information storage and retrieval problems. For maximum efficiency radixes other than 8 and 10 would be needed for some implementations.

ACKNOWLEDGMENT

The author would like to thank Dr J. K. MacLeod for his comments and Mrs. Greta Pribyl and Ms. Lorraine Scarr for their help in preparing the manuscript. The author also wishes to thank Dr. G. W. A. Milne and Dr. S. R. Heller for providing the mass spectral database.

REFERENCES AND NOTES

(1) C. T. Meadow, "The Analysis of Information Systems", Melville Publishing Co., Los Angeles, Calif., 1973, pp 314–344.
(2) N. Wirth, "Algorithms + Data Structures = Programs", Prentice-Hall, Englewood Cliffs, N. J., 1976.
(3) R. G. Dromey, "A Compact Free-Keyword File Structure for Author-Title-Keyword Searching", *J. Chem. Inf. Comput. Sci.*, **18**, 160 (1978).
(4) D. Lefkovitz, "The Large Data Base File Structure Dilemma", *J. Chem. Inf. Comput. Sci.*, **15**, 14 (1975).
(5) G. K. Zipf, "Human Behaviour and the Principle of Least Effort", Addison-Wesley Publishing Co., Cambridge Mass., 1949, pp 19–55.
(6) C. N. Mooers, "Zatocoding Applied to Mechanical Organization of Knowledge", *Am. Doc.*, **2**, 20 (1951).
(7) S. R. Heller, "Conversational Mass Spectral Retrieval System and Its Use as an Aid in Structure Determination", *Anal. Chem.*, **44**, 1951 (1972).
(8) G. M. Pesyna, F. W. McLafferty, R. Venkataraghavan, and H. E. Dayringer, "Statistical Occurrence of Mass and Abundance Values in Mass Spectra", *Anal. Chem.*, **47**, 1161 (1975).

A Simple Tree-Structured Line Formula Notation for Representing Molecular Topology

R. GEOFF. DROMEY*

Research School of Chemistry, Australian National University, P.O. Box 4, Canberra, A.C.T. 2600, Australia

Received June 15, 1977

A new linear notation for representing molecular topology is described. This canonical number-based approach is able to match the encoding power and systematics of existing systems (e.g., WLN) while using a much simpler encoding formalism. As such, the system should be much easier to use both manually and computerwise.

INTRODUCTION

Considerable effort has been expended in exploring linear notations for computer representation of molecular topology. Wiswesser line notation[1] (WLN), Dendral,[2] Hayward,[3] IUPAC,[4] and Skolnik,[5,6] are among the most concise and powerful general approaches to representation of molecular topology. These notations have become widely accepted among information scientists, and there would appear to be little point in adding another notation to the list unless it possessed some significant advantages over the current notations. A tree-structured "numeric" linear notation is proposed as a viable alternative *because it is able to offer a much simpler encoding (and decoding) formalism than existing methods.* The system has been designed to be computationally more attractive and to possess better indexing properties. The basis alphabet for this system is the hexadecimal character set (a 4-bit code), that is, the numerals from Ø to 9 and the characters from A to F. With this concise "alphabet" (cf. WLN which uses 40 characters—an 8-bit code on most computers) and, on a relative scale, only a minimal set of rules and conventions, it is possible to match the encoding power and systematics of existing notations. The first task that must be faced in designing a notation is the atom representation convention. A glossary of representation conventions is given in Tables I and II.

1. CONVENTIONS FOR REPRESENTING ATOMS

A very sound design criterion in existing systems has been to choose wherever possible symbols familiar to chemists. To try and meet this design criterion in the present system, the three most commonly occurring elements in organic compounds (apart from hydrogen), carbon, oxygen, and nitrogen, are assigned numeric representations according to their common valences, e.g., carbon = 4, oxygen = 2, and nitrogen = 3. The most abundant isotopes of all other elements in the periodic table take on a representation which is a function of their valency and atomic weight. The reason for this choice will become clear when the precedence rules are discussed. The basic format is

A[valence][atomic weight(rounded)]

Some examples are

A135 = chlorine A232 = sulfur A2Ø9 = beryllium

The symbol "A" (A for *A*tom) is used *only* for atomic representation. Elements with atomic weights greater than 99 and the representation of other than the most abundant isotopes are discussed in a more detailed report.[7]

2. RULES OF PRECEDENCE FOR ENCODING MOLECULAR STRUCTURES

To obtain an encoding scheme that is unique and unambiguous it is necessary to use a hierarchical set of precedence rules. In order to make the rules simple to apply they have been framed in terms of valency, connectivity, and atomic weight, concepts which are both very familiar to chemists and easy to work with. The central rules encompassing these concepts can be stated as follows.

General Rules of Precedence: At each stage in encoding a structure always choose to encode first

(A) *that connected path with an atom of smallest valence attached earliest*;

(B) *where minimum valence does not resolve the path choose to encode first the path with the atom of smallest atomic weight attached earliest*;

(C) *if resolution still has not been made, encode along the path that contains an atom with the least number of atoms attached earliest*;

(D) *finally, if none of the other constraints resolves the path, choose to encode along the path that has an atom with the least number of hydrogens attached earliest.*

This set of rules is applied hierarchically wherever precedence must decide which atom (or ring) is to be encoded next.

* Adress correspondence to author at Department of Computing Science, University of Wollongong, Wollongong, N.S.W. 2500, Australia.

