

## A Novel Organizational Code for Organic Structures Based on Functional Groups

JAMES F. FEEMAN

Research Department, Althouse Division, Chemicals Group, Crompton  
& Knowles Corporation, Reading, Pennsylvania 19103

Received April 19, 1966

**A novel code for organic structures is described which is easily learned by trained chemists, and which orders compounds within traditional classes according to functionality. The system has been used for efficient arrangement of a chemical storage room as well as in an optical-coincidence chemical information retrieval system.**

In our research laboratory we wished to arrange the chemical storage in a better manner than that afforded by the traditional alphabetical ordering by name. We first considered applying two registry number systems (1, 2) which have been described recently, but were not satisfied with the results.

The *Chemical Abstracts* Registry Numbers did not fulfill our requirements, for they have no obvious meaning and were apparently not designed to produce ordering of chemical structures according to classes or functional groups when listed in increasing numerical order. Also only a relatively small catalog (3) of numbers had been published when our work was initiated and many of the structures we wished to include were not in this catalog.

The Wiswesser BATCH Numbers (2), while readily derived, order organic structures primarily according to skeletal structure and atomic content. Although permutation of these numbers organizes structures according to elementary content, which often coincides with functionality, we wished to emphasize functional groups within structures.

Emphasis on functionality is important where synthetic organic chemistry is the primary concern of the laboratory or chemist involved. The practicing chemist thinks in terms of functional groups attached to skeletons having various shapes and sizes, and is frequently interested in correlation or retrieval by functional group when pursuing solutions to problems involving molecular modification.

Therefore, we developed and are using a novel coding system which organizes by functional elements of structure and which has given us excellent results.

Our code is most useful with small- to medium-sized collections of structures. Each structure is coded as a six-digit number, each digit having a value from zero to nine. These six digits are selected by consideration of: (1), skeletal structure; (2), (3), (4), and (5), location and composition of hetero connectives and functional groups; and (6), carbon count of the structure.

The first (left) digit indicates skeletal structure or ring number, and could be the same as the B in BATCH if desired. In our particular situation it was desirable to modify this digit to reflect the specialized collection of structures in our files. Since we are concerned primarily with textile dye intermediates, we have higher than normal percentages of naphthalene and anthraquinone structures to handle and, therefore, give these two classes special

treatment. Table I shows the value assignments we have chosen for this initial digit.

The next (center) four digits (2, 3, 4, 5) indicate directly the presence (to the maximum of four) of hetero atomic

Table I. First Digit Assignments



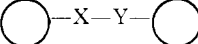


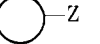
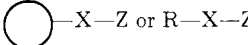
- 0 = no rings in structure; acyclic
- 1 = one benzene ring only in structure
- 2 = a plurality of nonfused benzene rings in structure
- 3 = one ring other than benzene only in structure
- 4 = a plurality of monocyclic rings in structure, at least one of which is not benzene
- 5 = one naphthalene ring in structure
- 6 = a plurality of naphthalene rings or naphthalene ring(s) with monocyclic ring(s) in structure
- 7 = with monocyclic ring(s) in structure bicyclic ring(s) other than naphthalene or bicyclic ring(s)
- 8 = anthraquinone ring(s) or anthraquinone ring(s) with mono- and/or bicyclic ring(s) in structure
- 9 = tricyclic (other than anthraquinone) ring(s) or higher ring(s) in structure

connectives and functional groups which are included within or attached to the basic skeletal structure. These digits and their order in the code are determined by visually scanning the structural formula according to the search order phases listed in Table II beginning with Phase 1. Hetero connectives and functionalities found during each of the search phases listed in Table II are cited in the order determined by their order of precedence established by Table III. After four numbers have been listed, thereby supplying digits 2, 3, 4, and 5, any remaining possibilities are ignored. Search order phases take precedence over functionality; for example, if more than four functions are found under Phase 1 any functions available under Phases 2 through 6 are ignored. As shown in Table III absence of functionality is indicated by zero; if less than four functions are found, the remaining digits are made zeroes to produce a five-digit number.

Digit 6 of the code number is the units digit of the carbon count of the empirical formula of the structure, and is the same as Wiswesser's C in BATCH.

If desired, suffixed arbitrary letters or numbers may be employed to generate a unique code number for each possible structure. In our application of this system we have chosen to suffix a letter which in many cases has an association with the common name of the structure. For example, in the naphthalene series the letters A, G,

Table II. Search Order

Phase	Structural Element to Be Searched <sup>a</sup>
1 Part of ring	X in 
2 Ring-to-ring connective, direct, or through another (nonaliphatic) connective	X in  X and Y in 
3 Ring-to-aliphatic connective, or Ring-to-terminal functional group connective, or Terminal functional group attached to ring	X in  X in  Z in 
4 Aliphatic-to-aliphatic connective	X in R-X-R
5 Aliphatic-to-terminal functional group connective or Terminal functional group attached to aliphatic structure	X in R-X-Z Z in R-Z
6 Other terminal functional group	Z in 

<sup>a</sup> X, Y, and Z represent groups listed in Table III. X and Y represent connectives and Z represents terminal functional groups. The circle is representative of any ring or fused ring system. R represents an aliphatic or substituted aliphatic radical.

