

A New Linear Notation System Based on Combinations of Carbon and Hydrogen

Herman Skolnik

Hercules Research Center, Hercules Incorporated

Dedicated to Professor Allan R. Day

Contribution Number 1492

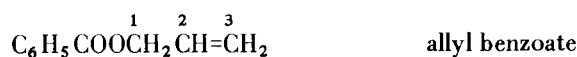
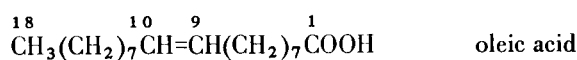
A new linear notation system is described which was designed to have a one-to-one correspondence with the chemical structures represented by the notations. Each notation is unique and unambiguous, yet simple and compatible with computer input and output characters. The symbols used in the notation system denote carbon in terms of bonds and attached hydrogen(s). The notation system is illustrated with acyclic, carbocyclic, and heterocyclic examples.

Although chemistry is outstanding among the sciences in its literature resource, there are many challenges yet to be solved in our communication and information retrieval mechanisms. For example, in conducting the initial literature searching for five- and six-membered heterocyclics containing oxygen and sulfur and two or more sulfurs (1), the authors were grateful for "The Ring Index" (2) and for the "Index of Ring Systems" in *Chemical Abstracts*. Yet beyond the one-ring system, almost every page of "The Ring Index" and every entry under the subjects in the *Chemical Abstracts* Index referred to by the "Index of Ring Systems" had to be scanned. This was a time-consuming, monotonous task.

As the authors delved into the original literature, they encountered many nomenclature problems. They also were confronted with nomenclature problems during the indexing of their book. These problems would have been considerably less burdensome if the heterocyclics they were interested in had been stored in a computer system by some reliable retrieval mechanism.

Ideally a computerized retrieval mechanism for chemical structures should be compatible with accepted practices in the way structural formulas are drawn and in the numbering of chain and ring moieties. It should also be compatible with keypunch input to computers and with the characters of computer print chains. Although many notation systems described in the literature are more or less computer-compatible, their topological delineations are not in harmony with the way structural formulas are drawn or with the accepted numbering, particularly of ring systems. This criterion of compatibility is an important one as there is a high degree of communication in the structural formulas of ring systems. One needs only to examine the cover of *This Journal* to appreciate how much information is transmitted by the structural formula.

In drawing structural formulas, the path is generally from the left to right, but functional groups are assigned the right-end position or as near to the right as possible in acyclics or to the top or right in rings. The numbering, however, generally proceeds from the right, starting with the carbon to which the functional group is attached or is a part of; numbering of the atoms in attachments, however, generally proceeds from the atom attached to the primary moiety. For examples, consider the following compounds:



Fortunately, "The Ring Index" is the bible for ring structures and their numberings. It is oriented to unsaturated ring structures, denoting only saturated atoms with hydrogen.

Basic Notation Principles.

Because structural formulas, as written by chemists, denote combinations of carbon and hydrogen, such as $-\text{CH}_3$, $-\text{CH}_2-$, $>\text{CH}-$, and $>\text{C}<$, it appeared to be logical to base the new notation system on these structural units. Furthermore, the ever increasing importance of nmr in structural and mechanistic studies gives added importance to carbon/hydrogen units.

Table I lists the symbols used in the new notation system for the various combinations of carbon and hydrogen that occur in structural formulas. Carbonyl is included in this group because of its wide occurrence, particularly in ring systems. Table II lists the notations for elements other than carbon.

TABLE I

Notation Symbols for Combinations of Carbon and Hydrogen and for Carbonyl

Single-bonded Carbons		Double-bonded Carbons	
-CH ₃	A	=CH ₂	E
-CH ₂ -	C	=CH-	B
$\begin{array}{c} \\ -\text{CH}- \\ \end{array}$	Y	>C=	D
$\begin{array}{c} \\ -\text{C}- \\ \end{array}$	X	>C=	R (fused >C=)
>CH-	J (fused or bridgehead >CH-)		
>C<	T (fused or bridgehead >C<)		
Triple-bonded Carbons		Carbonyl	
≡CH-	U	>C=O	K
≡C-	V		