0095-2338/78/1618-0225$01.00/0

Table I. Glossary of Representation Symbols

REPRESENTATION	MEANING
1. ATOM REPRESENTATION	
4	Carbon
3	Nitrogen
2	Oxygen
A[Valence][Rounded Atomic Weight]	For most abundant isotopes of atoms other than C, N, O.
A[Valence] D [3-digit Atomic Weight]	Atoms with rounded atomic weights greater than 99.
AA[Isotopic Weight] A [Valence][Ref. A.W.]	For other than most abundant isotopes e.g. ^{34}S.
AC[Effective Valence][Atomic Weight]	For singly charged atom.
ACC[Effective Valence][Atomic Weight]	Doubly charged atom.
2. BOND REPRESENTATION	
11	Double Bond.
111	Triple Bond.
CB	Bond linking repeating units of a polymer.
DB	Dative Bond ($\rightarrow$).
DC	Adduct connection.
DE	Pi-bond.
EF	Single Bond link between rings.
3. ACYCLIC BRANCHING CONVENTION	
5...9	The 5 and 9 parenthesize branches attached to the atom preceding the 5.
4. POLYMERS	
C	All acyclic polymers begin with a C.
ØC	All polymers with rings begin with ØC.

Parts A, B, and C of the rules of precedence are not applied to hydrogens.

As with most other notations a connectivity network is created, single bonds are not cited, and hydrogens are implied in accordance with common valence. Most of the tools necessary for encoding acyclic structures have now been developed.

3. ENCODING ACYCLIC STRUCTURES

The task of encoding acyclic structures can be simplified further by placing an additional constraint on the starting atom. To do this the concept of a *terminal atom* is introduced.

TERMINAL ATOM:

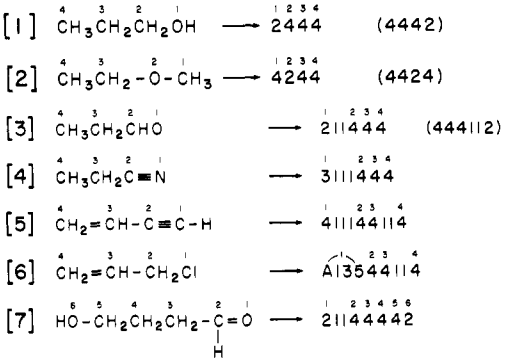
"A terminal atom is an atom bonded to only one other nonhydrogen atom."

The constraint required can now be expressed in the following rule:

RULE 1

Encoding for all purely acyclic structures must begin at a terminal atom.

Applying these encoding rules the following notations are obtained for some simple acyclics. Several alternative noncanonical representations are included in parentheses. Nonhydrogens have been numbered according to their order of encoding. In examples [1] → [6] atoms 1 and 4 satisfy the definition of a terminal atom.



Notice that the encoding has produced the smallest left-justified numeric representation for the structure that it is possible to write (e.g., 4244 < 4424). The notation has been designed so that in general this minimum numeric condition parallels with the canonical representation although on rare occasions for some complex ring systems and branched structures the minimum condition does not hold. Double and

triple bonds are represented by "11" and "111", respectively.

The other main rule needed to describe acyclics pertains to branching.

3.1 Branching Convention for Acyclic Molecules. The encoding of branched acyclic structures follows the principle of depth-first tree enumeration in that all segments of a chosen branch must be encoded before returning to the original branch point to begin the encoding of the next branch. This approach preserves the character of the individual branches. The numerals 5 and 9 are used to parenthesize individual branches. The degree of branching at an atom is indicated by the number of 5's that immediately follow the encoding for that atom. The 9's terminate all *successive branches except the last.* A pair of 5's (i.e., 55) indicates a tertiary branch point while a single 5 indicates a secondary branch. Rule 2 given below establishes the branching convention.

RULE 2

If after encoding a given atom "A" there are "n" unencoded nonhydrogen atoms attached to atom A then indicate the associated branch point with (n − 1) 5's. Branching groups are then encoded in order according to the rules of precedence and each branch except the last is terminated with a "9".

Examples are:

[8] $CH_3CH_2\overset{\overset{O}{\|}}{C}-O-CH_3$ → 2114524944

[9] $CH_3CH_2CH_2\underset{Cl}{CH}CH_2F$ → A119445A1359444

[10] $\begin{matrix} & CH_3 & & CH_3 & \\ & \diagdown & & | & \\ CH_3- & CH & -C- & CH_2CH_2CH_3 & \\ & \diagup & & | & \\ & CH_3 & & CH_2CH_3 & \end{matrix}$ → 44554549494449 44

Notice in example 10 how the canonical path is chosen. Terminal atoms are 1, 4, 5, 8, and 10 as numbered. Since each of the atoms is equivalent, it is necessary to look at the atoms to which each of the terminal atoms is attached in order to choose the canonical path. Since all of the atoms attached to the terminal atoms are of valence 4, have the same atomic weight, and have the same number of atoms attached, then precedence rule D concerning the minimum number of hydrogens attached must be applied in order to establish the canonical path.

To avoid the need to encode all atoms in a long unbranched alkyl chain, a multiplier convention is applied. This is discussed in detail elsewhere.[7]

Organic salts, acyclic systems with dative bonds, and hydrates, etc., are also discussed elsewhere.[7] The major aspects of acyclic molecular topology have been discussed and so the framework has been laid for dealing with the more complex task of encoding molecular ring topology.

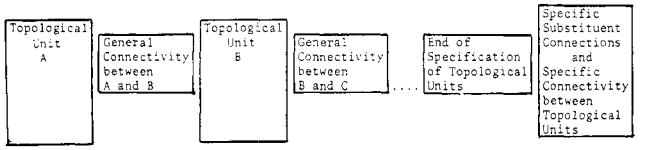
4. ENCODING RING SYSTEMS—INTERNAL TOPOLOGY

The more interesting and challenging aspects of the numeric linear notation are to be found in the encoding of ring systems. The underlying strategy is basically the same as that used for handling acyclic structures. Additional conventions are needed but the rules of precedence still provide the systematizing force behind all encoding.

Obviously a shorthand method for describing ring structures is essential since the atom-by-atom description employed for acyclics would be too cumbersome when extended to ring structures.

