Code for Chemical Ring Compounds with Application of Fusion Lines, Suited for Calculation of Their Physical Properties*

MICHAEL JOHN ROMANEC**
Moore School, University of Pennsylvania, Philadelphia, Pa. 19104

Received November 5, 1973

The boundary between two neighboring rings is called the fusion lines, which is enclosed by external segments called arcs. In the code, the separation of fusion line(s) from arcs takes place in such a manner that proper matching is achieved. By this procedure, the structure of a compound is recreated.

The development of automated processing of information has created an acute need for topological codes of chemical structures. While they must describe chemical structures, these codes should also be developed with their prospective uses in mind.

The code presented here has, as its main purpose, the computer-oriented calculation of physical properties of molecular species by group contribution methods, such as the method of Lydersen.[1] The principle of group contributions states that certain thermodynamical properties such as critical pressure, volume, and temperature are additive and can be calculated by summing the contributions of the groups (segments) that make up a particular chemical compound.[1,2]

Groups (segments), which Lydersen calls increments, are those atoms and collections of atoms which may be considered as building blocks of chemical compounds. For example, owing to its valency of four, the carbon atom can exist in the following segments: $-CH_3$, $-CH_2-$, $=CH_2$, $-\overset{|}{C}H-$, $-\overset{|}{\underset{|}{C}}-$, $=\overset{|}{C}-$, $=C=$, $-C\equiv$, and as a carbon with a charge. These segments can be present in a chain or a ring; and, of course, the hetero-atoms would have their segments too. Consequently, in order to calculate the physical properties of any compound, the Lydersen's increments must be represented in the code.

By substituting a specific value (from table) for each group (segment) in an equation, the particular property can be evaluated. Those tables and equations, as well as examples, are given in Reid's book.[1] Such calculations are useful as methods of estimating unknown properties of compounds. According to our alphabet, the code contains symbols for all Lydersen's increments of the ring compounds, so the basic thermodynamical properties (critical pressure, volume, and temperature) can be calculated straight from the code. No connectivity tables are required.

The codes now in use can be classified according to their purpose as follows:

(a) To furnish the shortest code possible[3,4,5]
(b) To be suitable for specialized classes of compounds and special purposes[6]
(c) To describe two- and three-dimensional structures through the use of matrices[2,7,8]
(d) To denote the carbon and hetero atoms of the structure in terms of bonds and attached hydrogen(s)[9]

THE PROPOSED CODE

For ring compounds a new code is presented, which emphasizes two different approaches in handling mono- and multi-ring compounds.

(a) The mono-ring compounds are encoded by starting with carbon- or hetero-atom No. 1 and proceeding around the ring in the direction of increasing enumeration (double bond insertions will be discussed later)

(b) Two-ring and multi-ring compounds are very complicated, and therefore their encoding is described in general terms as follows:

Two-ring and multi-ring compounds comprise two principal parts: internal fusion lines and external arcs, around the fusion lines.

By separating these two parts and simultaneously matching them together the structure of a chemical compound can be recreated—which is just what the proposed code system accomplishes. Or schematically: structure of a compound = fusion line(s) * arcs, where symbol * stands for the operation of matching and separating fusion line(s) from arcs. With naphthalene as an example we get:

or in code notation, it reads:

Naphthalene = V J UU / UU

Clockwise enumeration of atoms . . . . 1,2,3,4 and 5,6,7,8

For naphthalene, anthracene, and acenaphthene, the ratios of fusion line carbon atoms to total carbon atoms

* This code was developed in connection with Master-Thesis work in Computer and Information Science and presented at the Meeting of the Chemical Engineering Computing System Group, April 5, 1972. University of Pennsylvania, Philadelphia, Pa, 19104.

** Present address: Institute for Scientific Information, 325 Chestnut St., Philadelphia, Pa. 19106.

present are 20%, 28.5%, and 33%, respectively. Evidently, the percentage increases for more complex structures. Thus simplification of the code is achieved by separation of fusion lines, otherwise the fusion line atoms will seem to be among the arcs, which would obscure the code representation of the structure. On the other hand, the separated fusion lines and spiro atoms give an indication of the number of rings present, which renders the drawing of a structure simple and straightforward.