Table III. Hetero Connectives and Functional Groups

- 0 none present
- 1  $\text{—NH}_2$  (primary amino)
- 2  $\text{—NH—}$  (in secondary amine or amide)
- 3  $\text{—N=}$  or  $\text{=N—}$  (in tertiary amine, quaternary nitrogen, azo, nitro, nitroso, azine)
- 4  $\text{—OH}$  (hydroxyl)
- 5  $>\text{C=O}$ ,  $>\text{C=NH}$ ,  $>\text{C=S}$ ,  $\text{—C}\equiv\text{N}$
- 6  $\text{—O—}$  (in ether, ester, oxonium)
- 7  $\text{—F}$ ,  $\text{—Cl}$ ,  $\text{—Br}$ ,  $\text{—I}$  (halogen)
- 8  $\text{—S—}$  (in mercaptan, sulfide, sulfoxide, sulfone, sulfonic, sulfonium)
- 9 element other than listed above

H, J, K, and S have been used to terminate the codes for 1:2:4, gamma, H, J, K, and S acids, respectively. Position isomers of the benzene series have been likewise distinguished by suffixing A, B, or C. The additional use of a number or letter to indicate size of Wiswesser Line Notation could also aid generation of unique codes with no arbitrary letters used.

Additional arbitrary procedures we have used include X in position 6 for compounds having unknown or mixed C count (commercial mixtures of natural derivation), and a P suffix separated by hyphen to denote polymeric materials. The latter are assigned numbers from consideration of the monomeric repetitive unit of structure and the end functional groups. Sulfonic and carboxylic acids are coded as complex functional groups having the  $\text{—CO—OH}$  and  $\text{—SO}_2\text{—OH}$  structures; their salts are coded as the free acids.

All segments of ring-ring connectives (Phase 2) are cited according to the order of precedence in Table III. For example,  $\text{—N=N—}$  is coded -33-,  $\text{—CO—NH}$  is -25-,  $\text{—SO}_2\text{NH—}$  is -28-, and  $\text{—O—CO—}$  is coded -56-;  $\text{—NH—CO—C}_6\text{H}_5\text{—CO—NH—}$  is coded -2255-.

Obviously, all six phases of the search order (Table II) do not apply in assigning the code to any particular structure. Aliphatics require use of only Phases 4, 5, and

6. Conversely, ring structures are frequently coded by use of only one or two of the phases, for in many structures less than four functions are present or more than one of the functions is considered under a single phase of the search. When multiple functions are present in more than one phase, those in Phase 1 are all coded before proceeding to Phase 2, and so on.

Frequency of occurrence, synthetic importance, and desirability of proximity of certain compounds with their derivatives on lists dictated our choice of order and definition of the functions listed in Table III. The nitrogen function codes are readily remembered, for they follow the classical designations as primary, "1"; secondary, "2"; and tertiary, "3." The positions of hydroxyl, carbonyl, oxygen, and halogen functions in the order produce very useful, workable listings of carboxylic acids and their common derivatives such as esters, amides, and halides.

The  $>\text{C=O}$ ,  $>\text{C=NH}$ ,  $>\text{C=S}$  structural features are coded as "5" (phase 1) when present in ring structures, except in class 8 (anthraquinone) where the presence of two  $>\text{C=O}$  segments is obvious from the definition of the class. In this case citing "5" for each of the ring  $>\text{C=O}$  segments would serve no useful purpose but would decrease the number of digits available for describing other functions which might be present.

As will be obvious from a study of this system, the code numbers are not unique, but may be readily made so by suffixing an arbitrary letter or number as indicated above. We have coded approximately 2500 structures by these methods and have found little difficulty in generating the codes or in making them unique using only a small portion of the alphabet. The system lends itself to adaptation to any specialized laboratory's internal use. We have arranged our entire organic chemical storage according to this system and have had favorable reactions from the chemists using it. In addition, we have used the code numbers as key words or terms in an optical - coincidence chemical information retrieval system (Termatex-

Jonker Business Machines, Inc., Gaithersburg, Md.) which we have established recently in our laboratories for storage and retrieval of research notes.

While the information in Tables I, II, and III may seem complex initially, a brief study and some experience in assigning codes will rapidly give the user familiarity with the logic of the system and the ability to use the system efficiently. Some typical code assignments are shown in Table IV.

