

# A Chemical Notation and Code for Computer Manipulation\*†

DAVID LEFKOVITZ

The Moore School of Electrical Engineering, University  
of Pennsylvania, Philadelphia, Pa. 19104

Received May 4, 1967

**This paper describes and specifies the rules for generating a code to represent chemical structural formulas. It may be used either as a notation or for internal manipulation by computer for registration, screening, and atom-by-atom search. It is basically a connection table in a concise format which also contains abnormality information relating to specific atoms or relations between atoms. The notation produced by these rules unambiguously defines a structure (as a connection table), but the notation is unique only to a given numbering (or citation order) of the atoms. Hence, its uniqueness is dependent on the numbering uniqueness of the atoms in the structural formula. Techniques for the use of the code in registry and substructure search are also discussed.**

This paper describes and specifies the rules for generating a code to represent chemical structural formulas. It is referred to here as a mechanical chemical code (MCC) as it is primarily intended as a chemical code for computer automated storage and has properties that favor its manipulation in atom-by-atom search within the processor. It could also conceivably be used as a notation inasmuch as its rules of generation are relatively straightforward, but, except for use in compound registration, it does not appear to have any properties that would recommend it for manual or semi-automatic search procedures such as permuted listings.

The particular properties of the MCC that are of systematic interest are:

It has a one-to-one correspondence with a connection table (CT) and includes information such as abnormal mass, charge, and valence.

The conversion in either direction, from or to a CT, is completely automatic without requiring any chemical interpretation.

It is considerably more concise than a CT, in that it requires fewer characters (or bits) in digital storage.

If it is automatically generated from a CT, then it is unique if the CT from which it is generated is unique. (This is a consequence of the first property.) If it is manually generated then the rules of generation do not include uniqueness.

A modified molecular formula, called the coded molecular formula (CMF), can be generated from the MCC, which is useful for registration.

Because the MCC unambiguously represents a structural formula (to the extent that the corresponding CT unambiguously represents the structural formula) it is possible to perform

an atom-by-atom search of the structure when represented in the MCC.

The MCC formulation is directly based upon a chemical notation published by Hiz (1) in 1964. It follows any desired contour of the graph representing the structure and cites as a code symbol or implicitly accounts for every atom in the structure. Every symbol in the code has unique valence and all hydrogens are accounted for although not necessarily cited. The code formation rules and the unique valence of every symbol enables most bonds (of any type) to be accounted for without specific citation, which is the prime factor contributing to the conciseness of the code.

## MCC SPECIFICATION

The MCC specification is divided into three sections: (1) The MCC Alphabet, (2) The Use and Meaning of the Symbols, and (3) The Rules of the Code Formation. Each section is decimalized under its appropriate section number for easy cross reference.

1. **The MCC Alphabet.** The alphabet is a set of symbols which would be represented in digital storage by various binary bit combinations. The newer computers commonly use an eight-bit character (called a byte) so that 256 possible alphabetic characters can be readily represented by these machines.

The following alphabet contains only 56 symbols, but one category of symbols, called *descriptors*, is actually open-ended since this set of symbols may be augmented to suit special requirements. These descriptors denote additive information about specific atoms. The conventional abnormalities such as mass, charge, and valence are in this category. Stereo descriptors might also be devised, if suitable conventions could be established for their use.

\* Based in part on the paper "The Impact of third generation ADP Equipment on Alternative Chemical Structure Information System" presented before the Division of Chemical Literature, 153rd Meeting, ACS, Miami Beach, Fla., April 1967.

† This study was supported by the National Science Foundation under Contract No. NSF-C467.