The proposed code follows the orientation and enumeration as given in the "Ring Index."[10] In most cases, this permits arrangment of enumerated atoms in sequential order, and in some cases, in nearly sequential order. The sequence can be easily followed because the substituents on the ring will be furnished with positional numerals. The accepted orientation as found in the "Ring Index" results in a unique representation of fusion lines and arcs, thus contributing to the canonicity of the code.

Alphabet.

, Comma after a numeral indicates the position of a substituent on the ring. The numeral with a comma (e.g., 3,) is called a positional numeral. The length of a substituent should not exceed 10 code symbols. The positional numeral "1" is omitted.

: Semicolon after a positional numeral (e.g., 2,:) means that the substituent on the arc exceeds the length of 10 symbols or is a ring, and will be listed after the arcs section or after the crossing pseudo bond if the later is present. The long substituent will be prefaced by a semicolon followed by a positional numeral. The same applies to mono-ring compounds.

— Dash indicates a single horizontal bond to or from the top of fusion line V, VH, or elongated VH, or between V, VH, and E, or between two E's or two elongated EH's by themselves. After a positional numeral and followed by another positional numeral, the dash indicates the position of a single bond linkage between two rings. If a dash symbol is replaced by a double bond, it indicates double bond linkage between the two rings.

ø Zero symbol means zero or a single horizontal bond to or from the bottom of fusion line V, VH, or elongated VH.

= Number sign means a single horizontal bond to or from A, AH, D, and DH.

& Ampersand indicates a single bond to or from Y,X,N of elongated fusion lines. If the single bond goes, e.g., from the top of preceding V to Y of next VlYH, then the code reads : V-&VlYH.
Underlined atoms (Y,X,N, etc.) indicate the crossing pseudo bond between them. In this code, the bond crossing another bond is regarded as a pseudo bond. On the other hand, the single pseudo bond confined to a ring is represented as a fusion line (e.g., bicyclic compounds).

' Prime after a fusion line indicates the beginning of enumeration according to *Chemical Abstracts* and the beginning of arc codes matching the fusion line section. Ordinarily, the enumeration proceeds clockwise. For two-ring compounds, the single fusion line does not bear a prime symbol.

'' Double prime, bears the same meaning as a prime, except that enumeration proceeds counter clockwise, as with steroids.

" " Quotation marks designate two-letter elements other than B,F,G,H,I,M,N,O,P,S,Z. The following one-letter elements are designated differently:

potassium "PO" vanadium "VA"

uranium "UR" yttrium "YT"

/ Slash separates each particular arc, fusion line, or pseudo bond from long substituent.

| Vertical line means a single vertical bond connecting parallel equatorial fusion lines.

( ) Parentheses enclose the substituent on the fusion line and in the fusion line section, as well as on the pseudo bond atoms.

$ Dollar sign indicates that the last available atom in closing a ring should be disregarded.

* Star (asterisk) indicates closure of the ring (mono-ring or fusion lines ring).

= Double bond written out, also double bond linkage for two rings.

R Phenyl radical and fusion lines forming a benzene ring (see poly-fused aromatic ring compounds).

1. .2. .3 etc. Arabic numerals for —CH_2—, —CH_2—, CH_2—, etc.; with a comma (2, 3, 4,), they indicate positional numerals of substituents.

2',3', etc. Primed and double primed arabic numerals with commas,

2'',3'', etc. indicate positional numerals, as in dyes.

+ An atom if superscripted by plus symbol, means an onium atom.

A Arm (branch) of fusion line (see fusion line codes).

B Boron

C Carbon atom without hydrogen, double or triple bond as indicated.

D Arm (branch) of fusion line (see fusion line codes).

E Equatorial fusion line, (see fusion line codes) (owing to the benzene ring (how it is drawn), the equatorial fusion line could deviate by 30° from horizontality)

F Fluorine

G Chlorine

H Hydrogen, in saturated fusion lines, in terminal alkyl groups or for aldehyde, e.g. methyl 1H, ethyl 2H, etc., aldehyde KH.

I Iodine

J Terminal symbol for fusion lines and arcs.

