232-246	2, 3, 4, 5	+2/10	+3/6	+4/6	+4/8
			6	+5/6	+5/8	+6/8	+6/10
Ra	2	213-230		+2/4	+2/6		
Rb	1	79-95		+1/4	+1/6		
Re	7	177-192	3, 4, 5, 6	−1/8	−1/9	+0/6	+1/6
			7	+1/7	+1/8	+2/6	+3/4
				+3/6	+3/7	+4/6	+5/6
				+5/7	+5/8	+6/8	
Rh	2	96-110	1, 2, 3, 4	+0/5	+0/6	+1/4	+1/5
			6	+1/6	+2/5	+2/6	+3/5
				+3/6	+4/6		

Table II (*Continued*)

Element	Valence	Mass Range	Stock No.	Charge/Bond Values			
Rn	0	204-224					
Ru	2	93-108	2, 3, 4, 5	−1/8	+0/5	+0/6	+1/7
			6, 7, 8	+2/5	+2/6	+3/6	+4/6
				+6/8			
S	2	29-38					
Sb	3	112-135		−2/3	−1/3	−1/4	+1/4
				+3/4	+3/6	+5/6	+5/8
Sc	3	40-51		+3/2	+3/6	+3/8	
Se	2	70-87		−2/2	−1/2	−1/3	+4/6
Si	4	25-32		−2/4	−1/4	+4/5	+4/6
Sm	3	141-158	2, 3	+3/4	+3/6	+3/7	+3/8
				+3/9	+3/10		
Sn	4	108-132	2, 4	+2/3	+2/4	+2/5	+4/5
				+4/6	+4/7	+4/8	
Sr	2	80-95		+2/4	+2/6		
Ta	5	172-186	2, 3, 4, 5	−1/6	+2/6	+3/6	+4/6
				+5/6	+5/7	+5/8	
Tb	3	147-164	3, 4	+3/4	+3/6	+3/7	+3/8
				+3/9	+3/10		
Tc	7	92-107	4, 7	−1/8	+4/6		
Te	2	107-135		−2/2	−1/2	−1/3	+2/6
				+4/5	+4/6		
Th	4	223-235	3, 4	+3/6	+4/6	+4/8	+4/9
				+4/10			
Ti	4	41-52	2, 3, 4	+2/6	+3/4	+3/6	+4/5
				+4/6	+4/7	+4/8	
Tl	1	191-210	1, 3	+1/4	+3/4	+3/5	+3/6
Tm	3	153-176		+3/4	+3/6	+3/7	+3/8
				+3/9	+3/10		
U	6	227-240	3, 4, 5, 6	+2/4	+2/6	+2/7	+2/8
				+2/9	+2/10	+3/6	+4/6
				+4/7	+4/8	+5/6	+5/8
				+6/8	+6/10		
V	5	45-54	2, 3, 4, 5	−2/5	+0/6	+1/4	+1/8
				+2/6	+2/7	+3/4	+3/6
				+4/5	+4/6	+5/6	+6/6
W	6	174-189	2, 3, 4, 5	+0/6	+2/4	+2/7	+2/8
			6	+3/6	+3/7	+3/8	+4/5
				+4/6	+4/7	+4/8	+4/10
				+5/6	+5/8	+6/6	+6/8
Xe	0	118-144					
Y	3	82-97		+3/4	+3/6	+3/7	+3/8
				+3/9	+3/10		
Yb	3	154-177	2, 3	+2/4	+3/4	+3/6	+3/7
				+3/8	+3/9	+3/10	
Zn	2	60-72		+2/2	+2/3	+2/4	+2/5
				+2/6			
Zr	4	81-99	2, 4	+2/6	+2/7	+4/5	+4/6
				+4/7	+4/8		

provided below. Each element has standard values of valence, mass, Stock number, and, for coordination compounds, charge/bond values.

The valence has the traditional chemical connotation of combining capacity for an element. In Registry III the valence value for a given element is calculated by adding the bond connections (single = 1, double = 2, triple = 3), the hydrogen count (hydrogen atoms attached to noncarbon atoms are indicated explicitly), and the absolute value of any indicated charge. Except for the element carbon, values for valence that differ from the standard value by an even number (2, 4, ...) are acceptable. Other nonstandard valence values and any such valence for carbon must be indicated at input. The presence of invalid valence values is signaled by an error message.

For each element a range of acceptable mass values has been determined. Input isotopic mass values that fall outside such a range cause an error message.

For metallic elements which exhibit variable valence, the valence is indicated in simple metal salt representations by associating the appropriate numerical value, or Stock number, with the metal. For such metallic elements, acceptable values for the Stock number have been established. The input of any other value causes an error message.

The "charge/bond values" listed in Table II are coordination compound edits and originate from the CAS conventions for representation of this type of substance. Basically, the oxidation state of the central coordinating atom is associated with it as a charge, positive, negative, or zero. For each oxidation state there is one or more accepted bonding configuration represented by the values shown in the table. For example, the element silver (Ag) has associated values of +1/2, +1/3, +1/4, and +2/4. This means that when silver has a charge of +1, the total bond connection may be 2, 3, or 4, with a single bond = 1, double = 2, and triple = 3. Input values that violate the valid values cause an error message.

REFERENCES AND NOTES

(1) J. E. Blackwood, P. M. Elliott, R. E. Stobaugh, and C. E. Watson, "The Chemical Abstracts Service Chemical Registry System. III. Stereochemistry", presented at the 170th National Meeting of the American Chemical Society, Chicago, Ill., Aug 1975.

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

(3) R. G. Freeland, S. J. Funk, L. J. O'Korn, and G. A. Wilson, "The Chemical Abstracts Service Chemical Registry System. II. Augmented Connectivity Molform—A Technique for Recognition of Structure Topology Identity", presented at the 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 1975.

(4) P. G. Dittmar and J. Mockus, "An Algorithmic Computer Graphics Program for Generating Chemical Structure Diagrams", presented at the 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 1975.

(5) N. A. Farmer, J. E. Blake, and R. C. Haines, "An Interactive Computer Graphics System for the Input of Publication Quality Chemical Structure Diagrams", presented at the 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 1975.

(6) G. G. Vander Stouw, C. E. Watson, J. D. Rule, and C. Gustafson, "The Chemical Abstracts Service Chemical Registry System. IV. Use of the Registry System to Support the Preparation of CA Index Nomenclature", presented at the 170th National Meeting of the American Chemical Society, Chicago, Ill., Aug 1975.

(7) R. S. Cahn, C. Ingold, and V. Prelog, *Angew. Chem., Int. Ed. Engl.*, **5**, 385–415 (1966).