TABLE II

Notation Symbols for Atoms Other Than Carbon

Halogen		Nitrogen		Oxygen	
-F	F	>NH	M	-O-	Q
-Br	G	-NH ₂	MH	-OH	QH
-I	I	>N-	N	O ₂	W (as in NO ₂ or SO ₂)
-Cl	L	≡N or -N=	Z		
Other Alphabetical Symbols		Character Symbols			
-H	H	Fused or bridgehead atom (other than C) is denoted by * following symbol.			
-S- or =S	S	Ionic form is denoted by # following symbol.			
P	P	Substituent between bridgeheads is denoted by : Metal = & + Atomic Symbol, e.g., &NA for sodium.			

In representing a chemical structure, the notations are written as they would be drawn. The numbering system is that of *Chemical Abstracts* or "The Ring Index". Other principles will be noted in discussion of the examples that follow.

Acyclic Hydrocarbons.

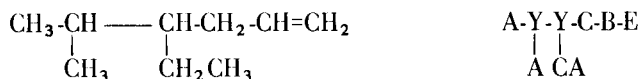
The following examples illustrate how the notation system is applied to acyclic saturated hydrocarbons:

Compound	Visual Notation	Written Notation
1. CH ₄		A.H
2. CH ₃ CH ₃	AA	A2
3. CH ₃ CH ₂ CH ₃	ACA	A2C
4. CH ₃ CH ₂ CH ₂ CH ₃	AC ₂ A	AC2A
5. $\begin{array}{c} \text{CH}_3 \\ \diagdown \\ \text{CHCH}_3 \\ \diagup \\ \text{CH}_3 \end{array}$	$\begin{array}{c} \text{A} \\ \diagdown \\ \text{Y} \\ \diagup \\ \text{A} \end{array}$	AYA.A (preferred) A3Y
6. CH ₃ (CH ₂) ₃ CH ₃	AC ₃ A	AC3A
7. $\begin{array}{c} \text{CH}_3 \\ \diagdown \\ \text{CHCH}_2\text{CH}_3 \\ \diagup \\ \text{CH}_3 \end{array}$	$\begin{array}{c} \text{A} \\ \diagdown \\ \text{Y} \\ \diagup \\ \text{A} \end{array}$	AYCA.A (preferred) A2YCA
8. (CH ₃) ₃ CCH ₃	A ₃ XA	AXA.A2 (preferred) A4X

In the case of isobutane, isopentane, and neopentane, the preferred notations are: AYA.A, AYCA.A, and AXA.A2, respectively. These conform to IUPAC rules and relate most directly to functional derivatives. The following cases illustrate notations for unsaturated acyclics:

CH ₂ =CH ₂	EE	or E2
CH ₃ CH=CH ₂		ABE
CH ₃ CH ₂ CH=CH ₂		ACBE
(CH ₃) ₂ C=CH ₂	$\begin{array}{c} \text{A} \\ \diagdown \\ \text{DE} \\ \diagup \\ \text{A} \end{array}$	ADE.A
CH ₂ =CHCH=CH ₂	EBBE	EB2E
CH ₂ =CHC(CH ₃)=CH ₂	$\begin{array}{c} \text{EBDE} \\ \\ \text{A} \end{array}$	EBDE.A
CH ₂ =C(CH ₃)=C(CH ₃)=CH ₂	$\begin{array}{c} \text{EDDE} \\ \\ \text{AA} \end{array}$	EDDE.A.A or ED2E.A2 (preferred)

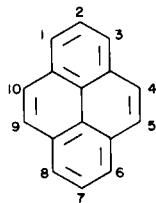
To illustrate the correspondence between the new notation system and the normal writing and numbering of a structure, let us consider 4-ethyl-5-methyl-1-hexene:



Notation: AYYCBE.CA.A



It is to be noted that for 5-ethyl-4-methyl-1-hexene, nomenclature rules require the citing of substituents in alphabetical order. This rule is ignored in the notation system, as substituents are more appropriately positioned



B2R4=B3R=B2R=B3
(same as above)

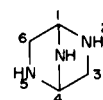
Heterocyclics.