Unlike the approach employed in WLN, a single convention is applied to all ring systems including those with bridge substructures. It is convenient to divide ring topology into two fundamental parts, internal topology, and that which relates a ring to other rings. In this context fused ring structures are considered as basic topological units (as for rings) while at the same time the member rings of a fused system all retain their own identity. In this respect the present system differs significantly from WLN, where, for fused systems, the identity of individual rings is not so clearcut. From a practical standpoint the present system treats fused rings in much the same way as the chemist works with such systems.

Treating fused ring systems as basic topological units is important when it comes to expressing both the general and specific aspects of inter-ring connectivity. The overall connectivity hierarchy for ring systems can be generalized by the following diagram:



Both within isolated systems and for relations between topological units the trend is always from the more general to the more specific aspects of the overall structure. For example, the specific positions of substituents are not encoded in the notation until well after their structure has been indicated. The idea of structuring the notation from the general to the specific is essential if the notation is to possess good indexing properties and at the same time provide straightforward and effective substructure retrieval.

The notation treats both isolated and fused rings in a similar manner. The major advantage of this strategy is that it provides appreciable economy in the encoding of fused ring structures. The parameter most basic to internal ring topology, ring size, is assigned the numeral corresponding to the number of atoms in the ring (e.g., 5 is used in the notation when encoding a five-membered ring). Another basic internal feature is the presence and position of heteroatoms. They take on the same representation as in acyclic molecules (e.g., oxygen = 2, and sulfur = A232, etc.). Each heteroatom is followed by its associated "position-in-the-ring" numeral which is derived directly from the rules of precedence. The level of unsaturation of a ring can be classified as fully conjugated, partially saturated, or fully saturated. Acyclic substituents are considered to be part of a ring's topology and so they are represented internally. Ring fusion positions are also considered to be internal.

With respect to the actual notation an ordered set of fields is set up to describe each aspect of a ring's topology. Separate pieces of information within a field are punctuated by a subfield terminator (7). Different fields are separated by an end-of-field terminator (8). The end-of-ring terminators B, C, and BC, are used respectively to terminate fully conjugated (including aromatic), saturated, and partially saturated rings.

The encoding for all structures containing at least one ring must begin with a zero. This is directly followed by a numeral indicating the number of rings in the structure. Then follows the encoding for the first ring as determined by the rules of precedence. The basic format is

Ø[ring count][encoding for first ring]. . .

Should there be a heteroatom present in any of the rings in the structure then an extra zero is added directly after the ring count and so we get

Ø[ring count]Ø[encoding for first ring]. . .

If there are more than six rings in a structure, a "D" precedes the ring count, and so we have (D for *D*ouble)

ØD[ring count one or two numerals]
[encoding for first ring]. . .

The "D" is used widely throughout the notation to indicate that the number directly following it is greater than 6. A single D followed by a numeral is used *exclusively* for this purpose.

To follow on the practice adopted for acyclic systems, the representation conventions for cyclic structures have been designed so that the order of encoding for different substructures will correspond in the main to the smallest left-justified numeric representation (not necessarily the smallest number of characters) that it is possible to write for the structure.

4.1 Basic Encoding Formalism for Rings. There are two basic considerations in encoding ring systems. In a multi-ring system an unambiguous procedure is needed to define which ring should be encoded first. With this decision made, a method is needed to determine the order for encoding each aspect of the ring's topology. In many cases these two considerations cannot be differentiated and so the same principles are applied to answer both questions. This makes the encoding methodology straightforward in that the one set of basic principles is used to make all the decisions.

To simplify further the procedure for encoding ring systems, two additional constraints (Rule 3) are used to supplement the rules of precedence.

RULE 3

When encoding a ring system

(a) *Encode first at each stage in the connected path the ring of smallest size.*

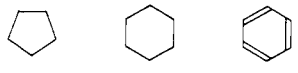
(b) *A fused ring takes precedence over an isolated ring of the same size.*

These last two constraints play an important role in helping to make the canonical notation also the smallest left-justified numeric notation. In resolving which ring to encode first, Rule 3 should be applied before general precedence. The connected path referred to in the rules of precedence is, in the ring context, the path that takes in all aspects of internal ring topology. *Preference is given to ring heteroatoms over substituent heteroatoms in applying the rules concerned* with *valence and atomic weight.* Isolated fused rings take their numbering from the rules of precedence.

Always in encoding a ring system only the combination of ordered fields present is encoded. For isolated rings that have no heteroatoms, substituents, or double bonds that are not fully conjugated, the only fields needed are ring size and the degree of saturation indicator. The encoding format for individual rings is then just

[ring size][degree of saturation indicator]

Some simple rings have the following notations.



ϕ15C ϕ16C ϕ16B

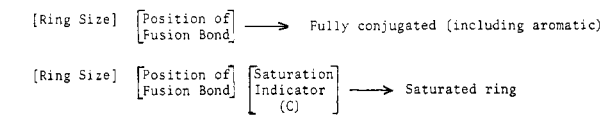
[11] [12] [13]

Rings [11] and [12] are fully saturated and are consequently terminated with a "C" whereas ring [13] is fully conjugated which requires that it be terminated with a "B". Because the discussion is following the order in which the fields are specified, partially saturated rings will be treated later (section 4.3C)

4.2 Encoding Fused Ring Systems. A conscious effort has been made in developing the present notation to make the handling of fused rings systems as simple and compact as possible. In other notations both the encoding and decoding of complex fused ring structures have proved to be difficult tasks.

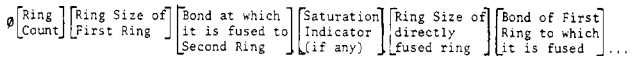
The principle of the depth-first tree enumeration used earlier for acyclic structures translates most naturally to a breadth-first tree enumeration in the context of fused ring systems when it is taken together with the rules of precedence. As more complex fused ring structures are discussed, the power of this simple algorithmic approach will become clearer.

The simplest type of fused ring systems is that where the rings contain no heteroatoms or substituents. Individual fully conjugated and saturated rings then have the format



For fully conjugated fused rings of this most simple type, the "B" end-of-ring terminator is dropped.

