

# An Encoding System for a Group Contribution Method<sup>†</sup>

Delin Qu,<sup>‡</sup> Bao Fu,<sup>‡</sup> Massaaki Muraki,\* and Toyohiko Hayakawa<sup>§</sup>

Department of Industrial Engineering and Management, Tokyo Institute of Technology, Tokyo, Japan 152

Received February 10, 1992

In order to use the group contribution method for physicochemical property estimation, a suitable chemical language (encoding system) for the expression of the structural information of chemical compounds is required. An advanced encoding system (AES) is proposed by improving the WLN (the Wiswesser line notation) notation rules for the cyclic compounds, except benzene, in order to reflect the structural information of atomic level in a readable form. The AES emphasizes the valence bonds and the internal connection relations between encoded atoms with common rules for coding cyclic compounds (except benzene). In order to construct the AES, three working processes (locant path, structure information, and uniqueness regulation) have been developed.

## INTRODUCTION

The group contribution method, which is often used for physicochemical property estimation in process design,<sup>1,2</sup> possesses an extended application domain and better accuracy.<sup>3,4</sup> When it is used to estimate physicochemical properties of a chemical compound, the functional groups must be taken apart according to the requirement of the relevant group contribution method based on its structure analysis.<sup>5,6</sup>

Evidently one of the prerequisites to the structure analysis is the definition of a chemical compound in terms of a chemical language (encoding system) which can be accepted by computers.<sup>7-9</sup> Therefore, a suitable chemical language to express the structure information of a chemical compound is required. This is usually related to the classes of atom or superatom (aggregates of atoms such as a carbon atom with its immediate neighboring hydrogen atoms), valence bond, and saturation of this compound. For this chemical language, three conditions are necessary. First, the language can definitely describe the structure information of atomic level in human-readable form, i.e., the microinformation and the topological structure of a compound can be displayed. Second, it should possess a compact structure and simple notation rules. Third, it should be convenient to computer handling. There are three broad categories of chemical languages by which the structure information is represented and communicated.<sup>10-12</sup>

**(1) Fragment Codes.** These are used as the primary means of structural information communication for chemical compounds, but they fail to reflect structural information at the atomic level.

**(2) Line Notations.** With the introduction of computers, these become widely used to describe the chemical structure.<sup>13</sup> They offer a compact way of completely representing the structure of a chemical compound, but it is difficult to read the structure information from line notations.

**(3) Connection Tables.** They are used extensively to represent the structure information at the atomic level and usually provided as an internal topological description of the molecular structure by most

computer-based transformation techniques.<sup>14</sup> They are not however compact or convenient for computer management.

Based on this comparison, the most feasible method would appear to be to select a line notation method as the chemical language and to provide it with the functions of connection tables.

There are several line notation methods available, such as the Wiswesser line notation (WLN)<sup>15</sup> and the simplified molecular line entry system (SMILES).<sup>16,17</sup> The WLN divides chemical compounds into two classes (branch chain including benzene and cyclic compounds) for coding. For the branch chain compounds, the structure information can be better represented. But for the cyclic compounds, except benzene, the structure information of atomic level cannot be reflected visibly. This is due mainly to its notation method and complicated notation rules. The SMILES system only requires some simple rules for generating original linear notation, and ancillary computer programs are necessary to determine the unique SMILES. Though these line notation methods enjoy some popularity among chemists for their relevant application objectives, neither of them can reflect the structure information of atomic level in a human-readable form. However, based on the requirement of group contribution method, the WLN possesses the better characteristics because its notation rules and method are more easily improved than SMILES.

The purpose of this study is to develop an advanced encoding system (AES) for the group contribution method by improving WLN's notation rules and method for cyclic compounds except benzene. The AES takes the same approach as WLN,<sup>15</sup> dividing chemical compounds into two classes for coding. For branched chain compounds, the notation rules are similar to those of WLN with only minor revision. For cyclic compounds except benzene, because there are many kinds of rings and the WLN strictly obeys the traditional classification convention in developing notation, the program must distinguish the class of cyclic compounds first, then divide cyclic compounds into fused rings, bridged rings, spiro rings, perifused rings, pseudobridged rings, and so on. The different notation rules are necessary for the different cyclic compounds. The AES strives to describe the structure information at the atomic level. On this basis, our emphasis is to represent the valence bonds and the internal connections between the encoded atoms. The key to solving this problem is to break down the convention by which the notation is implemented according to the strict

<sup>†</sup> Presented in part at the 4th APCChE'87, Singapore, May 1987.