# A CHEMICAL NOTATION AND CODE FOR COMPUTER NOTATION

## 1.1 The Alphabet

|                  |   |
|------------------|---|
| Letters:         | A to Z (upper case)<br>a,b,c (lower case) |
| Numerals:        | 0 to 9 (full size)<br>0 to 9 (subscripts) |
| Descriptors:     | -<br>:<br>*                               |
| Special Symbols: | (<br>)<br>+<br>,                          |

## 1.2 Special Combinations and Bond Types

| Symbol     | Valence   | Denotation  |
|------------|-----------|---|
| C          | 4         | $\begin{array}{c}   \\ -C- \text{ or } =C- \text{ or } =C= \text{ or } -C\text{I} \\   \end{array}$ |
| E          | 1         | Br  |
| F          | 1         | F   |
| G          | 1         | Cl  |
| I          | 1         | I   |
| J          | 5         | Pentavalent N   |
| M          | 2         | H<br>—N—  |
| Q          | 1         | OH  |
| L          | 2         | $\begin{array}{c} O \\    \\ -C- \end{array}$   |
| Z          | 1         | —NH <sub>2</sub>  |
| a          | 3         | CH  |
| b          | 2         | CH <sub>2</sub>   |
| c          | 1         | CH <sub>3</sub>   |
| $\alpha X$ | $\bar{2}$ | $\begin{array}{c} O \\    \\ \alpha = O, \alpha \text{ is any allowable element} \end{array}$       |
| R          |           | Benzene ring  |
| D          |           | Double bond   |
| T          |           | Triple bond   |

## 2. Use and Meaning of the Symbols

2.1 All elements use their atomic symbols unless they appear in one of the special combinations of 1.2. Two-letter element symbols are written with both letters as capitals, and are preceded by +.

2.2 Full-sized numerals denote locants or descriptor quantifiers.

2.3 Three descriptor symbols are at present included in the code. These are:

- abnormal valence
- : abnormal mass
- \* abnormal charge

2.3.1 All elements except N and C are assumed to have their lowest possible valence value. N is assumed to have valence 3, and C is assumed to valence 4. Any element having another valence assignment in a compound is said to have an abnormal valence, which is cited as  $-n\alpha$ , where  $n$  is the abnormal valence of  $\alpha$ .

2.3.2 Any element for which a mass is to be specifically cited is represented as  $:n\alpha$ , where  $n$  is the mass of element  $\alpha$ . The decision to cite such a mass is beyond the scope of these rules.

2.3.3 A charged atom  $\alpha$  is cited as  $*n\alpha$ , where  $n$  is the charge of  $\alpha$ .

2.4 Subscripts denote repetition of the subscripted element symbol or repetition of an entire expression enclosed in subscripted brackets.

Example (1)

$b_3$  means  $-(CH_2)_3-$

$(\beta)_2$  means  $\beta\beta$ , where  $\beta$  is any string of characters.

2.5 The full-sized numerals are used to designate locants and as descriptor quantifiers, as described in 2.3. Two or more contiguous locants are separated by a comma.

3. Rules of Code Formation. The following set of rules will enable any structural formula that can be represented as a connection table (CT) to be encoded as an MCC. The exact string of symbols in the MCC for a given structure is a function only of the particular numbering of the nonhydrogen atoms in the CT; hence a given structure may produce as many MCC's as there are ways of numbering its atoms. The element combinations CO and O<sub>2</sub> are represented by the single MCC symbols L and X, respectively, so that only a single CT number would be assigned to these combinations.

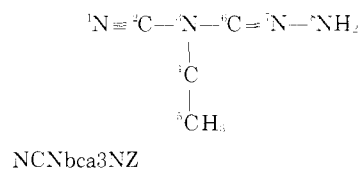
3.1 Every element symbol in an MCC has a unique valence, in accordance with 1.2 and 2.3.1.

3.2 The element symbols are cited in a sequence, from left to right, according to the CT numbering sequence.

3.3 A bond is cited only where an element of the CT is connected to an element with a lower number which is *not* the immediately preceding number. The bond is cited as a locant number immediately following the higher numbered element symbol, and the locant number is the number of the lower numbered element symbol. If the bond is single, no letter precedes the locant number. If it is double, the letter D precedes the locant number, and if it is triple, a T precedes the locant number. If an element is bonded to two or more lower numbered (but not immediately preceding) elements, the locant numbers may be cited in any order and are separated by commas. (The proof that this rule is sufficient to enable the unambiguous encoding of any structure is presented in the Appendix.)