A very common fused ring system is that where the first ring, as established by the rules of precedence, has only rings present that are directly fused to it. In this situation the basic format for the notation is



For encoding acyclic structures the procedure was simplified by requiring that the encoding start at a terminal atom. An analogous constraint that requires the definition of a terminal fused atom is employed to encode fused ring structures.

TERMINAL FUSED ATOM:

"A terminal fused atom is a ring atom that is shared by only two fused rings."

The constraint required can now be expressed in the following rule.

RULE 4

(a) *To encode a fused ring system the numbering of the atoms in the first ring encoded must, wherever possible, begin at a terminal fused atom and follow the path as constrained by the rules of precedence that takes it* first *through a fused bond.*

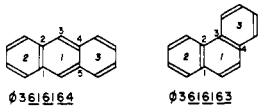
(b) *In fused systems the positional numbering of a ring other than the first is always taken from the numbering of atoms in the ring to which it is fused that has been encoded earliest.*

The numbering of the first fused ring is thus able to completely define the numbering of atoms for the rest of the fused rings present in the topological unit. Once the numbering of a pair of atoms in a ring has been derived from a previously encoded ring, it extends accordingly to cover the range from 1 → ring size for that ring (see ring 3 of example [17] below). Applying the principles that have been given, some simple systems can be encoded as follows. The encoding for individual rings has been underlined.

ϕ2616I ϕ2616IC

[14] [15]

Where an otherwise saturated ring is fused to an aromatic ring, it (the saturated ring) is encoded as though it were fully saturated. Some other examples are



ϕ3616164 ϕ3616163

[16] [17]

ϕ251C61 ϕ251C61C

[18] [19]

In example [16] the fused bond "4" is taken as being between ring atoms 4 and 5.

Many fused ring systems occur where there are rings present that are fused to rings other than the first encoded ring. In this situation the following rule applies.

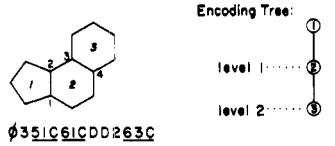
RULE 5

When, after encoding all rings directly fused to the first ring, there are still other fused rings not yet encoded this condition is signalled by an "end-of-direct-fusion" indicator "DD" followed by a numeral indicating the earliest *encoded ring that still has unencoded fused rings attached. This numeral is followed by the encoding for the rings that are directly fused to it and so on recursively.*

Rule 5 embodies the essence of the breadth-first tree enumeration mentioned earlier. The tree formalism amounts to first establishing the root (ring 1) of the tree by the rules of precedence; the branches corresponding to directly fused rings are then enumerated according to precedence—this corresponds to level 1 of the tree. Branches connected to level 1 of the tree are then enumerated in order according to the earliest branch of level 1 to which they are attached. This corresponds to encoding rings fused to other than the first ring. The process is repeated recursively until the complete structure is encoded. This mechanism can be most easily understood by referring to the encoding trees used below. The notational scheme takes the format

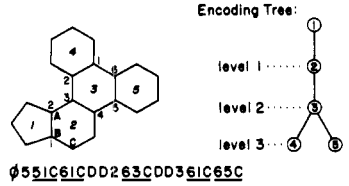
[First Ring] [Block of directly fused rings] DD [Numeral Indicating next earliest encoded ring with rings directly fused to it] [Next Block of directly fused rings] DD [Numeral... .....] ...

Two examples are



ϕ351C61CDD263C

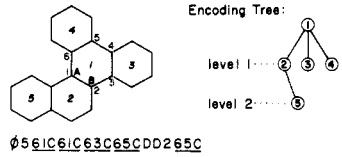
[20]



ϕ551C61CDD263CDD361C65C

[21]

If the five-membered ring in example [21] is replaced by a six-membered ring, it is interesting to see how this changes the encoding.



ϕ561C61C63C65CDD265C

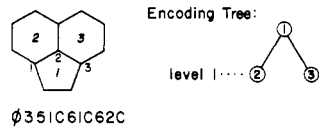
[22]

The next type of fused system to consider is that where a basis ring has two rings fused to adjacent bonds. This is referred to as a perifused system. These structures easily fit within the formalism that has been outlined so far provided the following convention is adopted.

RULE 6

Fusion is implied between any pair of rings that are fused to consecutive bonds of a basic ring.

For example, consider the structure [23] below:

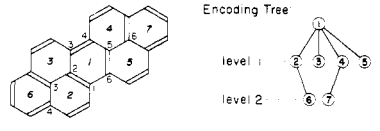


ϕ351C61C62C

[23]

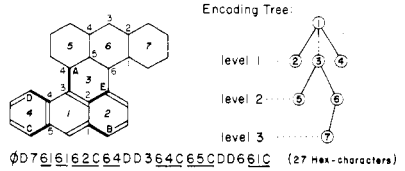
Rule 6 indicates that since rings 2 and 3 are fused to consecutive bonds of basis ring 1, they are, by implication, fused together at the common bond between them.

Some other slightly more complicated peri-fused systems are given below.



ϕD761616264650D263DD466

[24]



ϕD761616264DD364C65CDD661C (27 Hex-characters)

[25]

It is instructive to see how precedence decides which is the first ring and its numbering in example [25].

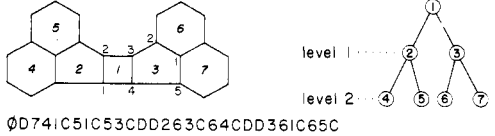
STEP 1. All the rings are six-membered and they all consist solely of carbon and so neither minimum valence nor atomic weight can play any part in deciding ring 1 (Rules A and B).

STEP 2. Rule C indicates that the first ring must come from among rings 1, 2, and 4 as numbered since these all have six atoms attached while ring 3 has nine atoms attached, and rings 5, 6, and 7 have twelve atoms attached.

