A Decoding System for a Group Contribution Method

Delin Qu,† Jianmin Su,† Masaaki Muraki,* and Toyohiko Hayakawa‡

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

Received February 10, 1992

Both a computerized chemical language (encoding system) and a computerized decoding system are necessary for physicochemical property estimation using the group contribution method. An efficient decoding system with two basic procedures, which takes chemical compounds apart into the corresponding functional groups, has been developed on the basis of the AES (advanced encoding system, a line notation method). The first step involves the generation of an encoded adjacency matrix based on the structure information analysis of three steps (information analysis, reverse deduction, and matrix generation). The second step is the identification of the relevant functional groups based on structure decomposition in three steps (scan and search, fragment identification, and group generation). The decoding system converts the structure information of the AES into the relevant functional groups. The effectiveness and reliability of the decoding system are illustrated with examples.

INTRODUCTION

Physicochemical properties, which are fundamental to chemical process calculations, may be estimated by several methods. The group contribution method is often adopted for such estimations, because it possesses proper robustness and an extended application domain with satisfactory accuracy. In order to use the group contribution method, there are two prerequisites. One is a suitable computerized chemical language (encoding system) to express the structure information of a chemical compound, and the other is a computerized decoding system to disassemble chemical compounds into the corresponding functional groups based on the structure information. Along with the development of database techniques, the line notation methods have become widely used to describe chemical structures because they occupy less memory space than connection tables. As a suitable chemical language, three conditions must be met. The code must describe the structure information of atomic level in a human-readable form, possess a compact structure and simple notation rules, and be convenient to computer memorizing and handling. There are several line notation methods such as the Wiswesser Line Notation (WLN)[1] and the Simplified Molecular Input Line Entry System (SMILES) methods,[2,3] but they cannot reflect the structure information of atomic level in a readable form. Based on the requirement of the group contribution method, the AES (advanced encoding system),[4] which can reflect the microinformation and the topological structure (i.e., describe the structure information of atomic level visibly), was developed by improving the WLN's notation rules and method for cyclic compounds except benzene.

The decoding system does not convert the line notation (structure information) into the molecular structure,[5-10] but into the relevant functional groups for various group contribution methods. Therefore two functions should be achieved by the decoding process. One is to analyze the structure information of a chemical compound, and the other is to decompose it into the relevant functional groups. In order to implement this conversion, as a result of the structure information analysis, the generated relationship of the encoded atoms must be stored in the computer memory.

There are few publications concerning conversion of WLN notation to a connectivity matrix.[11,12] Some in-house software for searching the structure information of WLN was developed for the molecular structure display,[13-19] but further decomposition into functional groups was never mentioned. Moreover, generating connection tables for WLN is said to be time-consuming,[20] because its encoding rules are complicated. Recently the fragment generation method from the connection tables was proposed for the purpose of recognition of the chemical structure,[21] but the searching efficiency is not satisfactory for the group contribution method. No paper covers this topic of a decoding system which converts the line notation into the relevant functional groups.

The purpose of this study is to develop an efficient decoding system for the group contribution method based on the AES method.[4] It is required to convert the structure information of the AES into the relevant functional groups. Because the AES can express the structure information of atomic level visibly and offers advantages for physicochemical property estimation with the group contribution method, it is possible to disassemble the AES structure notation into corresponding functional groups using an efficient decoding system. Therefore, the decoding system is developed with two basic procedures based on the AES. The first one reflects the relationship and the relevant number of the encoded atoms or superatoms (aggregates of atoms such as a carbon atom with its immediate neighboring hydrogen atoms) based on the structure information analysis; then the encoded adjacency matrix, (a kind of connection table) whose elements are the encoded atoms or superatoms, is generated as the most convenient working form for the computer. The second procedure accomplishes the identification of the functional groups based on the structure decomposition. The identification of the functional groups is implemented for various group contribution methods, because they require different functional groups.