\* To whom all correspondence should be addressed.

<sup>‡</sup> Present address: Department of Chemical Engineering, Tsinghua University, Beijing, China 100084.

<sup>§</sup> Present address: Department of Industrial Engineering, The Nishi Tokyo University, Yamanashi, Japan 409-01.

classification of chemical compounds. Common rules are established by combining the advantages of the line notation and the connection table methods. There is no need for the AES to distinguish different classes of cyclic compounds; the types of rings (including heterocyclic and carbocyclic rings) need only be identified with different symbols. In order to construct the AES, the sufficiency of the reflected structure information and the uniqueness of the structure notation are the premises on which the notation rules are determined. In order to ensure these two premises, three working processes are proposed. First, the locant path that affects the unique structure notation is determined. Second, the contents which are related to the sufficiency of structure information and topological expression are defined. Finally, the reasonable regulations which guarantee the uniqueness of the structure notation are developed. These processes are the notation method for the AES.

The AES proposed here can definitely reflect the microinformation and the topological structure of chemical compounds; the effectiveness and uniqueness of the reflected structure information are illustrated by the encoding examples. Because they can express the structure information at the atomic level visibly with only simple notation rules, the corresponding functional groups for various group contribution methods can be easily identified using the computerized decoding system.<sup>6</sup>

### LOCANT PATH

Before structure coding, it is necessary to determine a proper and unique locant path. This provides the base for describing the notation contents and guarantees the uniqueness of the structure notation. In order to obtain this unique locant path, two principles for determining a locant path are proposed for the AES. The first one is the choice of a locant path, which should include all cyclic atoms whenever possible. If one locant path cannot include all ring atoms, the branched locant path must be indicated. The second one is to make a choice of a branched locant path, which should take the earliest locant and connect with the branched locant path whenever possible.

According to the previous principles, determination procedures of a locant path in the AES are proposed. First, taking some ring atom as a starting point "a", the other neighboring ring atoms are marked with sequential letters "b, c, d, ..." to indicate the position of the corresponding atom. For a large cyclic structure where the number of ring atoms is over 26, adding "&" after the sequential letters to increase the count range is adopted. The 27th locant is written as "a&", the 28th as "b&", and so on. If it is still not enough, two or more "&" symbols are used after the alphabetic character. A lowercase letter is used to prevent confusion with the element symbol when locating a locant path in the structure formula; in formal notation the capital letter is usually adopted.

For a more complex cyclic structure, when one locant path cannot include all ring atoms, branched locant paths must be used. The method of marking is to add one "\*", two "\*\*", ... after some locant. Figure 1 shows the marking method for a branched locant path where "a-i" is the longest locant path to include cyclic atoms and "bb\*" is the branched locant path in which "b" is the first locant to join with the branched locant path.

These principles of determining a locant path and the specific procedures provide the sufficiency for the defining a unique locant path.

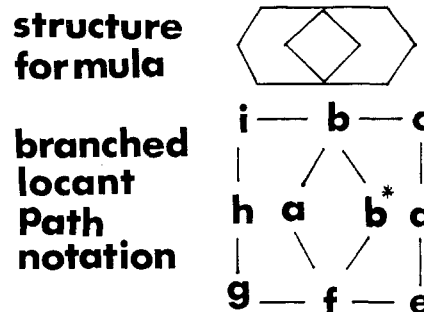


Figure 1. Branched locant notation method.

### STRUCTURE INFORMATION

The contents of the structure notation are directly related to the sufficiency and the completeness of the structure information, which reflects the microinformation and the topology of a chemical compound. The valence bonds and the internal connection relations between the cyclic atoms are emphasized to express the structure of a cyclic compound. Thus, the two kinds of information, corresponding to the integrative topological structure and the structural microinformation, are the necessary contents which visibly indicate the structure information at the atomic level. In particular, the following eight items are proposed to guarantee the sufficiency and the completeness for the representation of the structural information.

Figure 2 shows the examples for the cyclic compound notation except benzene, and 2 is taken mainly to illustrate the procedures of the AES.

(1) **Beginning Notation.** T and L indicate the heterocyclic and carbocyclic rings, respectively. Because 2 is a carbocyclic compound, its beginning notation is L.