STEP 3. From Rule D it is found that ring 1 has one hydrogen attached, ring 2 has three hydrogens attached, and ring 4 has four hydrogens attached and so ring 1 must be the first ring encoded.

STEP 4. The problem is now to determine the numbering of ring 1. Atoms 1 and 5 (as marked) are possible starting points (Rule D) since they both represent the start of paths that have no hydrogens directly attached until a path of six atoms has been traversed. However, when reference is made to the atoms *attached* to atoms A, B, C, D, and E, it is found that there are no hydrogens attached at E (the *second* atom on the path 1, 2, 3, . . .) whereas there is one hydrogen attached at atom D (the *second* atom on the path 5, 4, 3, . . .) and so the numbering 1, 2, 3, . . . represents the canonical path since Rule D is operative.

The numeric notation for example [25] requires 27 characters, whereas WLN uses 32 characters. In computer terms, however, the numeric representation (in this case) takes *less than half* the space for the WLN representation, because only 4 bits are needed to encode hexadecimal characters, whereas 8 bits are usually needed for each alphanumeric character.



ϕD741C51C53CDD263C64CDD361C65C

[26]

Having looked at the most common aspects of fused ring systems, it is now possible to move on to the discussion of some

other internal topological properties of rings.

4.3 Encoding Ring Systems Containing Heteroatoms, Substituents, and Unsaturated Bonds. The field specification for individual rings must be extended to represent features in addition to ring size, fusion connections, and complete saturation. This is accomplished by a set of ordered fields each terminated by an 8. The letters B, C, and BC are still used for end-of-ring termination.

The basic format for an isolated ring is

[Ring Size] 7 [Ø Heteroatoms] 8 [Substituents] 8 [Unsaturations] [End-of-ring terminator]

For fused rings the format becomes

[Ring Size] [Fusion Connection] [Ø Heteroatoms] 8 [Substituents] 8 [Unsaturations] [End-of-ring Renumerator]

The "7" is included in the isolated ring format so that its template will match that of the fused ring. The fusion connection takes the place of the 7 in the latter.

4.3A. Conventions for Heterocyclic Rings. Within the heteroatom field it is necessary to specify both the atom representations and ring positions of all heteroatoms in the ring. A ring position numeral is assigned to each heteroatom and is placed directly *after* the heteroatom. The specification of each heteroatom/position is separated by a subfield terminator 7, and by convention the heteroatom field is always begun with a zero. The reason for this will become clear when substituents are discussed.

...Ø [Representation for Heteroatom 1] [Position of Heteroatom 1] 7 [Representation of Heteroatom 2] [Position of Heteroatom 2] 7...8

Heteroatoms are encoded according to precedence (e.g., a heterocyclic oxygen is encoded before a heterocyclic nitrogen). Some simple examples are given below.

Ø1Ø67Ø31B

[27]

Ø1Ø67Ø31C

[28]

Ø1Ø67Ø31733B

[29]

Ø1Ø57ØA2321733C

[30]

Ø1Ø67Ø2173 3C

[31]

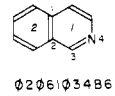
Ø1Ø57Ø217A2323C

[32]

For fused-ring heterocyclics the heteroatom encoding has the same format as for isolated rings except that here it follows directly after the specification of the fusion connection.

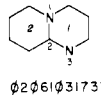
Ø2Ø61Ø33B61

[33]



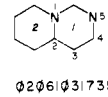
Ø2Ø61Ø34B61

[34]



Ø2Ø61Ø31733C61C

[35]

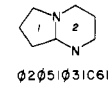


Ø2Ø61Ø31735C61C

[36]

Ø2Ø61Ø23731C61C

[37]

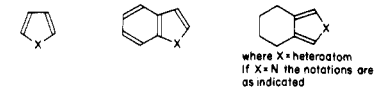


Ø2Ø51Ø31C61Ø33C

[38]

Rule 4(a), which states that the numbering for fused rings must begin at a terminally fused atom and pass initially through a fused bond, governs the ring numbering for the examples above (note in particular structure [36]). The rules of precedence clearly define which rings "shared" heteroatoms must belong to.

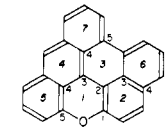
Because of the very common occurrence of heterocyclic five-membered rings of the form



Ø2Ø57Ø31B [39] Ø2Ø51Ø33B61 [40] Ø2Ø51Ø34B61C [41]

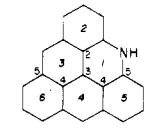
a convention is adopted whereby these systems are treated in a similar way to fully conjugated and aromatic systems in that they are terminated with a B.

Some more complicated fused rings containing heteroatoms are



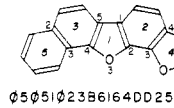
ØD761Ø26C61626364DD263DD364

[42]



Ø6Ø61Ø36C61C62C63C64CDD364C

[43]



Ø5Ø51Ø23B6164DD253Ø228DD362

[44]

4.3B. Conventions for Ring Substituents. A very important part of internal ring topology is the specification of acyclic ring substituents. Interconnected rings and chained ring systems are treated as external and so are discussed later (section 5).

Individual acyclic substituents are represented the same way as in acyclic compounds. The starting point for encoding a substituent is defined by the following rule.

RULE 7(a)

The coding for individual acyclic ring substituents must start at the atom (or double bond) connected directly to the ring and then proceed according to the rules of precedence.

The basic format in the substituent field is

...[Substituent (1)] 7 [Substituent (2)] 7...

The rules of precedence also determine the order for encoding substituents in the same way as they determine the order for encoding heteroatoms. Rule 7(b) summarizes the procedure.

RULE 7(b)

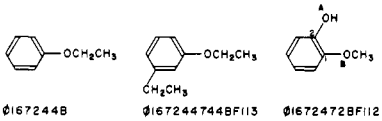
To resolve the order for encoding a set of substituents attached to a ring, the rules of precedence are applied first to the atoms attached directly to the ring. If this does not provide resolution among the substituents the rules are again applied to the atoms, one atom removed (away) from the ring and so on recursively until resolution is achieved.