The first procedure is the reverse of encoding. Hence, the encoding rules of the AES are used to generate an encoded adjacency matrix. The proposed method of generating the encoded adjacency matrix consists of three steps: information analysis, reverse deduction, and generation of the encoded adjacency matrix.

* To whom correspondence should be addressed.
† Present address: Department of Chemical Engineering, Tsinghua University, Beijing, China 100084.
‡ Present address: Department of Industrial Engineering, The Nishi Tokyo University, Yamanashi, Japan 409-01.

0095-2338/92/1632-0448$03.00/0

The second procedure is identification of the functional groups. The proposed method of generating the functional groups is also composed of three steps: scan and search, fragment (functional group) identification, and generation of the functional groups.

The decoding process is illustrated by examples, showing that the proposed decoding system is effective and reliable for decomposing the structure information of the AES into the required functional groups. It constitutes a superior computerized decoding system for the group contribution method.

GENERATION OF AN ENCODED ADJACENCY MATRIX

Generating an encoded adjacency matrix is an indispensable procedure for the decoding system: it is the reverse process of encoding. In the encoding process, provided that the encoding rules and method are given, the notation of a chemical compound can be determined. In the decoding process, the computer must analyze the structure information first and then implement the decoding. Therefore, the computer should be instructed with the encoding rules and method. In other words, the rules and method for encoding have to be enunciated quite clearly in the computer program. Then the reverse deduction can be carried out.

Because the AES divides chemical compounds into two classes for encoding, namely (i) branch chain chemical compounds including benzene rings, and (ii) cyclic compounds except benzene, the decoding process should normally decipher the corresponding classification of chemical compounds. However the decoding conception and procedures are common to any kind of chemical compound. One proceeds through the general decoding program to generate an encoded adjacency matrix. Moreover, the AES improves the notation of cyclic compounds other than benzene based on the WLN, and the decoding process based on the AES becomes time-saving and easier than the WLN method. Consequently, cyclic compounds (except benzene) are selected to illustrate the handling process and the method for generating an encoded adjacent matrix.

Information Analysis. The first step is to analyze the information provided by the AES. The first information which can be identified by the computer is the classification of the structure notation. The notation items are read one by one, and the structure information is extracted from the relevant notation items.

In the case of cyclic compounds except benzene, because the AES emphasizes the microinformation and the topological structure of a chemical compound, the information analysis and extraction should reflect the structure characteristics and interactive relationships. There are eight items of notation:

(1) Beginning Notation. This offers the first information about the classification of cyclic compounds.

(2) End Locant of a Locant Path. Using the encoding rules, the information on the number of encoded atoms or superatoms which are involved in the longest notation path can be extracted.

(3) Locants of All Branched Locant Paths. This item reflects the number and the circumstance of the branched locant paths connected to the longest locant path.

(4) Locant Pair of Ring Atoms. The locants of a locant pair, which are interactive but not continuous, reflect the connection relation and circumstances of the encoded atoms or superatoms.

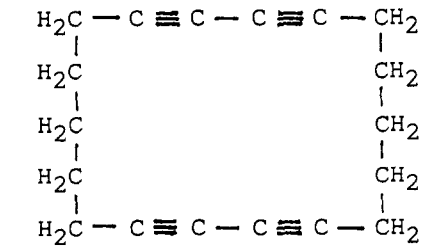

Figure 1. Structural formula of a cyclic unsaturated compound.

(5) Locants of Heteroatoms or Special Ring Segment. The locant positions of the heteroatoms or special ring segments are obtained.

(6) Locants for Saturated or Unsaturated Heteroatoms and Carbon Atoms. Information on unsaturation is extracted.

(7) End Notation. It indicates the end of the cyclic ring structure.

(8) Notation for Substituent on the Cyclic Ring. On the basis of it, the position and name of the substituent can be determined.