Procedures and mental processes involved in the assignment of code numbers will be clarified by consideration of the detailed examples which follow.

## EXAMPLES OF ASSIGNMENT OF CODE NUMBERS

1. Diethanolamine,  $\text{HOCH}_2\text{CH}_2\text{—NH—CH}_2\text{CH}_2\text{OH}$ , is assigned the code number 024404. The first digit is "0" (from Table I) since no ring structure is present. Following the search order of Table II, Phases 1, 2, and 3 do not apply because no ring is present. Under Phase 4 the  $\text{—NH—}$  group is an aliphatic-to-aliphatic connective, so digit 2 is "2" (from Table III). Proceeding to Phase 5 of Table II, the  $\text{—OH}$  groups are terminal functional groups attached to aliphatic structures. Therefore, digits 3 and 4 are both "4" representing these  $\text{—OH}$  groups.

Table IV. Some Typical Code Numbers

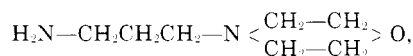
010001 Methylamine	055112 Oxamide
010002 Ethylamine	055442 Oxalic acid
010004 Butylamine	057102 Chloroacetamide
011002 Ethylenediamine	057402 Chloroacetic acid
011003 Propylenediamine	057702 Chloroacetyl chloride
011005 Pentanediamine	057742 Dichloroacetic acid
011006 Hexanediamine	057772 Trichloroacetic acid
014002 Aminoethanol	077701 Chloroform
014003 Aminopropanol	080002 Dimethyl sulfoxide
014004 Aminobutanol	110006 Aniline
014404 Aminomethylpropanediol	110007 Toluidine
014405 Aminoethylpropanediol	110008 Xylidine
014444 <i>tris</i> -(Hydroxymethyl)-aminomethane	111006 Phenylenediamine
015402 Glycine	111306 Nitrophenylenediamine
015544 Aspartic acid	111607 Methoxyphenylenediamine
020002 Dimethylamine	112508 Aminoacetanilide
020004 Diethylamine	113006 Nitroaniline
020006 Dipropylamine	113306 Dinitroaniline
021001 Methylhydrazine	113357 Dinitroanthranilic acid
021104 Diethylenetriamine	113406 Aminonitrophenol
021404 <i>N</i> -(2-Hydroxyethyl)-ethylenediamine	113476 Aminochloronitrophenol
022116 Triethylenetetramine	114006 Aminophenol
024404 Diethanolamine	154007 Benzoic acid
025102 <i>N</i> -Methylurea	156009 Ethyl benzoate
025104 <i>N</i> -Acetyleneethylenediamine	157007 Benzoyl chloride
025152 Dicyandiamide	158447 Sulfobenzoic acid
025506 Iminodipropionitrile	158847 Disulfobenzoic acid
028403 <i>N</i> -Methyltaurine	160008 Phenetole
030006 Triethylamine	165707 Phenylchloroformate
031005 Dimethylaminopropylamine	165708 Phenoxyacetyl chloride
033006 Tetramethylethylenediamine	197706 Phenylphosphonic dichloride
034004 Dimethylaminoethanol	211002 Benzidine
034446 2,2',2''-Nitrilotriethanol	225004 Acetaminobiphenyl
040001 Methanol	225103 Aminobenzanilide
040002 Ethanol	355006 Benzoquinone
040003 Propanol	335006 Vinylpyrrolidone
040004 Butanol	336107 <i>N</i> -(3-Aminopropyl)-morpholine
040008 Octanol	433500 Phenylmethylpyrazolone
045403 Lactic acid	440002 Cyclohexylphenol
045408 Ricinoleic acid	514840 Aminohydroxynaphthalenesulfonic acid
045544 Malic acid	518400 Aminonaphthalenesulfonic acid
047002 Chloroethanol	548840 Hydroxynaphthalenedisulfonic acid
047703 Dichloropropanol	628406 Phenylaminonaphthalenesulfonic acid
048002 Mercaptoethanol	656403 Phenyl salicylate
048402 2-Hydroxyethanesulfonic acid	738008 Methylbenzothiazole
050001 Formaldehyde	760008 Styrene oxide
050003 Acetone	817004 Aminobromoanthraquinone
050004 2-Butanone	844004 Alizarin
054001 Formic acid	956003 Xanthone
054002 Acetic acid	

There is no other function present, so digit 5 is "0", and digit 6 is "4" (representing four carbon atoms in the empirical formula of the compound).