3.3.1 Given the substructure  $\alpha\beta\gamma$ , where  $\alpha$  is a branched element with CT number  $n$ ,  $\beta$  is a single element symbol branch with CT number  $n+1$ , and  $\gamma$  is an element attached to  $\alpha$ , with CT number  $n+2$ , then it is *not* necessary to cite the  $\alpha$ - $\gamma$  bond as the locant number  $n$ .

Example (2) (Rule 3.3)



Based upon the analysis presented in the Appendix, the bonding implied by this code may be illustrated graphically by the following diagram and three application rules:

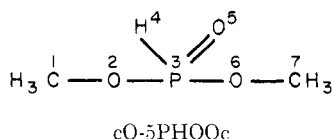
Generate all bonds explicitly cited by locant numbers.

Generate all bonds, starting from the left, by requiring that a given element symbol be bonded to the next succeeding symbol that has not had its valence satisfied, with the highest possible bond value where the predecessor gives as high a bond as it can consistent with its currently unassigned valence, and the successor takes as high a bond



3.4 All hydrogens not implied by combination symbols are cited explicitly.

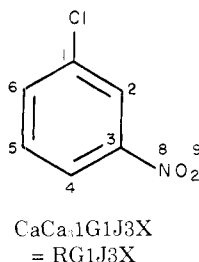
Example (8)



3.5 The benzene ring may be represented by the symbol R if all of its atoms are numbered consecutively. That is,

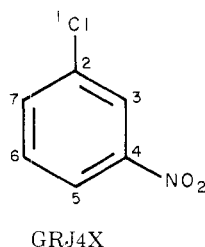
$R = \{a \text{ or } C\} \{a \text{ or } C\} \{a \text{ or } C\} \{a \text{ or } C\} \{a \text{ or } C\} \{a \text{ or } C\}$

Example (9)



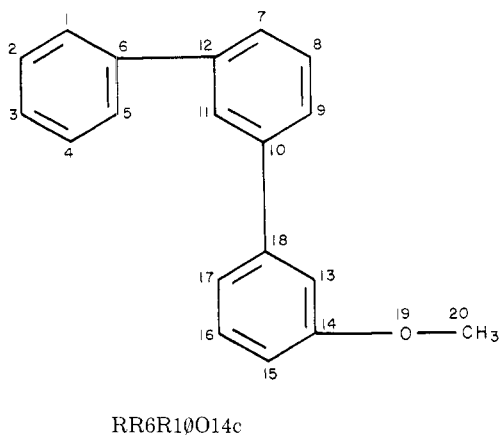
This structure could be renumbered to give

Example (10)

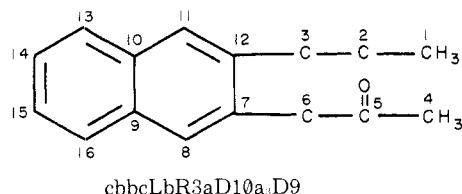


3.5.1 If R is used to represent a benzene ring then at most one element of the ring may be attached to a lower (but not immediately preceding) numbered element, and this element must be the highest numbered atom of the benzene ring.

Example (11)



Example (12)



## SYSTEM APPLICATIONS OF THE MCC

**The Coded Molecular Formula.** The molecular formula (MF) of a compound provides an element by element count of the atoms in the structural formula; analogously the *coded molecular formula* (CMF) provides a symbol by symbol count of the element (or combination) symbols in the MCC. For example, the MF of Example 2 is  $C_4H_8N_4$ ; The CMF is  $abcCN_4Z$ . As another example, the MF of Example 5 is  $C_5H_{12}N_2OS$ ; the CMF is  $bc_1CMNO-4S$ . Note that the unique valence of every symbol is retained in the CMF, so that the sulfur appears in the CMF as  $-4S$ .