Consequently, the analysis is the process by which the structure information using the notation rules and method is judged, and it is also the process for extracting the information. Taking the notation of a cyclic compound as an example, the information extraction process is as follows.

For example, according to the AES, a cyclic compound (Figure 1) has following notation (LR;/AR;!AC!BC!CC!DC!JC!KC!LC!MC;U#AB#CD#JK#LM:). From the beginning symbol of the notation "L", a carbocyclic ring compound can be determined. From the second item "R", the longest locant path contains 18 encoded atoms or superatoms (from A to R). Because of the nonexistence of a third item, it indicates no branched notation locant. Therefore, this compound contains a total of 18 encoded atoms or superatoms. From the fourth item "/AR", it is known that the locant "A" is connected with the locant "R" in the compound. In other words, this is a large cycle which contains 18 encoded atoms or superatoms. The fifth item "!" shows that there are a total of eight "special carbon atoms". According to the notation rules, special carbon atoms may have two kinds of structures =C= and —C≡. From the sixth item "U" and "#", it is determined that the special carbon atoms are —C≡. The end symbol ":" indicates the end of the cyclic ring structure, and there is no substituent on the ring as would be described by an eighth item.

Reverse Deduction. After extraction of the useful information from the structure notation, the reverse deduction and the synthesis process are implemented based on the notation rules and method.

In the case of cyclic compounds (except benzene), the information of the second item is first combined with the information of the third item to get the total number of encoded atoms or superatoms. Then the information from items 4 and 5 was used to locate the position of the special carbon atoms and to define the valence bond of the special carbon atoms and the other carbon atoms. Because the connectivity relation of the encoded atoms or superatoms is provided, one can determine that beside the special carbon atoms, the other atoms are carbon atoms based on the notation rules. Moreover, according to the circumstances of the valence bonds of the special carbon atoms, the bonds of the other carbon atoms and the number of hydrogen atoms which are connected to each carbon atom can be determined. Finally, all the information may be synthesized to deduce the structural formula from the structure notation.

	C	C	C	C	CH_2	CH_2	CH_2	CH_2	CH_2	C	C	C	C	CH_2	CH_2	CH_2	CH_2	CH_2
C		3																1
C	3		1															
C		1		3														
C			3		1													
CH_2				1		1												
CH_2					1			1										
CH_2						1			1									
CH_2							1			1								
CH_2								1			1							
C									1			3						
C										3		1						
C											1		3					
C												3		1				
CH_2													1		1			
CH_2														1		1		
CH_2															1		1	
CH_2																1		1
CH_2	1																1	

* The values of remaining matrix elements are zero

Figure 2. Encoded adjacency matrix of the example of Figure 1.

The reverse deduction is now carried out using the previous example. First a total of 18 encoded atoms or superatoms are counted in the compound. Eight special carbon atoms have the structure —C≡, and their positions are determined successively. Finally because the locant "A" is connected to locant "R", the whole molecule must be a large cycle in which, beside the special carbon atoms, all other atoms are carbon atoms connected with saturated bonds. Through atom-by-atom deduction, the structural formula with this structure notation can be obtained and expressed as shown in Figure 1.

Generation of an Encoded Adjacency Matrix. Through the reverse deduction, the structure and the relation between the encoded atoms or superatoms are determined. The molecular structure that has been analyzed is stored temporarily in the computer as an encoded adjacency matrix, which is composed of the encoded atoms or superatoms and is a suitable type of connection table. The element of an encoded adjacency matrix expresses the connection relation and the valence bond. Zero indicates no connected relation, and the figures represent the multiplicity of the bond.

After the procedures mentioned above, the following encoded adjacency matrix can be obtained for the previous example as shown in Figure 2, where the first column and first row represent the encoded atoms or superatoms.

A general decoding program is developed with Fortran 77 to accomplish this procedure.