2. Glycine,  $\text{H}_2\text{N}-\text{CH}_2-\text{CO}-\text{OH}$ , is assigned the code 015402. The structure is acyclic (digit 1 = 0). From Table II, Phases 1, 2, 3, and 4 do not apply since no ring is present and there is no aliphatic-to-aliphatic connective. Proceeding to Phase 5, the  $-\text{NH}_2$  is terminal functional on aliphatic, and the  $-\text{CO}-$  is an aliphatic-to-terminal functional connective. Since they have equal rank in Phase 5, they are ordered according to their position on Table III, the  $-\text{NH}_2$  taking precedence. This produces digit 2 = "1" and digit 3 = "5." The remaining  $-\text{OH}$  is an additional terminal functional group (Table II, Phase 6) and is coded as "4" at digit 4. No other functions are present so digit 5 = "0," and digit 6 = "2" because two carbons are present.

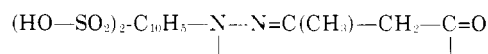
3. Aminochloronitrophenol,  $\text{C}_6\text{H}_3(\text{Cl})(\text{NO}_2)(\text{NH}_2)(\text{OH})$ , is assigned the code 113476 from the following facts: Digit 1 = "1" (from Table I) for there is a single benzene ring present. The four terminal functional groups are all associated with Phase 3 of Table II and from Table III their relative rank is determined to be  $-\text{NH}_2$ ,  $-\text{NO}_2$ ,  $-\text{OH}$ , Cl. Therefore, digits 2, 3, 4, and 5 are "1" (for  $-\text{NH}_2$ ), "3" (for  $\text{NO}_2$ ), "4" (for  $-\text{OH}$ ), and "7" (for Cl). The carbon count is 6.

4. Aminopropylmorpholine,



is assigned the code number 336107. Digit 1 = "3" (from Table I) because the structure contains one ring which is not a benzene ring, but is monocyclic, and contains no other rings. Phase 1 of Table II reveals a tertiary N atom and an ether oxygen which are part of the ring structure. Since the  $=\text{N}-$  comes first in Table III it is cited before the  $-\text{O}-$ , digit 2 = "3" (for  $=\text{N}-$ ) and digit 3 = "6" (for the  $-\text{O}-$ ). Phases 2, 3, and 4 do not apply in this example, but in Phase 5 an  $-\text{NH}_2$  group is present as a functional group attached to an aliphatic structure. This gives digit 4 = "1" (for  $-\text{NH}_2$ ). There are no other functions present; therefore, digit 5 = "0", and the carbon count is 7 (digit 6).

5. Amino J pyrazolone,  $\text{C}_{14}\text{H}_{12}\text{N}_2\text{O}_7\text{S}_2$ , is assigned the code number 633584. Digit 1 = "6" (from Table I) indicating the presence of a



naphthalene ring plus a monocyclic ring. The two "3"s and the "5" (from Phase 1, Table II) indicate the two  $=\text{N}-$  atoms and the  $=\text{C}=\text{O}$  function which are part of the ring structure. The "8" (from Phase 3, Table II) indicates an  $-\text{S}-$  function ( $-\text{SO}_2$ ) which is a ring-to-terminal functional group connective. Finally, the "4" is the units digit of the carbon count from the empirical formula. The additional  $-\text{SO}_2-$  and  $-\text{OH}$  functions present cannot be coded because of the four-digit limitation and are, therefore, ignored.

Certain fine details remain to be clarified, but the system has proved most useful in its present form. Modifications to fit local conditions of chemical specialization may be made quite readily where required. These code numbers are also adaptable to punched card methods. They should be usable for the preparation of indexes which would then furnish answers to questions concerning presence or absence of functional groups in files of coded structures. For such uses it would be desirable to cite all functions present rather than to restrict the field to four digits. We expect to investigate these aspects in the future, as well as the possible use of the Wiswesser Line Notation code letters as replacements for the numbers in Table III. This departure might yield a nonunique, but nevertheless useful code having value for correlation or retrieval by functional group with better discriminatory power than that described here.

#### ACKNOWLEDGMENT

The author wishes to thank his colleagues, G. F. Garcelon, R. H. Horning, R. E. Eltonhead, and M. C. Bernier for helpful suggestions and constructive criticism, W. J. Wiswesser for encouragement and advice, Mrs. E. H. Larreau for assistance in establishing the optical-coincidence system, and Mary Ellen Beitzel for technical assistance during the summer of 1965.

#### LITERATURE CITED

- (1) Leiter, D. P., Jr., Morgan, H. L., Stobaugh, R. E., *J. Chem. Doc.*, **5**, 238 (1965).
- (2) Barnard, A. J., Jr., Kleppinger, C. T., Wiswesser, W. J., *ibid.*, **6**, 41 (1966).
- (3) Chemical Abstracts Service, "SOCMA Handbook, Commercial Organic Chemical Names," American Chemical Society, Washington, D. C., 1965.