K Carbonyl group as in ketone, in aliphatics or on the ring.

L Left orientation of double bond (see fusion line codes)

M NH

N Nitrogen free of hydrogen.

O Oxygen

P Phosphorus.

Q Hydroxy group (OH).

S Sulfur.

T Triple bond.

U —CH=CH— in arc or in chain.

W Dioxo group as in NO_2, SO_2, ClO_2, CO_2.

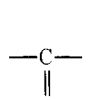
X Tetravalent carbon atom, or spiro carbon atom in fusion lines.

Y —CH<, —CH=, always with hydrogen, double bond indicated.

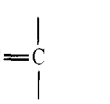
Z NH_2

Alphabet of Fusion Lines

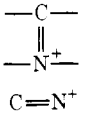
(A) Unsaturated Fusion Lines



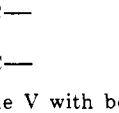
V (vertical)



E (equatorial)

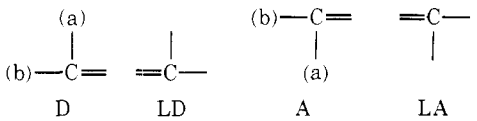


$C{=}N^+$



LV, Fusion line V with both double bonds to left.

(B) Terminal Segments of Fusion Lines



D means, downward from vertical or equatorial fusion lines, double bond to right.

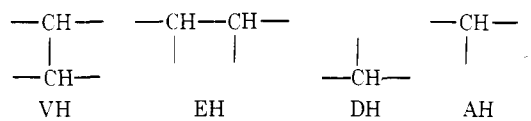
LD means, downward from vertical or equatorial fusion lines, double bond to left.

A means, upward from vertical or equatorial fusion lines, double bond to right.

LA means, upward from vertical or equatorial fusion lines, double bond to left.

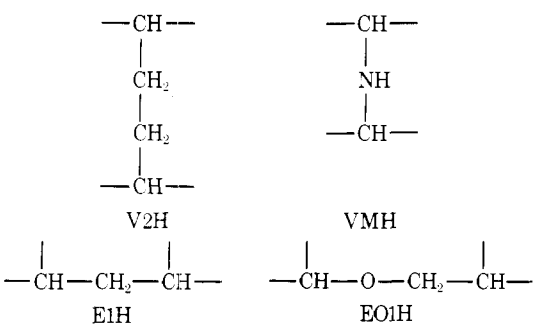
The terminal fusion lines are connected to other fusion lines (or among themselves) by a vertical single bond (a), or by a horizontal single bond (b), the latter (b) case is indicated by linkage symbol =.

(C) Hydrogenated Fusion Lines and Terminal Fusion Lines



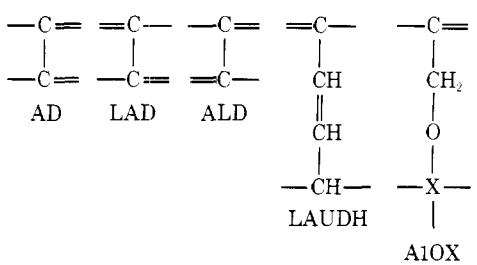
(D) Elongated Fusion Lines

(Three and more atoms fusion lines)



In addition to those indicated, other atoms can exist in the middle of elongated fusion lines.

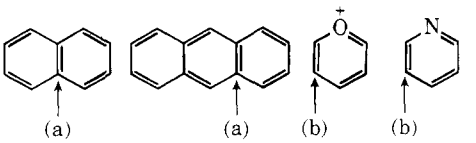
(E) Elongated Terminal Fusion Lines



Orientation and Double Bond Insertions of Ring Compounds. The manner in which a compound is drawn in the plane is also quite important, which is why the "Ring Index" [10] orientation and enumeration of the chemical compounds have been adopted, and why, to facilitate the encoding procedure, the "Ring Index" standardization already familiar to chemists is accepted. Standardization in this manner enhances the code's canonicity, which will be even further improved by a rule for depicting double bonds.