GENERATION OF THE RELEVANT FUNCTIONAL GROUPS

The second procedure of the decoding system is to decompose a chemical compound into the functional groups which are required for physicochemical property estimation. In some sense the decomposition problem is the process of identification of the functional groups. For physicochemical property estimation, the kinds of the functional groups and the requirements are given by the corresponding group contribution method, which is different from the discovery of the functional groups for optimal synthetic pathway.[21,22] Therefore, it is necessary to develop a more efficient and rational method for generating the relevant functional groups.

Because the basis of functional group identification is the explicit expression of a chemical structure, the encoded adjacency matrix provides the base for functional group identification. In addition, different estimation methods require different functional groups for the physicochemical property estimation. Each estimation method therefore has its own, specific decomposition program. Though the decomposition programs are different, the idea of generating the relevant functional groups is the same. Therefore the

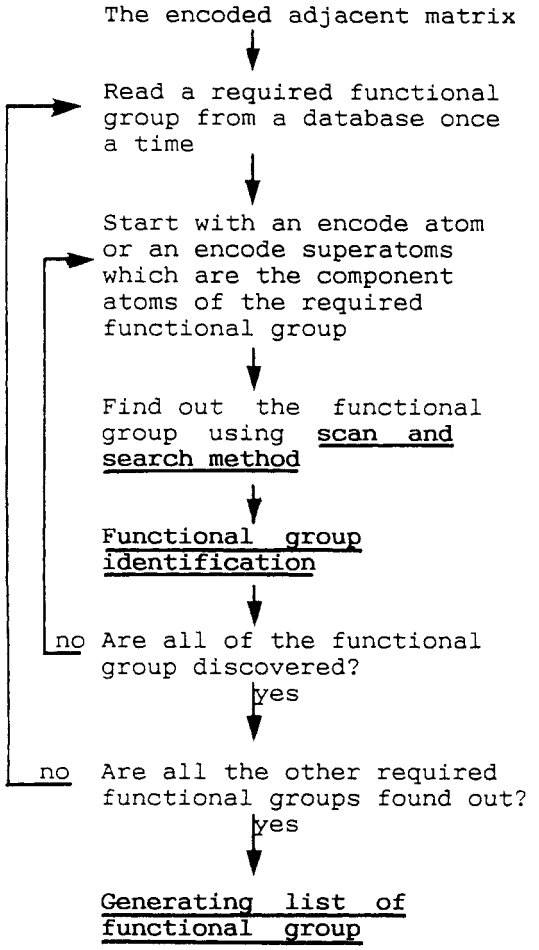

Figure 3. Generation of relevant functional groups.

decomposition method shown in Figure 3 is developed here based on the encoded adjacency matrix. In the figure, the boldface letters indicate the three main steps (scan and search, functional group identification, and generation of the final list).

Scan and Search. To estimate some physicochemical property, the functional group list of some group contribution method is first read from a database in regular order. Once one functional group is removed, a scan down the first column of the encoded adjacency matrix is carried out. In the simplest case, the functional group is found by scanning down the first column. In the general case, if there is no encoded atom or encoded superatoms which are the component atoms of a functional group, the next functional group is read off. Otherwise, the encoded atom or superatoms which are found are selected as the starting points for searching. Searching is carried out on the row of the encoded adjacency matrix where the starting point exists. The purpose of the search is to see whether there is any encoded atom or encoded superatoms forming the functional group with the starting encoded atom or encoded superatom. The searching direction is from left to right along this row, and any non-zero element of the matrix indicates that there is a bond between two encoded atoms or superatoms. There are two cases. The first case is that the functional group cannot be constituted. Then scanning down the first column from the starting point again, the procedure mentioned above is repeated until the end position of the first column is reached. If this functional group is not found in this final scan, then the next functional group of the list should be taken into account. The second case is that a functional group can be found by row searching. Then searching should be continued until all of the functional groups

Table I. Final List of Relevant Functional Groups for Example in Figure 1

no.	main functional group	subfunctional group	no. of relevant functional groups
1	CH_2	CH_2	10
		C	4
2	C≡C	C≡C	4

are discovered. Generally speaking, searching on the encoded adjacency matrix is carried out on the basis of the corresponding row, when only non-zero elements should be examined.