In the following examples, which illustrate the new notation system for heterocyclics, an asterisk following a symbol denotes a fused ring atom other than carbon (R):

Ring Index No.	Compound	Notation
124	 Pyrazole	B3ZM.
125	 3H-Pyrazole	B2CZ2.
127	 Imidazole	B2ZBM.
128	 2H-Imidazole	B2ZCZ.
881	 2H-Imidazo[4,5-d]oxazole	^{6 5 4} MBZR] = ^{3 2 1} ZCQ.
883	 1H,3H-Imidazo[1,5-c]oxazole	^{7 6 5 4} BZBN *R = ^{3 2 1} CQC.
918	 Imidazo[4,5-d]imidazole	ZBZR2=ZBZ.
919	 1H-Imidazo[1,5-a]imidazole	^{7 6 5 4} BZBN *R = ^{3 2 1} B2M.
923	 1,4-Diazabicyclo[2.2.1]heptane	^{6 5 4} C2N *C2N *C. ^{3 2 1} 7

925

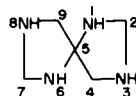
Trimidine



^{6 5 4 3 2 1} 7
CMJCM] :M.

927

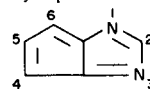
1,3,6,8-Tetraazaspiro[4.4]nonane



^{9 8 7 6 5 4 3 2 1}
C*MC*M=T=C*MC*M.
(π Tπ = spiro C)

948

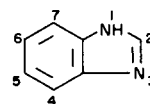
Cyclopentimidazole



B3R2=ZBZ.

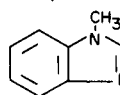
1213

Benzimidazole



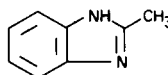
^{7 4} B4R2=ZBM.

1-Methylbenzimidazole



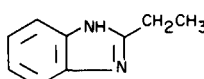
B4R2=ZBN.A

2-Methylbenzimidazole



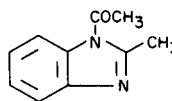
B4R2=ZDM.A

2-Ethylbenzimidazole



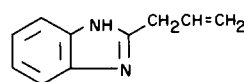
B4R2=ZDM.CA

1-Acetyl-2-methylbenzimidazole



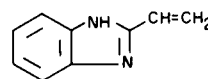
B4R2=ZDN.KA.A

2-Allylbenzimidazole

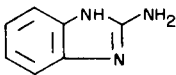
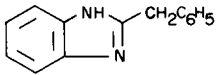
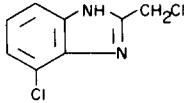
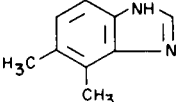
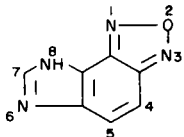
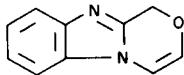
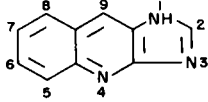
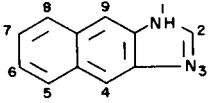
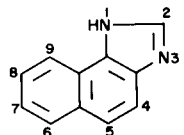


B3R2=ZDM.CBE

2-Vinylbenzimidazole



B4R2=ZDM.BE

	2-Aminobenzimidazole	B4R2=ZDM.MH
		
	2-Benzylbenzimidazole	B4R2=ZDM.C/DB5/
		
	4-Chloro-2-chloromethylbenzimidazole	B3DR2=ZDM.CL.L
		
	5,6-Dimethylbenzimidazole	B2D2R2=ZBM.A2
		
2282	8H-Furazano[4,5-e]benzimidazole	MBZR2=B2R2=ZQZ.
		
2730	1H-[1,4]-Oxazino[4,3-a]benzimidazole	B4R2=N*R.Z=B2QC.
		
2754	1H-Imidazo[4,5-b]quinoline	B4R2=ZR2B=ZBM.
		
2785	1H-Naphth[2,3-d]imidazole	B4R2=BR2B=ZBM.
		
2788	1H-Naphth[1,2-d]imidazole	B4R2=B2R2=ZBM.
		

Discussion.