(2) **End Locant.** This content involves the integrative topological information of a cyclic compound and gives the total number of encoded atoms in the determined locant path, which is the base of structure microinformation. The end locant is written just after the beginning notation without the punctuation symbol. The end locant of the locant path of 2 is X, so up to this step its notation is LX.

(3) **Locant of All Branched Locant Paths.** The information of the branched locant paths also reflects the integrative topological structure information of the compound, which constitutes the important part of the topological structure representation.

The locant in a branched locant path is punctuated with a ";" introduced as the final symbol of the individual item. After ";", "B" is adopted to indicate the locant in a branched locant path, then the locant number in the branched locant path is written one by one. No punctuation symbol is written between the two locants.

For the example, the locant of the branched locant path is I\*I\*\*, and the notation up to this step becomes LX;BI\*I\*\*.

(4) **Locant Pair of Ring Atoms.** This information reflects the structural information pertaining to ring atoms and provides the internal connection relationships between the ring atoms. If there is more than one locant pair, priority rules must be used to determine the coding sequence in order to guarantee the uniqueness of the structure notation. The coding method is as follows. First, the symbol ";" is used to separate the prior item. Then the "/" symbol is written, to distinguish the different locant pairs. Finally, the locant pair is written. If there is more than one locant pair, "/" is inserted before the next locant pair. For the coding sequence, the following priority rules are used. In each pair, the earlier

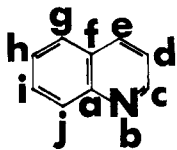
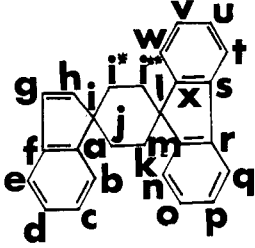
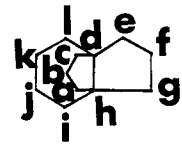
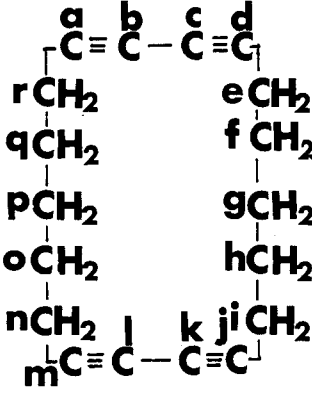
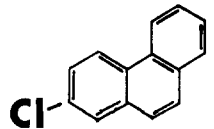
No	compound structure	structure notation
1		<b>TJ;/AF/;!BN; U/BC :</b>
2		<b>LX;BI*I**;/AF /AI/I**L/LX/ MR/SX;-II* I**JKL:</b>
3		<b>LL;/AH/ DH/DL; = :</b>
4		<b>LR;/AR;!AC! BC!CC!DC! JC!KC!LC! MC;U#AB# CD#JK#LM:</b>
5		<b>LN;/AJ/AN/ BG;-;!EG</b>

Figure 2. Encoding method for cyclic compounds except benzene.

alphabetic locant is coded first; for different pairs, the earlier locant with alphabetic sequence for first locant is coded. If the first locants are identical, the earlier locant with alphabetic sequence is coded for the second locant. If a branched locant path is involved, the terminal locant of the branched locant path is paired with the other locant.

In the example, there are six locant pairs, and each pair is interactive, but the locants are not continuous. According to the principle of the earlier alphabetic locant coding first, the notation will be LX;BI\*I\*\*;/AF/AI/I\*\*L/LX/MR/SX up to this step.

**(5) Specification of Heteroatom or Special Ring Segment.** This information indicates the structure information of the valence bond of the ring atoms, which is one of the most important pieces of microinformation. A special ring segment involves special carbon atoms, which have two kinds of structures:  $=C=$  and  $-C\equiv$ . First, a ";" is written followed by a "!" symbol. This serves as the distinction symbol to locate the locants of heteroatoms or special ring segments. The locant of the heteroatom or the special ring segment locant is coded first, and finally, the coding symbol of the heteroatom

or the special ring segment coding symbol, based on the locant sequence, is coded.

In the previous example, there is no heteroatom or special ring segment, but the compound 1 in Figure 2 possesses a nitrogen, and the heteroatom locant should be written as TJ;/AF/AJ;!BN up to this step.

**(6) Indication of Saturation.** This information relates to the saturation of ring atoms and provides the characteristic of the valence bonds and internal connection relations between the encoded atoms.