Only element symbols (with + for two-symbol elements), combination symbols and abnormality (with associated quantifier and element) symbols appear in the CMF. Locants, parentheses, bond symbols, and R do not appear. If these rules for the formation of CMF are followed, *all* of the possible MCCs for a given structure will produce the same CMF. That is, a particular structural formula has one and only one CMF regardless of how its CT is numbered and therefore regardless of the form of its MCC. For example, the CMF of Example 3 is  $bcCQS$ . Example 4 illustrates the same structure renumbered. The resulting MCC is different, but the CMF is still  $bcCQS$ .

As a final example, the MF of Example 9 is  $C_6H_4ClNO_2$ ; the CMF is  $a_4C_2GJX$ .

The CMF can serve two purposes in the system. First, it can be used as a course screen for substructure search. This topic will be discussed further in the next subsection.

Second, it can be used as a more efficient registry technique than isomer sort registration. That is, all compounds would be sorted by CMF instead of MF. Each CMF isomeric group obviously is a subset of the corresponding MF, and the isomeric CMF group will probably be very small because C in the MF is partitioned into C, a, b and c, and other differentiations such as by Z, Q, M, J, and abnormalities are also made. For example, the CAS registry system contains 20 compounds in the isomeric group  $C_6H_{12}O_6$ . Aside from stereoisomers, there is only one pair of compounds in this group that is not distinguished by the CMF.

The isomeric group  $C_{10}H_{18}O$  has 16 members and aside from stereoisomers, there are three different compounds having the same CMF; all of the remaining compounds are distinguished.

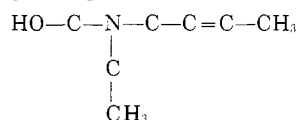
Within a CMF isomeric group structures could be distinguished by the symbol connectivity search to be discussed in the next subsection.

**Substructure Search.** As there exists a one-to-one relation between the CT and MCC it should be possible to perform an atom by atom search in the MCC with the same definitiveness as in the CT. However, this type of iterative search should be called a *symbol* rather than an *atom* connectivity search. It appears that not only can the equivalent of an atom by atom search be performed, but the search may be more efficiently implemented, for the same reason that the CMF is a more efficient classifier of compounds than the MF. That is, the homogeneous C in the CT, which leads to most of the combinational difficulty (and backtracking) in the iterative search, is differentiated in the MCC as four separate symbols, each

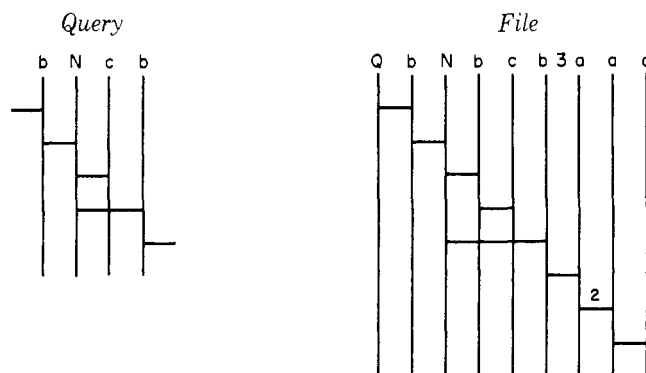
with its own unique valence. Similarly the other special combinations provide further power in reducing the backtracking (or, in the case of the Sussenguth (2) method, of reducing the sets).

There would, however, be one imposition placed upon the form of the substructure query in order to most efficiently perform this type of search—namely, that the bordering (structural) environment of the substructure be stated in the query. That is, the kind of bonding allowed at the open ends of the query should be stated. If the bonding is indefinite, then the search may still be effected by the logical OR, but it is less efficient. Table I presents a set of substructure queries; in the first column is the imprecise statement of the query; in the second column is a possible correct form of the query, and in the third column is the MCC for the query.

Consider the application of the second query in Table I to the following example.



The topology of query and file compound are as follows:



The CMF of the file compound is  $a_2b_3c_2NQ$ ; the CMF of the query is  $b_2cN$  and being a subset of the compound CMF passes this screen. (It *must* pass this screen in order to be a substructure.)