The group contribution method (UNIFAC for liquid–vapor equilibrium) is taken as an illustration of the decomposition process using the previous example.

Through the first procedure described in the previous section, the encoded adjacency matrix is obtained as shown in Figure 2. The list of main and subfunctional groups is read from the developed database in the proper order. For instance the main functional group to be read is CH_2. First, scanning down the first column of the encoded adjacent matrix, it is found that the first row is an encoded atom C, but searching along the first row shows that it is not a component atom of this main functional group. The fifth row is the main functional group CH_2, and searching for the relevant subgroup is carried out along this row. Because only non-zero elements are the objects of the search, this searching is fast. Finally, the subfunctional groups CH_2 and C are detected by searching of the fifth row.

Functional Group Identification. After discovering a functional group using scan and search, the relationship between this functional group to other functional groups has to be determined; this process is called the functional group identification. The functional group identification is also based on the encoded adjacency matrix and the requirement of the correspondent group contribution method. Through this step the relevant functional group is identified.

In the previous example, searching from the fifth row, a main functional group CH_2 is found and identified through the functional group identification routine. Then subfunctional groups are identified. Finally, it is determined from row searching that there is one CH_2–CH_2 and one CH_2–C .

Generating the Final List. Searching and identifying a functional group are repeated until one kind of functional group is clearly identified. Subsequent functional groups are examined until all the required functional groups are identified. Finally, the detailed list of the relevant functional groups, which includes the main and subfunctional groups and their number, is generated. With the detailed list of the functional groups, the corresponding physicochemical property estimation program can be called to calculate the relevant physicochemical property.

A decomposition package which contains several programs for the individual group contribution method is developed to complete this procedure.

In the previous example, through searching all UNIFAC functional groups,[23,24] the final list of relevant functional groups can be obtained as shown in Table I.

APPLICATION OF DECODING SYSTEM

Four chemical compounds with the AES notation, shown in Table II, are used to generate the relevant functional groups in order to calculate their critical constants with the Lyderson group contribution method.[25] By the procedure for generation of an encoded adjacency matrix, these compounds can be determined to have the corresponding structure formulas as shown in Figure 4, and their encoded adjacency matrices are generated by the general decoding program as shown in Figure 5. Then according to the second procedure for generating Lyderson functional groups, the functional relevant groups for each compound are determined by the decomposition program for Lyderson group contribution method based on the encoded adjacency matrix, and the result is indicated in Table III.