For a heterocyclic compound or a compound with a special carbon atom, when the unsaturation of a heteroatom or special carbon atom is coded, first a ";" is used. Then the symbol "U", the distinction symbol for locating the heterocyclic or special carbon atom unsaturation is written. After that, the locant pair of the heteroatom or special carbon atom and the atom which forms a double bond with the heteroatom or special carbon atom is coded. If more than one locant pair of heteroatoms or special carbon atoms and the other atoms that constitute double bonds exist, each locant pair is separated with the symbol "/". A triple bond should also be coded in this term, and the coding method is the same as above, only the distinction symbol "/" becomes "#". In the previous example, the unsaturation of the heteroatom nitrogen of 1 is indicated as TJ;/AF/AJ;!BN;U/BC up to this step.

For a carbocyclic compound, if the number of the unsaturated carbon atoms are a majority of the cyclic carbon atoms, the saturated carbon atoms are coded instead for simplicity. The coding method is to write the symbol "-" after the punctuation ";". The locants of the saturation carbon atoms are then coded without any punctuation symbol between each locant. If there are no saturated carbon atoms, only a "-" symbol is coded and no other content need be written. If the number of saturated carbon atoms is a majority of the cyclic carbon atoms, the unsaturated carbon atoms are coded instead. The coding method is the same as the saturated cyclic carbon atoms, only changing the "-" symbol into a "=" symbol is required. Compound 4 of Figure 2 possesses a triple bond whose carbon atoms should be written as LR;/AR;!AC!BC!CC!DC!JC!KC!LC!MC;U#AB#CD#JK#LM up to this step.

In the original example, the locant of the saturated carbon atoms of 2 should be indicated. The coding procedure is to write the locant of the saturation carbon atoms after the ";" and the "-". Thus, the notation is LX;BI\*I\*\*;/AF/AI/I\*\*L/LX/MR/SX;-II\*I\*\*JKL up to this step. Compound 3 of Figure 2 only has saturated carbons, and so the notation is written as LL;/AH/DH/DL;= up to this step.

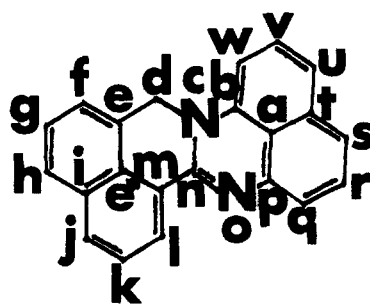
**(7) End Notation.** The end notation for the cyclic structure is the symbol ":". After this step, the notation of 2 is completed, and the final expression is as follows: LX;BI\*I\*\*;/AF/AI/I\*\*L/LX/MR/SX;-II\*I\*\*JKL:.

**(8) Notation for Substituent on a Ring.** This information represents the structure information for the atom of a ring that is related to substituents. If there are substituents on the ring, their positions and names should be expressed in this item. The notation method is to cite the locant and name of the relevant substituent followed by a "!" symbol. Different substituents are separated with a "!" symbol. In Figure 2, compound 5 has a -Cl substituent; its notation is written as LN;/AJ/AN/BG;-;!EG, where E and G indicate the position and name of this substituent, respectively.

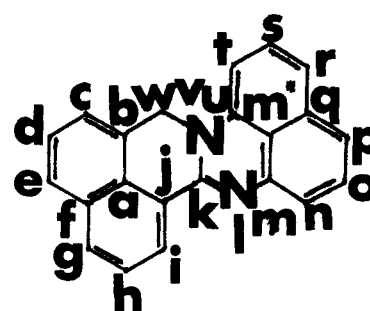
Aside from the beginning and end notations, if there is not any item of the above contents, the expression for that item is omitted. The symbols which have different meanings with

Table I. Meaning of Special Symbols

symbol	meaning
B	beginning symbol of branched locant path for coding cyclic ring structure except benzene
C	carbon disconnected to hydrogen includes $\text{=C=}$ and $\text{—C=}$ structures
L	beginning symbol of carbocyclic ring notation except benzene
T	beginning symbol of heterocyclic ring structure notation
U	identified symbol for indicating the unsaturation of heterocyclic atom
&	end symbol of branched locant notation
	end symbol for ring substituent contained
	sign of number of branched chain
	locant expansion symbol
-	distinction symbol for two-character element
	distinction symbol of two-digital number
/	sign of saturated carbon atom in ring notation
	sign of noncontinuous locant pair
:	sign of nonsaturated bond locant pair in ring notation
:	punctuation symbol for each item in ring notation
:	end symbol of ring notation
*	locant mark of branched locant path in ring notation
<	mark of branched symbol excluding branch symbols X, Y, and Z <sup>3</sup>
!	locant mark; when citing substituent, heteroatom, and special ring segment in substance-contained ring, cite according to locant, add "!" before each locant
=	mark of nonsaturated carbon atom in ring notation
#	distinction symbol between two locant pairs of heteroatoms or special ring segment which constitutes a triple bond



(a)



(b)

Figure 3. Choice of notation locant.