Remember that in applying the rules of precedence ring hydrogens are not considered as substituents. To minimize indexing scatter in the notation the attachment positions of substituents are not encoded until *all* rings have been encoded.

The letter "F" (F for Finish) must always precede the encoding for substituent positions, ring interconnections, and bridges. The ring position template must identify both the ring to which the substituent(s) is attached and the associated ring position.

F	Ring Number of First Encoded Ring with Substituents	Ring Position of First Encoded Substituent	Ring Position of Second Encoded Substituent	...8	Ring Number of Second Encoded Ring with Substituents	...
Finish of Ring System Specification						

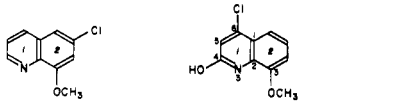
If there is only one substituent attached to a ring and all ring positions are otherwise equivalent, specification of the substituent position is ignored (see example [45] below). When relative ring positions of substituents are not known, the ring/position field is left out.



ϕI67244B ϕI67244744BFII3 ϕI6724728FII2

[45] [46] [47]

Two fused examples are



ϕ2ϕ6IϕЗЗB6IAI35724BF253 ϕ2ϕ6IϕЗЗ8AI3572B6I24BFI64823

[48] [49]

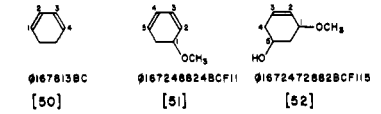
The code . . .F164823 in example [49] indicates that on ring 1 the first encoded substituent is at position 6, and the second encoded substituent is at position 4. On ring 2 the substituent is at position 3. The ring positions of substituents in examples [48] and [49] are encoded in the *same* order as the substituents have been encoded.

The handling of multiple occurrences of a substituent on a ring is very straightforward and is included in the discussion on multiplier conventions.[7]

4.3C. Encoding Partially Saturated Rings. To complete the major conventions for internal ring topology it is necessary to consider the specification of partially saturated rings. The format employed is

	Position of the earliest encoded unsaturated bond	Position of the next encoded unsaturated bond	
...8			...BC

Ring unsaturation is always encoded after heteroatoms and substituents in accordance with precedence. Remember that partially saturated rings terminate with a "BC".



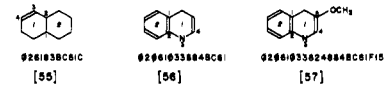
ϕI67813BC ϕI67248824BCFII ϕI6724728828CFII5

[50] [51] [52]

ϕIϕ67ϕ3I6248835BCFI2 ϕIϕ67ϕ3I824CFI2

[53] [54]

The leading "8" in the unsaturation field ensures that unsaturations can be differentiated from substituents. In example [52] valence determines ring position 1 but the unsaturation determines the path taken to the OH group. Partially saturated fused systems are handled similarly.



ϕ26I83BC8IC ϕ2ϕ6Iϕ33884BC8I ϕ2ϕ6Iϕ3382484BC8IFI5

[55] [56] [57]

5. INTERCONNECTED RING SYSTEMS - EXTERNAL TOPOLOGY

Thus far only isolated rings and fused ring systems have been considered. These structures form the basis for specification of interconnected ring systems. However, it is necessary to explore the various ways for linking ring systems together.

5.1 Single Bond Links between Ring Systems. The simplest type of bond between ring systems is a single bond without any intervening acyclic substructures. The general interconnection format for this case is

Ring System (1)	EEE...F	Ring System (2)	F	Ring System (3)	F	Ring System (4)	...

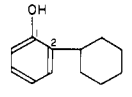
In the case when precedence allows a chain of rings to be formed or there are just two ring systems involved, the format simplifies to

Ring System (1)	EF	Ring System (2)	EF	Ring System (3)	EF...

The "EF" indicates that there is a single bond link between two ring systems. This represents a *general* connection between ring systems with no reference to actual positions of connection.

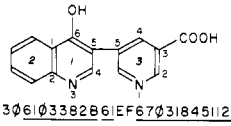
The specific positions for inter-ring connections are expressed only *after* the substituent positions for all rings have been encoded. *The order for encoding inter-ring connections coincides with the order in which the rings have been encoded.* To mark where the encoding for substituent positions ends and the inter-ring position information begins, an extra "8" leads the inter-ring information. The format is

...8	Ring Number of First Inter-connected Ring	Position of Inter-connection at First Ring	8	Ring Number of Ring connected to First Ring	Position of Inter-connection at Second Ring	...	9 End of Branch or Ring Chain	...



ϕ2672BEF6CFII88I282I

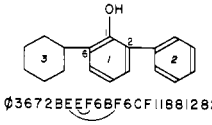
[58]



ϕ3ϕ6IϕЗЗ82B6IEF67ϕ3I8451I292BFI6833B8I5835

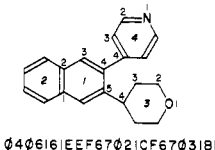
[59]

Other examples are



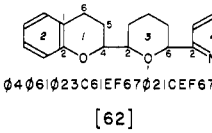
ϕ3672BEEF6BF6CFII88I282I9I683I

[60]



ϕ4ϕ6I6IEEF67ϕ2ICF67ϕ3IBF8I5834914844

[61]



ϕ4ϕ6IϕЗЗC6IEF67ϕ2ICEF67ϕ3IBF8I48326842

[62]

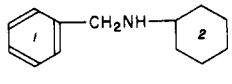
5.2 Ring Systems Linked by Acyclic Chains. The . . .EF. . . convention for representing single bond inter-ring connections is easily extended to cover the case of ring systems linked by acyclic chains. The "gap" between the E and the