Table II. Example for Generating Lyderson Functional Groups

no.	AES notation
1	LJ;/AF/AJ;!CVQ!DG;–:
2	TI;/AE/AI;!BM!HO1;–:
3	TF;/AF;!AV!BM!CV!DM!EV!F2&2;=:
4	G2N<3&2

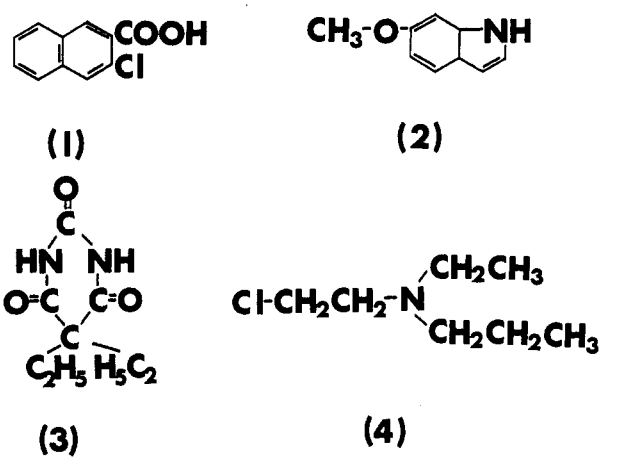

Figure 4. Structural formulas of the examples of Table II.

```
(1)
      C  CH  C  C  CH  C  CH  CH  CH  CH  CO  OH  Cl
C     1              2              1
CH 1     2
C     2     1
C           1     2                                  1
CH                2     1
C  2                 1     1
CH                      1     2
CH                            2     1
CH 1                                2
CO          1                                    1
OH                                         1
Cl                1

(2)
      C  NH  CH  CH  C  CH  CH  C  CH  O  CH3
C     1              2              1
NH 1     1
CH    1     2
CH          2     1
C  2              1     1
CH                      1     2
CH                            2     1
C                                   1     2   1
CH 1                                      2
O                                         1       1
CH3                                             1

(3)
      CO  NH  CO  NH  CO  C  C2H5  C2H5
CO    1              1
NH 1     1
CO          1     1
NH          1           1
CO                1     1
C  1                    1     1     1
C2H5                    1
C2H5                    1

(4)
      Cl  C2H5  N  C3H7  C2H5
Cl    1
C2H51        1
N     1           1     1
C3H7               1
C2H5               1

* The values of remaining matrix elements are zero
```

Figure 5. Encoded adjacency matrix of the examples of Table II.

Table III. Final List of Relevant Functional Groups for Example in Table II

functional groups	no. of relevant groups 1	2	3	4
=CH (in the ring)	6	5		
—C— (in the ring)	4	3	1	
–Cl	1			1
–COOH	1			
—NH (in the ring)		1	2	
–O–		1	2	
$-CH_3$		1	2	
–C=O (in the ring)		1	3	
$-CH_3$ or $-CH_2$			2	7
—N—				1

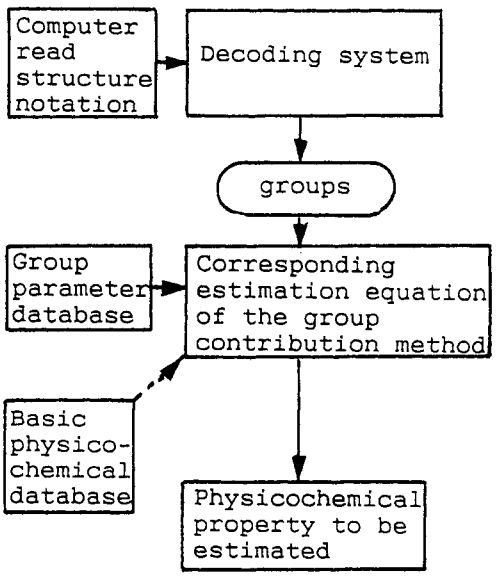

Figure 6. Position and effectiveness of the decoding system in a database.

The decoding system succeeds in decomposing any chemical compound into the individual functional groups. Provided that the parameters of functional groups are read from a group parameter or a basic physical property database, the required property can be estimated by the relevant group contribution methods. The position and the effectiveness of the proposed decoding system in a database can be indicated by Figure 6. From this figure, it is clear that the decoding system is an indispensable part for a physicochemical property database. Therefore, the proposed decoding system can find extensive use in the physicochemical property estimation.

Moreover, the design idea of the decoding system is completed by two basic procedures (generating the encoded adjacency matrix and identifying the relevant functional groups) as described here. For the convenience of users who are not familiar with the AES, the proposed decoding system provides another option which allows the user to input the encoded adjacency matrix instead of the AES structure notation.

CONCLUSION

The proposed decoding system succeeds in converting the structure information of the AES into the relevant functional groups for the correspondent group contribution method.

Illustrating the decoding procedures of examples, it shows that the proposed decoding system is effective and reliable. It can be applied in the group contribution method for the physicochemical property estimation.