WLN are summarized in Table I where the special symbols for the notation of cyclic compounds except benzene are indicated.

#### UNIQUENESS REGULATIONS AND COMPARISON

The final criterion for judging a new encoding system is the uniqueness of the structure notation produced. In order to guarantee the uniqueness of the structure coding, the regulations must be worked out. Therefore, the following priority rules (earliest locant rules) have been developed to provide

the order of choice. The first three regulations ensure the uniqueness of the integrative topological structure information, and the subsequent five regulations guarantee the uniqueness of the structure microinformation, which relates valence bonds and connection relations between encoded atoms. Provided that the encoding is implemented according to this priority sequence, the uniqueness of the structure notation can be assured.

- (1) the least locant path for branch locant path
- (2) the earliest locants for coding the branched locant path
- (3) the earliest locants for coding the interactive non-continuous locant pairs
- (4) the earliest set of locants for coding the heteroatom or the special cyclic ring segments. If there is more than one heteroatom or special ring segment to be encoded, the one which has the latest position in an alphanumeric list should be cited first.<sup>15</sup>
- (5) the earliest set of the locants for coding the unsaturation of the heteroatoms
- (6) the earliest set of locants for coding the saturation or the unsaturation of the cyclic carbon atoms
- (7) the earliest set of the locants for coding the substituent. If there are more than one substituent to be encoded, the one which has the latest position in an alphanumeric list should be cited first.<sup>15</sup>

In these "earliest locant" rules, the first one is considered initially according to the priority sequence. Choosing the least locant path for the branch locant path means to search a locant, including all cyclic atoms whenever possible. For a chemical compound with a complex structure, there may be several potential choices (alternatives) which are difficult to assess. An example of a complex structure shown in Figure 3 is used to illustrate this choice. For this example, it is impossible to choose a locant including all cyclic atoms, but there are two alternatives, each with only one branch locant path, shown in Figure 3a and 3b, respectively. Then according to the earliest locant for coding the branch locant path, the one shown in Figure 3a has the earlier locant. Consequently, Figure 3a is determined as the notation locant. Therefore, for the worst case, all the possible alternatives may be required, and then after the comparison, using these rules, the unique AES notation can be finally obtained.

Through these three working processes, the notation of cyclic compounds except benzene has been modified, and a rational encoding system has been constituted. The comparison between the AES and the WLN encoding method is shown in Figure 4, where one branch chain chemical compounds and two cyclic compounds, except benzene, are involved. It is clear that the proposed AES indicates the structure information of atomic level visibly from Figure 4, where the distinction symbols provides the definite expression of the individual content. From the examples in Figure 4, the WLN distinguishes the type of cyclic compounds first, then the corresponding rule is used for the notation. The AES uses the common rules for the notation of cyclic compounds; therefore, the encoding procedures are simpler and user friendly. Moreover, the AES emphasizes the information of the valence bonds and the internal connection relations between atoms, which reflects the structure information of atomic level visibly.

#### DISCUSSION

Judging a notation method by its usefulness for estimating physicochemical property can be considered from two aspects.

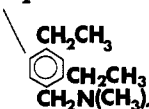
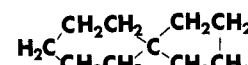
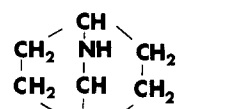
structure formula	WLN structure notation	AES notation
$(\text{CH}_3)_2\text{NCH}_2$ 	1N1&1 2 ER B2 D2	1N<1&1R<  B2 D2 E 1N<1&1
	L6XTJ A & AL5XTJ	LJ; /AE/EJ; =:
	T66 A B A MTJ	TH; /AF/CH;  AM;=:

Figure 4. Comparison between the AES and the WLN encoding methods.