Table II. Glossary of Representation Symbols

REPRESENTATION	MEANING
5. RING TERMINATORS, ETC.	
Ø	All rings begin with zero.
B	Terminator fully conjugated (including aromatic) rings. Is neglected for fused rings with no substituents, and no heteroatoms.
C	Terminates all saturated rings.
BC	Terminates all partially saturated rings.
F	End-of-ring specification must be present if there are any substituents, or inter-ring or bridge connections.
7	Sub-field terminator (also "filler" for isolated rings).
8	Field terminator.
9	End-of-positional specifications for bridge, chain of rings or inter-ring connection.
DD	End-of-direct-fusion indicator.
DF	End-of-fusion indicator where no implied fusions.
BB	Start of positional bridging specifications.
6. RING INTERCONNECTIONS	
...EF...	Single bond ring interconnection.
...E[Acyclic Substructure]F...	Rings interconnected by acyclic substructure.
...EØF...	Spiro inter-ring connection.
...EBF...	Single bond bridging interconnection.
...EB[Acyclic Substructure]F...	Ring(s) interconnected by an acyclic bridge.
...ECCF...	Catenene ring interconnection.
7. MULTIPLIER CONVENTIONS	
Ø[Number of Occurrences]	Multiplier indicated by a zero followed by number of occurrences.
ØD[Number of Occurrences]	Number of occurrences greater than 6.
...EØ[Number of Occurrences]F...	Ring-chain multiplication.
...EØ[Number of Occurrences]ØF...	Spiro ring-chain multiplication.

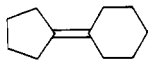
F is used to hold the acyclic connector, and so the general form is

...Ring System] E [Acyclic Structure] F [Ring System...



Ø26BE43F6C

[63]



Ø25CE11F6C

[64]

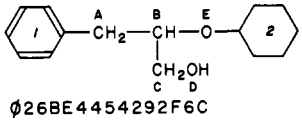
When all ring positions are equivalent for all rings attached to the acyclic connector it is not necessary to specify the connections.

If the acyclic substructure linking ring systems exhibits branching the following convention applies.

RULE 8

The encoding for the an acyclic substructure linking two or more ring systems must begin at the atom directly attached to the first encoded ring. Encoding then proceeds according to the rules of precedence with the overriding restriction that acyclic branches that are not directly involved in the inter-ring connections take precedence.

In example [65] Rule 8 is operative.



Ø26BE4454292F6C

[65]

5.3 Encoding Ring Systems with Spiro Atoms. A rather more unusual inter-ring connection is that involving a spiro atom. For spiro links the . . .EF. . . single bond convention is replaced by . . . EØF. . . *Rings that are bonded by a spiro connection are treated similarly to rings linked by single bonds*

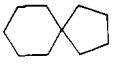
in that they are considered as isolated for the purposes of determining their individual encodings. The general format is

[Ring System (1)] EEE ... ØF [Ring System (2)] ØF [Ring System (3)] ØF [Ring System (4)] ...

Where there is a chain of spiro-connected rings or just two ring systems the format simplifies to

[Ring System (1)] EØF [Ring System (2)] EØF [Ring System (3)] EØF...

An example is



ϕ25CEØF6C

[66]

6. ENCODING RING SYSTEMS WITH BRIDGES

At a first glance it would seem that a procedure for handling the variety and complexity of bridging topology may be cumbersome. Fortunately by formalizing the definition of what constitutes a ring it is very easy to extend the methodology so far enumerated to incorporate bridging topology.

RECURSIVE RING DEFINITION:

"A ring is the shortest closed path which shares at most one bond with any other ring."

This definition together with the rules of precedence enables identification both of the ring systems and bridging components in a molecular structure. The actual enumeration of all the rings in a bridged structure is achieved by applying the definition recursively to closed paths attached to previously encoded rings. The application of this principle to fused rings with bridges will be discussed in the next section. Once a closed path has been established which satisfies both the ring definition and precedence, then any other attached closed paths which share more than one bond with the ring are considered to be part of a bridging system.

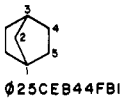
Actual bridging components are not encoded until all rings have been specified. The atoms in a bridging substructure are encoded via a connected path according to the rules of precedence. As for rings connected by branched acyclic substructures, the branches of bridges with terminal atoms are encoded first so as to ensure there is no ambiguity in deciding which bridging atoms form connections to rings.

If there is more than one bridge present the order for encoding the bridges is again resolved by the rules of precedence. To bring the ring count into line with other notations and the Ring Index, it is set equal to the number of rings plus bridge contributions. *A bridge with N ring connections contributes the equivalent of (N − 1) rings to the ring count.* Most bridges have just two ring connections and so they contribute one ring. Heteroatoms in bridges are treated like heteroatoms in rings in that they require a zero to be encoded in the notation directly after ring count. The general format for parenthesizing bridge fragments is

... EB[Acyclic Bridge] F .... BB

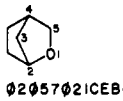
This reduces to . . .EBF. . . if the bridge is just a bond instead of an acyclic bridge. The "BB" rather than the "88" is used to indicate the start of "ring number-ring position" specifications for bridges. It reduces to . . .FB. . . if there are no substituents and no nonbridging acyclic connections. *Positional encoding follows the order of encoding of the corresponding bridge atoms. Bridging positions are only specified after substituents and inter-ring connections.*

6.1 Bridging Systems with No Fused Bonds. The simplest bridging systems are those that involve no "fused" bonds.



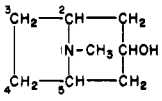
ϕ25CEB44FB113

[67]



ϕ2ϕ57ϕ21CEB44FB124

[68]



ϕ2ϕ57ϕ3184CEB445294F11BB125

[69]

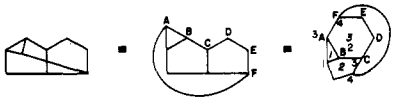
ϕ34CEB44FEB44FB1139113

[70]

Applying the ring definition to example [68] indicates that a choice must be made between two five-membered rings. Precedence confirms that the ring containing a heteroatom must be the first choice. This leaves a $-CH_2-CH_2-$ bridge. Ring numbering follows the path of minimum valence and so begins at the heteroatom. The code . . .FB124 indicates that ring 1 has a bridge connected to atoms 2 and 4. The "9" in . . .FB1139113 of structure [70] denotes the end of positional specification for a bridge.