REFERENCES AND NOTES

(1) Smith, E. G. *Wiswesser Line-Formula Chemical Notation Method*: McGraw-Hill: New York, 1968; p 77.
(2) 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.
(3) Weininger, D.; Weininger, A.; Weininger, J. C. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
(4) Qu, D.; Fu, B.; Muraki, M.; Hayakawa, T. An Encoding System for a Group Contribution Method. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, preceding paper in this issue.
(5) Vander Stouw, G. G.; Naznitsky, I.; Rush, J. E. Procedures for Converting Systematic Names of Organic Compounds into Atom–Bond Connection Tables. *J. Chem. Doc.* **1967**, *7*, 165–169.
(6) Vander Stouw, G. G.; Elliott, P. M.; Isenberg, A. C. Automated Conversion of Chemical Substance Names to Atom–Bond Connection Tables. *J. Chem. Doc.* **1974**, *14*, 185–194.
(7) Vander Stouw, G. G. Computer Programs for Editing and Validation of Chemical Names. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 232–236.
(8) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, I. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 2. Development of a Formal Grammer. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 106–112.
(9) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, I. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 3. Syntax Analysis and Semantic Processing. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 112–118.
(10) Kirby, G. H.; Lord, M. R.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 6. (Semi)-automatic Name Correction. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 153–160.
(11) Hyde, E.; Matthews, F. W.; Thomson, L. H.; Wiswesser, W. J. Conversion of Wiswesser Notation to a Connectivity Matrix for Organic Compounds. *J. Chem. Doc.* **1967**, *7*, 200–204.
(12) Thomson, L. H.; Hyde, E.; Matthews, F. W. Organic Search and Display Using a Connectivity Matrix Derived from Wiswesser Notation. *J. Chem. Doc.* **1987**, *7*, 204–209.
(13) Hansch, C.; Leo, A. Pomona College Medicinal Chemistry Laboratory, Claremont, CA; Newsletter details WISCT 1980.
(14) Warr, W. A. Diverse Uses and Future Prospects for Wiswesser Line-Formula Notation. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 98–101.
(15) Rosenberg, M. D. Introduction to the Symposium on the Uses and Applications of the Wiswesser Line Notation Today. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 87–88.
(16) Wiswesser, W. J. How the WLN Began in 1949 and How It Might Be in 1999. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 88–93.
(17) Rosenberg, M. D.; Debardeleben, M. Z.; Debardeleben, J. F. Chemical Supply Catalog Indexing: Now and the Future. An Ideal Place for Use of the Wiswesser Line Notation. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 93–98.
(18) Bond, V. B.; Bowman, C. M.; Davison, L. C.; Roush, P. F.; Young, L. F. Applications of the Wiswesser Line Notation at the Dow Chemical Company. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 103–105.
(19) Osinga, M.; Verrijn Stuart, A. A. Documentation Chemical Reaction. IV. Further Applications of WLN Analysis Programs: A System for Automatic Generation and Retrieval of information on Chemical Compounds (AGRICC). *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 26–32.
(20) Eakin, D. R. Graphics Challenge WLN. Can WLN Hold Fast? *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 101–103.
(21) Takeuchi, M. Study on Identification of Chemical Structure. Ph.D. Thesis, University of Tokyo Institute of Technology, Japan, 1990.
(22) Malcolm, B.; Esack, A. Functional Group Discovery Using the Concept of Central Atoms. *Chem. Scripta* **1976**, *9*, 211–215.
(23) Zhao, C. The New Development of UNIFAC Method. *Chem. Eng. (China)* **1984**, *1*, 1–16.
(24) Kojima, K.; Tochigi, K. *ASOG and UNIFAC*; Chemical Engineering: Tokyo, 1987; p 14.
(25) Reid, R. C.; Prausnitz, J. M.; Sherwood, T. K. *The Properties of Gases and Liquids*; McGraw-Hill: New York, 1966; pp 6–33.