The iterative search proceeds in the following steps.

Query N matches File N

Query N connects by a single bond to a c but file N does not. Note: The c connection in the query is chosen as the next test over the b connection because there is one c and two b's; hence, less ramification.

There are no more Ns in the file; therefore, no substructure match.

As another example, the third query *vs.* Example 5 would be screened out by the CMF, but the fourth query would pass both the CMF and the symbol connectivity search.

**Compression of the Code in Digital Storage.** A principal reason for using the MCC is its conciseness.

The MCC lends itself to a further compression due to the high frequency of use of a subset of the 56 alphabet symbols. The computer character in the direct access

Table 1. Examples of Generic Search Queries in MCC



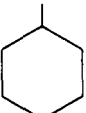

|   |  |   |
|---|--|---|
| Query<br>$\text{N}=\text{C}=\text{N}$   | Corrected for MCC<br>$-\text{N}=\text{C}=\text{N}-$  | MCC of the Query<br>NCN   |
| $\text{C}-\text{N}-\text{C}$<br> <br>$\text{CH}_3$  | $-\text{C}-\text{N}-\text{C}-$<br> <br>$\text{CH}_3$   | bNcb  |
| $\text{N}-\text{S}-\text{C}$  | $=\text{N}-\text{S}-\text{C}-$<br>or<br>$-\text{N}-\text{S}-\text{C}-$   | NSa   |
| $\text{N}-\text{S}-\text{C}$<br> <br>$\text{O}$   | $=\text{N}-\text{S}-\text{C}$<br> <br>$\text{O}$   | N-4SO{ a or b or c or C }   |
| $\text{H}_2\text{N}-\text{C}$<br> | $\text{H}_2\text{N}-\text{C}-\text{Ring}$<br> | $\text{Z}\{ \text{a or C} \}$<br><br>$(\{ \text{a or b or C} \})_5$ |
|                                   |   | $\text{a}(\{ \text{a or b or C} \})_5$                              |

Table II. The Bi-Octal Code for MCC

|  |        | Four-Bit Code |                       |
|--|--------|---------------|-----------------------|
|  |        | Bit 1         | Bits 2,3,4<br>(Octal) |
| MCC Symbol   |        |               |                       |
| a  | $\phi$ | 0             | 0                     |
| b  | 1      | 0             | 1                     |
| c  | 2      | 0             | 2                     |
| C  | 3      | 0             | 3                     |
| N  | 4      | 0             | 4                     |
| O  | 5      | 0             | 5                     |
| Q  | 6      | 0             | 6                     |
| M  | 7      | 0             | 7                     |
| Z  | 8      | 1             | 0                     |
| L  | 9      | 1             | 1                     |
| X  | E      | 1             | 2                     |
| R  | ( F    | 1             | 3                     |
| D  | ) G    | 1             | 4                     |
| T  | , I    | 1             | 5                     |
| Following code is a subscript<br>or - ( ) , P S          | P J    | 1             | 6                     |
| Following code is a full sized<br>numeral or E F G I J H | S H    | 1             | 7                     |

storage is eight binary bits, but it can also be used in a four-bit mode. Table II presents a four-bit coded symbol set that will enable all compounds containing the elements C,H,N,O,S,P,Cl,Br,F,I (Four more elements could be added if the numeral digits were 0 to 7 instead of 0 to 9) exclusively to be coded almost completely in four-bit characters.

The MCC characters for full and subscripted numerals -, ( ), PSEFGIJ and H all require eight bits because they

must be preceded either by the 16 or 17 code. That is, E is the eight-bit bi-octal code 1712; the subscript 6 is 1606, etc. All of the other symbols in Table II are represented by four-bit codes. For example, the MCC for Example 6 is:

10 03 00 03 01 05 00 06 03 05 03 1704 04 00 14 1702  
Z C a C b O a Q C O C 4 N a D 2

Compounds containing elements other than those found in Table II would be encoded by a straight eight-bit character code, but as a large proportion of the compounds in an organic file of 3 million compounds would be covered by the elements of Table II, the average number of bits per MCC symbol would lie between four and eight, and very possibly closer to four.