As models for depicting double bonds, naphthalene or anthracene may be used, so that the first fusion line (right to left) must have a double bond, drawn as indicated. In some cases, the presence of a hetero-atom partially predetermines the location of the double bond; but even here the double bond should be similarly indicated (if possible) on the first fusion line from right.

In the case of mono-ring compounds, the double bond will be shown on the left vertical of the structure.



(a); first fusion line (right to left)

(b); double bond in question on left

Fusion Line. If two or more rings are fused, then they have one or more fusion lines between them. The most frequently encountered fusion lines are vertical, but a few are horizontal, and the remainder are oblique. The vertical and horizontal directions are of course accepted as fundamental in our representation of fusion lines. Other directions of fusion lines are disregarded, and will be transposed to vertical or horizontal, depending on the encoding requirements.

Chain or branched chain forms of fusion lines will therefore be represented in code by a linear series of vertical and horizontal lines.

The accepted vertical and horizontal directions in turn predetermine the form of fusion line ring which in the code will appear as a rectangle, even if the ring is bigger than a 4-membered ring.

Encoding of Fusion Lines. The fusion lines and spiro atom(s) are encoded topologically from left to right, as they appear in the chemical structure. If a compound has two or more separated (not interconnected) fusion lines, then they are separated by a slash. At a higher complexity, however, the fusion lines can be interconnected.

If two or more fusion lines are interconnected by an outside bond with no carbon or hetero-atom between, then this (outside bond) is recorded in the fusion line code section. In such a case, the fusion lines may form a chain or branched chain.

In the case of cyclic interconnected fusion lines, the procedure of encoding is as follows: Starting with the outer left fusion line (or spiro-atom), move to the right and try to close the first ring of fusion lines, and then the second, and so on. The preferred direction from the starting point is toward the top of the fusion lines. The closure of a ring is indicated by a star and referred by a type of bonding to a particular part of the fusion line.

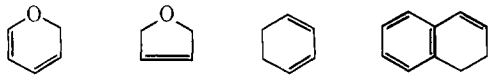
Thus, two interconnected or separated fusion lines will be represented as vertical ones, even if one of them deviates from verticallity. But, it can happen, that the fusion line bypasses two or more rings. Then, three or more atoms lie on the fusion line, which are encoded by symbol V (VH) with terminal ending A, D or by elongated VH.

ENCODING OF ARCS

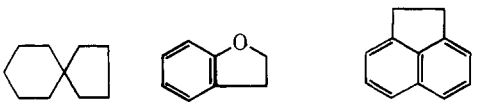
The next step is to match the line notation for the fusion lines with the arcs resting upon them. Thus the first arc is linked to the primed (or double primed) fusion line and to the neighboring one. Beginning with the prime marked fusion line, encoding of arcs proceeds clockwise. On the other hand, the double prime marked fusion line indicates counterclockwise encoding of arcs, as in steroids. Both approaches will result in an increasing order of enumerated atoms on the arcs, as cited by the "Ring Index." Another excellent approach is to prefix each substituent on the arc with a positional numeral.

If in a chemical structure a crossing pseudo bond and a long substituent exists, then the arc section is terminated by symbol J, after which the pseudo bond is listed, followed by a long substituent. The code for the crossing pseudo linkage starts with an underlining symbol.

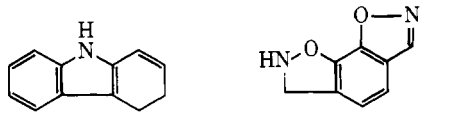
A few examples will now be given to illustrate the code.