First, it is easy to generate functional groups for the relevant group contribution method. Generally speaking, the first step for generating functional groups is to convert the line notation into a corresponding connection table. Because the notation method and rules of the WLN are complicated for cyclic compounds except benzene, generating the connection table is a time-consuming job;<sup>14</sup> therefore, the WLN is not suitable for the requirement of physicochemical property estimation in a database. Second, it saves computer memory for storing structure information. It is known that conversion of chemical substance names into a connection table is an important means for processing chemical information for the computer.<sup>18,19</sup> However, it is difficult to store this kind of structure information in a computer because the memory space of the computer is limited. For example, in compound 2 of Figure 2, more than 100 characters are required to store the structure information of this simple compound using the connection table method, but the AES stores the same structure information with only 12 characters. Generally, the AES notation can save memory storage by about a factor of  $n^2$ , where  $n$  is the number of encoded characters. The AES can easily generate the encoded adjacency matrix (one kind of connection table). Then on the basis of this, the relevant functional groups may be generated smoothly, because the structure information of the valence bonds and connection relations for the relevant functional groups is involved in the AES notation.

For branch chain chemical compounds (acyclic compounds), the basic notation rules of the AES are similar to the WLN, but WLN's simplification rules such as multiplier and methyl contractions are eliminated and the distinction symbols are introduced for definite expression of the individual notation content. Therefore, AES for acyclic compounds is more easily identified by the computerized decoding system for generating the relevant functional groups.

## CONCLUSION

AES uses common notation rules for coding cyclic compounds except benzene. In order to construct the AES, three working processes are proposed for the unique representation of the microinformation and the topology of a compound. Because there are no ancillary computer programs required for implementation of the structure notation, it may be convenient for more chemists and chemical engineers to use it.

This paper is intended to describe the AES. The decoding system for the group contribution method is proposed based on the AES,<sup>6</sup> both of which have been used for physicochemical property estimation system with success.

## REFERENCES AND NOTES

- Reid, R. C.; Prausnitz, J.; Poling, B. *The Properties of Gases and Liquid*, 4th ed.; McGraw-Hill: New York, 1987; p 388.
- Leonet, H.; Melli, T.; Montagna, J. M.; Vecchiotti, A.; Cerro, R. L. SIMBAD: A Process Simulator Linked to A DBMS-III. The Physicochemical Properties Package. *Comput. Chem. Eng.* **1987**, *11*, 217-226.
- Richard, S. H. Mah. *Foundations of Computer-Aided Chemical Process Design*; Engineering Foundation: New York, 1981; p 3.
- Kajima, K.; Tochigi, K. *ASOG and UNIFAC*; Chemical Engineering: Tokyo, 1987; p 20.
- Fredenslund, A.; Jones, R. L.; Prausnitz, J. M. Group Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures. *AIChE J.* **1975**, *21*, 1086-1099.
- Qu, D.; Su, J.; Muraki, M.; Hayakawa, T. A Decoding System For Group Contribution Method. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, following paper in this issue.
- Attias, R. DARC Substructure Search System: A New Approach to Chemical Information. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102-108.
- Balaban, A. T. Application of Graph Theory in Chemistry. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334-343.
- Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS Online Search System. 1. General System Design and Selection, Generation and Use of Search Screens. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93-102.
- Stobaugh, R. E. Chemical Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 271-275.
- Fugmann, R. Peculiarities of Chemical Information from a Theoretical Viewpoint. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 174-180.
- Cooke-Fox, D. I.; Hirby, G. H.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 1. Introduction and Background to a Grammar-Based Approach. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 101-105.
- Wiswesser, W. J. Historic Development of Chemical Notation. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 258-263.
- Eakin, D. R. Graphic Challenge WLN. Can WLN Hold Fast? *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 101-103.
- Smith, E. G. *Wiswesser Line-Formula Chemical Notation Method*. McGraw-Hill: New York, 1968; p 77.
- Weininger, D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31-36.
- Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97-101.
- Vander Stouw, G. G.; Naznitsky, I.; Rush, J. E. Procedures for Converting Systematic Names of Organic Compounds into Atom-Bond Connection Table. *J. Chem. Doc.* **1967**, *7*, 165-169.
- Vander Stouw, G. G.; Elliott, P. M.; Iserberg, A. C. Automated Conversion of Chemical Substance Names to Atom-Bond Connection Tables. *J. Chem. Doc.* **1974**, *14*, 185-193.