## APPENDIX

### A FORMALIZATION OF THE SYMBOL ORDERING SYNTAX

The implicit definition of bonds as a consequence of the symbol ordering rules 3.2 and 3.3 is basically a function of the fact that every symbol has unique valence and these rules are actually an implementation of the principle of *first in first out* (FIFO), which can be stated more formally as follows:

#### Principle of FIFO.

The structural formula is reconstructed from the MCC by first inserting all bonds explicitly designated by locants.

All remaining bonds are implicitly defined according to a left to right scan of the MCC employing a FIFO principle. That is, the valence of a given element in this scan must be completely satisfied by as many succeeding or preceding symbols before proceeding to the next symbol. The bond type between two symbols  $\alpha$  and  $\beta$  is determined as follows:

$\alpha$  has valence  $v_\alpha$ .

$\beta$  has valence  $v_\beta$ .

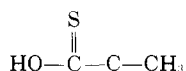
$\alpha$  has  $m_\alpha$  of its valence units already satisfied either by preceding element symbols or locants, and  $\beta$  has  $m_\beta$  of its valence units similarly satisfied.

The bond type between element symbols  $\alpha$  and  $\beta$  is then given as:

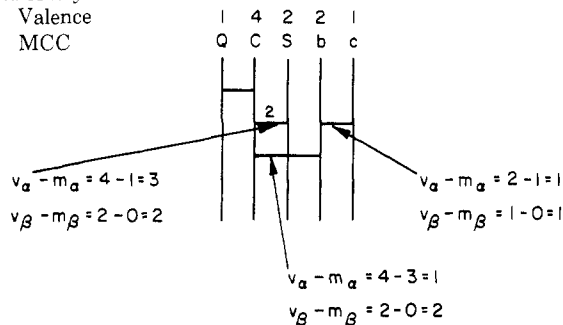
$$\text{Min } \{ (v_\alpha - m_\alpha), (v_\beta - m_\beta) \}$$

The following four examples illustrate this principle. For convenience the notation for  $(v_\alpha - m_\alpha)$  and  $(v_\beta - m_\beta)$  is shortened to  $\Delta_\alpha$  and  $\Delta_\beta$ , respectively, in Examples A2 through A4.

#### Example (A1)

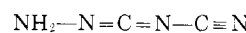


Bond Analysis  
Valence  
MCC

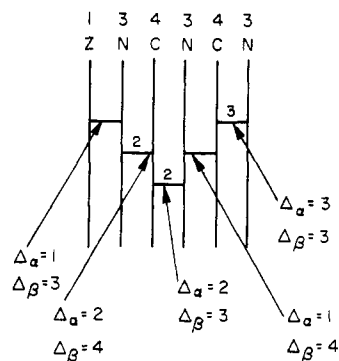


NOTE: The last connection must satisfy the relation  $v_\alpha - m_\alpha = v_\beta - m_\beta$ , or there is an error in the code.

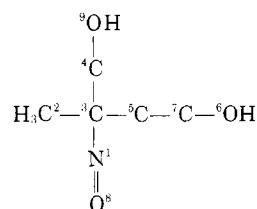
#### Example (A2)



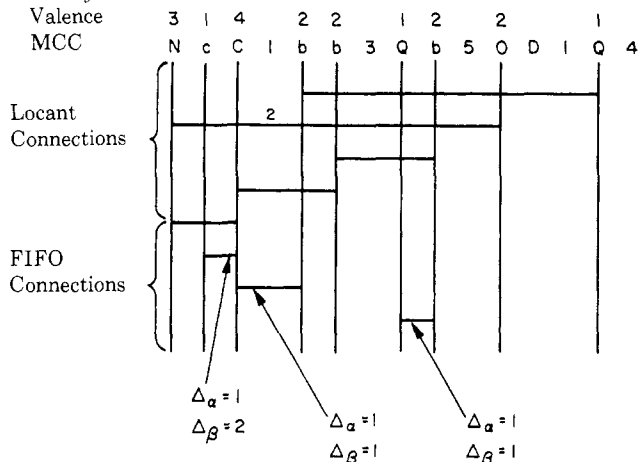
Bond Analysis  
Valence  
MCC



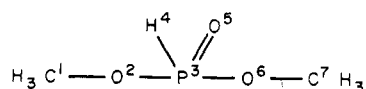
#### Example (A3)