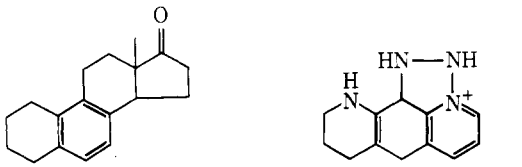
O1UU* O1N=Y1* UU2* VJU2/UU



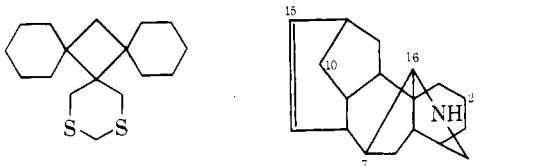
XJ5/6 VJO2/UU LA'VAJ2/=YU/UY=



VØV'JU2/UU/M V-V'JON=Y/U/1MO



V''-VØV-(1H)J4/U/2K/2 V-#A'HV-N^{+}JMM/=YU/1/3M

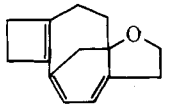


XXX'J5/S3S/5/1 V1YH&AH#-V'-(Y)J3Y/1Y/U/1J M1/-

An Alternative for Matching Arcs with Fusion Lines. To facilitate matching arcs, the fusion lines should be furnished with dot (a period), > (across), and comma symbols. The sequence of matching for both clockwise and counterclockwise directions is dot, > (across), comma, >, dot, which is a quasi seniority sequence of matching symbols. The same sequence should be maintained for inscribing those symbols at a fusion line. If any symbol is missing from a given section, then the next symbol of the sequence is used. The matching starts with marked fusion line which will coincide with the right upper part of the chemical structure for clockwise direction, and with the left lower part for counterclockwise direction of enumeration.

The following meanings are ascribed to the symbols: dot indicates that arc comes on the top of the chemical structure, > or < indicates that arc comes on the side of the chemical structure comma indicates that arc comes underneath of the chemical structure.

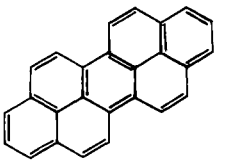
In proceeding around the chemical structure, two dots, or two commas, or one > (across) symbols are required for one arc. From each end of a fusion line or spiro atom, however, one dot or one comma may indicate the match of an arc. The symbol > (across) refers to both ends of a fusion line or to a single spiro atom.



>.VØ#,D1..X'.,LDJO2/=YY=/2/2

The arc section starts at X.
First arc requires two dots at X and LD.
Second arc requires two commas at LD and D.
Third arc requires one > at V.
Fourth arc requires two dots at V and X.

Polyfused Aromatic Ring Compounds. In polyfused aromatic compounds, the fusion lines inside of the structure, might form a benzene ring. It is now essential to enumerate the benzene ring corners in order to allocate the extreme left branch (a fusion line) and other branches on the ring. If the extreme left branch (a fusion line) is on the corner of the benzene ring where the double bond begins, then this position is considered as No. 1; if next to it, then this will be No. 2 position. The enumeration proceeds clockwise.



ELD-R4,ØV'DJUY=/U/U/=YU/U/U

By means of the code presented, it will be possible to encode the majority of known ring compounds. The code would appear novel. It does not require connectivity tables, and, it is suitable for evaluation of physical properties.

LITERATURE CITED

(1) Reid, R. C., and Sherwood, T. K., "The Properties of Gases and Liquids," p. 9, McGraw-Hill., New York, 1966.
(2) Brasie, W. C., and Liou, D. W., "Estimating Physical Properties; Chemical Structure Coding," *Chem. Eng. Progr.* **61** (5), 102–8 (1965); **61** (7), 16 (1965).
(3) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, 1968.
(4) Silk, J. A., "An Improved System for the Enumeration and Description of Ring Systems," *J. Chem. Doc.* **1**, 58–62 (1961).
(5) Silk, J. A., "A Linear Notation for Organic Compounds," *Ibid.*, **2**, 189–95 (1963).
(6) Gould, D., Gasser, E. B., and Rian, J. F., "Chemical Search—An Operating Computer System for Retrieving Chemicals Selected for Equal, Analogous or Related Character," *Ibid.*, **5**, 24–32 (1965).
(7) Lefkovitz, D. A., "A Chemical Notation and Code for Computer Manipulation," *Ibid.*, **7**, 186–92 (1967).
(8) Lefkovitz, D. A., "Use of a Non Unique Notation in a Large Scale Chemical Information System," *Ibid.*, **7**, 192–210 (1967).
(9) Skolnik, H., "A New Linear Notation System Based On Combinations Of Carbon And Hydrogen," *J. Heterocycl. Chem.* **6**, 689–95 (1969).
(10) Patterson, A. M., Capell, L. T., and Walker, D. F., "The Ring Index," Amer. Chem. Soc., Washington, D. C., 1960–5.