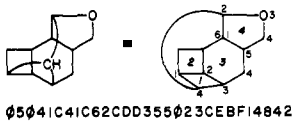
6.2 Fused Ring Systems with Bridges. *Fused ring systems that possess bridges are handled similarly. The important thing is to begin by correctly identifying the first ring to be encoded. Care must then be taken to enumerate all directly attached fused rings before proceeding to identify bridges. The critical step is to rigidly apply the definition of a ring in deciding the fused rings that are attached to the first ring.*

Example [71] drawn in three different forms clearly illustrates this point. The encoding is most easily seen to follow from the third representation (right-hand side). Since the bridge is just a single bond between two ring atoms, the bridge indicator reduces to . . .EBF. . ..



ϕ431C51C62CEBFB24834

[71]



ϕ5ϕ41C41C62CDD355ϕ23CEBF14842

[72]

A QUALITATIVE COMPARISON WITH WISWESSER LINE NOTATION

The formalism described represents an attempt to develop a system for the linearly encoding of molecular topology that is simpler than existing systems (in particular, WLN). It is difficult to give quantitative evidence to measure any simplification that has been made relative to other notations. However, there are a number of qualitative factors which suggest that considerable simplification may have been achieved. The first is that since the formalism is underpinned by the most fundamental concepts of valence it might be reasonable to expect that considerable systematization and simplification follow naturally in any encoding effort, either manual or automatic.

The decision to maintain individual ring identities which can accomodate substituents and heteroatoms, etc., would appear to be simpler and more natural than separating such components. Viewing complex fused ring systems as a "tree-of rings" also seems easier for encoding purposes than mapping out peripheral paths.

It might be said in very general terms that WLN is based on an encoding formalism that at least, in part, is a function of the characteristics of the notation itself. In the present system the encoding formalism is *independent* of the notation. Valence establishes the canonical notation for each structure. In contrast, in WLN, the principle of latest position is applied to achieve canonical notations. The former should represent a simplification of the encoding procedures because of the direct dependence on structure rather than notation conventions. Requiring that the encoding depend on the notation must, by its very nature, introduce an additional, and seemingly unnecessary, level of complexity.

CONCLUSIONS

The aim of the present work has been to develop a simple chemically guided (as opposed to a notation guided) encoding formalism for representing molecular topology. The major design constraints have been that the system should be easy to use both manually and by computer. To remove uncertainty and ambiguity, emphasis has been placed upon the consistent application of very basic principles and strategies. At the expense of slightly longer notations in some instances, complete atom-by-atom descriptions have been used for other than ring representation. Consequently any judgment as to what configurations of atoms or functionalities are important has been avoided. To counter any additional computer storage requirements that might stem from this atom-by-atom approach, a 4-bit hexadecimal character code has been employed in preference to the 8-bit codes of other notations. It would certainly be very easy to adapt a 40-character alphabet to the present notation. The convention of dealing with the general aspects of connectivity prior to the specific connections ensures that the system has good indexing properties.

A detailed treatment of more unusual aspects of molecular topology is given elsewhere.[7] As with all formalisms many value-judgments have been made both with respect to representation and encoding strategy. In carrying out these judgments the aim has always been to make the system as simple and easy to work with as possible because, almost invariably, the usefulness of a system deteriorates as it increases in complexity.

With the amount of effort and resources that has been committed to other notations and connection table representations, it would seem highly desirable (if the present system were to play a useful role in future chemical information systems) that a straightforward method exists for conversion to and from other representations. To this end a new canonical connection table formalism that is directly compatible with the present notation has been developed. This formalism is discussed in detail in a separate publication.[8] The simplicity of the encoding rules should make the task of automatic derivation of the notation from connection tables much easier than in systems like Wiswesser line notation where the complexity of the rules has made this mode of conversion very difficult.[1,9]

ACKNOWLEDGMENT

I would sincerely like to thank Dr. J. K. MacLeod for his encouragement, his painstaking reading of the manuscript, and his helpful suggestions and criticisms. Drs. J. Christie and J. Traeger have also made helpful suggestions. I would like to thank Carol Jacobs, Betty Moore, Greta Pribyl, Byam Wight, Lorraine Scarr, and Gary Brown for their patience and assistance in preparing the figures and the manuscript.

REFERENCES AND NOTES

(1) E. J. Smith, "W.J. Wiswesser's Line Formula Chemical Notation", McGraw-Hill, New York, N.Y., 1968.
(2) J. Lederberg, NASA Report, N65-13150, 1964.
(3) H. W. Hayward, Patent Office Research and Development Report, No. 21, Patent Office, Washington, D.C., 1964.
(4) G. M. Dyson, M. F. Lynch, H. L. Morgan, *Inf. Storage Retr.*, **4**, 27 (1968).
(5) H. Skolnik, "A New Linear Notation Based on Combination of Carbon and Hydrogen", *J. Chem. Doc.*, **6**, 689 (1969).
(6) H. Skolnik, "A Chemical Fragment Notation Index", *J. Chem. Doc.*, **11**, 142 (1971).
(7) R. G. Dromey, University of Wollongong, Computing Science Department, Technical Report, CS78-1 (1978) (copies are available from the author on request).
(8) R. G. Dromey, *J. Chem. Inf. Comput. Sci.*, in press.
(9) M. F. Lynch, J. M. Harrison, W. G. Town, and J. E. Ash, "Computer Handling of Chemical Structure Information", MacDonald, London, 1971.