Bond Analysis  
Valence  
MCC

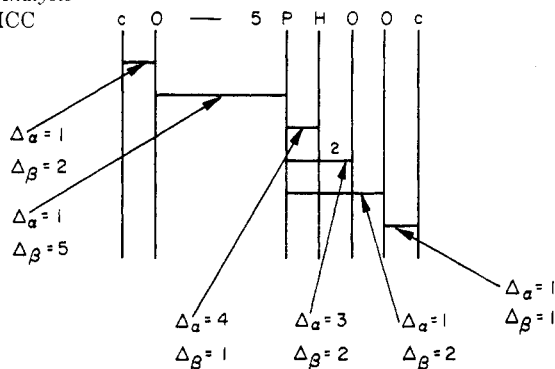


Example (A4)



Bond Analysis

MCC



## ACKNOWLEDGMENT

The author gratefully acknowledges the suggestions of and consultations with the following persons. D. Loev, M. Plotkin and J. Munz, and C. T. Van Meter of the University of Pennsylvania, A. Genarro of the Philadelphia College of Pharmacy and Science, Sylvan Eisman of the Frankford Arsenal, William Wiswesser of Fort Detrick, P. Olejar, S. Rhodes, and T. Quigly of the National Science Foundation, J. Mitchell of the Edgewood Arsenal, and Fred Tate and his staff of the Chemical Abstracts Service.

## LITERATURE CITED

- (1) Hiz, J., J. CHEM. DOC. 4, 173 (1964).
- (2) Sussenguth, E. H., *Ibid.*, 5, 36 (1965).

## Use of a Nonunique Notation in a Large-Scale Chemical Information System\*†

DAVID LEFKOVITZ

The Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, Pa. 19104

Received May 4, 1967

This paper examines the functional requirements of an automated chemical information system as it might be implemented on third generation ADP hardware. Of primary concern in this examination is the formal representation of the structural formula within the automated files and the requirements placed upon this representation by the five system functions of Input, Registry, Storage for Search, Search, and Display. The various representations are divided first into the two broad categories of connection tables and notations. These are then broken into categories of unique and nonunique representations. Also examined are ease of automatic generation and manipulation of the various representations. The paper also presents a discussion of a desired systems approach to registry and search for real time, interactive operation and a final recommendation for structural formula representation in this type of system.

The primary objective of this paper is to discuss the combination of the concept of a nonunique notation or code for representing chemical structural formulas and the functional characteristics of third generation ADP equipment as they relate to the design of a large scale automated chemical information system. The most significant feature of such a system, which sets it apart from most other information storage and retrieval systems, is the requirement to search a large number of chemical structures, wherein the decision process to accept or reject

a given structure for a given query can be complicated enough to require several seconds of processor time. Because a large scale system would ultimately have to handle about 3 million compounds, retrieval on a serial search basis would be prohibitive; therefore, screens, or retrieval keys, are assigned to each compound, analogous to the retrieval keys of a document in a document retrieval system, so that a more rapid decision can be made on a large percentage of the file, simply on the basis of a match between the keys in the query and corresponding keys in a given file compound.

The new generation of computer hardware provides direct access storage for several hundred million characters of information. That is, access may be had to any single record among a several-hundred-million character store

\*Based in part on the paper "The Impact of Third Generation ADP Equipment on Alternative Chemical Structure Information System" presented before the Division of Chemical Literature, 153rd Meeting, ACS, Miami Beach, Fla., April 1967.

†This study was supported by the National Science Foundation under Contract No. NSF-C467.