Chemical notation systems have been with us since the 1940's (3, 6), yet none really has caught the fancy of bench chemists. Nevertheless, the Dyson (4) and the Wiswesser (5) systems have experienced some acceptance by literature chemists for chemical indexes. This acceptance, even though limited, indicates that linear notation systems can contribute to the solution of problems in nomenclature, indexing, and communication vis-a-vis the computer.

The basic weakness of notation systems has been the complexity of the many rules invoked to ensure that a unique and unambiguous notation is assigned to a given structure. These rules have forced upon the potential users the need to learn a different way of visualizing chemical structures and of numbering chains and rings. In designing the notation system described in this paper, the primary objectives were to retain accepted numbering schemes, to have the notations conform to accepted chemical structures, and to invoke relatively few rules. These three objectives have been achieved.

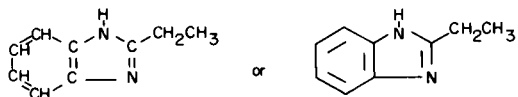
Nine notation symbols define the two important parameters associated with carbon in organic chemical structures, viz., the carbon bonding and the number of hydrogens attached to the carbon. These nine notation symbols are: A (-CH₃), B (-CH=), C (-CH₂-), D (>C=), E (=CH₂), U (≡CH), V (≡C-), X (>C<), and Y (>CH-). An additional three notation symbols are used to define carbon in condensed rings or as bridgeheads: J (>CH-), R (>C=), and T (>C<); the symbol \square indicates a condensed ring situation, and $\square T \square$ a spiro carbon; the punctuation symbol : indicates a substituent between bridgehead atoms. The notation symbol K is used to describe the carbonyl function.

Denoting carbon in terms of bonds and hydrogens is a logical approach as these units comprise the major part of organic chemicals. Because other systems tend to treat carbon as acyclic and carbocyclic, they must use symbols to indicate carbon bonding other than single bonding (although some systems, such as atom connection tables, also indicate single bonding). Because the bonding in the new system is included in the symbol, molecular formulas and molecular weights can be calculated easily with a computer and given in the printout along with other information.

Because the notation system described in this paper conforms to practice, visual inspection of a notation reveals more structural information than do other notation systems. With a little experience, one soon is able to project the linear notation into a two dimensional structure, pretty much in the following manner for 2-ethylbenzimidazole:

B4R2=ZDM.CA

B	M	
B	R	D.CA
B	R	Z
B		



This one-to-one correspondence between a notation system and the structure it represents is a major advantage. At the same time, the notation is unique and unambiguous as a communication. The symbols used are simple, can be inputted by keypunch or typewriter, and are on most computer print chains. It is particularly suitable for computer processing, and requires relatively simple programming for manipulating in the computer from various viewpoints. Its use in computer operations, however, is a subject for a subsequent paper.

REFERENCES

- (1) D. S. Breslow and H. Skolnik, "Multi-Sulfur and Sulfur and Oxygen Five- and Six-Membered Heterocyclics", Volume 21 (in two parts) of "The Chemistry of Heterocyclic Compounds", John Wiley & Sons, Inc., 1966.
- (2) A. M. Patterson, L. T. Capell, and D. F. Walker, "The Ring Index", 2nd Edition, American Chemical Society, 1960.
- (3) "Survey of Chemical Notation Systems", Publication 1150, National Academy of Sciences/National Research Council, 1964.
- (4) "Rules for I.U.P.A.C. Notation for Organic Compounds", John Wiley & Sons, Inc., 1961.
- (5) E. G. Smith, "The Wiswesser Line-Formula Chemical Notation", McGraw-Hill Book Co., 1968.
- (6) Other references to notation systems are:
 - (a) M. M. Berry and J. W. Perry, *Chem. Eng. News*, 30, 407 (1952).
 - (b) R. Fugmann, *Nachr. Dokumentation*, 12, 69 (1961).
 - (c) W. Gruber, *Angew. Chem.*, 61, 429 (1949).
 - (d) H. W. Hayward, Patent Office Research and Development Reports, No. 21, 1961.
 - (e) J. A. Silk, *J. Chem. Doc.*, 1, 58 (1961).
 - (f) H. Skolnik and A. Clow, *ibid.*, 4, 221 (1964).

Received July 30, 1969

Wilmington, Delaware 19899