case, however, we observe a sudden spurt of interest in the structural properties of binary fluorides following a major discovery some steps removed. The interest, as evidenced by the rate of publication, has been sustained ever since. This observation exemplifies the wide and profound influence Bartlett's discovery has had. It is a noteworthy correlation of the stimulating effects of an exciting landmark discovery in chemistry.

REFERENCES

(1) D. T. Hawkins, L. S. Bernstein, W. E. Falconer, and W. Klemperer, "Binary Fluorides: Free Molecular Structures and Force Fields. A Bibliography 1959-1975", NSRDS Bibliographic Series, IFI/Plenum, New York, N.Y., 1976, 238 pp.

(2) N. Bartlett, "Xenon Hexafluoroplatinate", *Proc. Chem. Soc.* (*London*), 218 (1962).

(3) B. C. Bennion and L. A. Newton, "Epidemiology of Research on "Anomalous Water"", *J. Am. Soc. Inf. Sci.*, **27**, 53–56 (1976).

An Efficient Design for Chemical Structure Searching. III. The Coding of Resonating and Tautomeric Forms[1]

ALFRED FELDMAN

Walter Reed Army Institute of Research, Washington, D.C. 20012

Received February 10, 1976

A simple method for coding resonating and tautomeric structures is described. The method does not require the introduction of artifical bonds and facilitates the recovery of the original input structures.

1. INTRODUCTION

In retrieval systems using atom-by-atom codes, chemical structures are treated as finite graphs. Classical valence bonds and atoms are easily equated with edges and nodes. Exceptions are the hybridized bonds of resonating groups and the shifting structures of tautomers. These are not so readily equatable and risk complicating an otherwise simple coding process.

The difficulty might be avoided by the heroic means of generating, coding, and storing all possible resonance and tautomer isomers of the input compounds, or by having every submitted query undergo a similar transformation. Such a solution would greatly burden processing and storage facilities, and thus far no system has implemented it.

Most current systems have solved the problem by creating artificial bonds to replace the hybrid and shifting ones. This requires a number of new bond types, for example, "delocalized", "alternating", and "tautomeric" bonds.[2] Using these, structures can be normalized.

The use of artificial bonds does not represent an ideal solution. The determination, in a search, as to which of these bonds are equivalent to which, requires more complex programming and lengthier processing. Furthermore, the replacement of the hybrid and shifting bonds by normalized ones results in a loss of information. No longer will it be possible to recover the structure originally used for input. For this reason, normalization methods must be used with caution. For example, the most common type of tautomerism, the keto–enol tautomerism, is not normalized by *Chemical Abstracts*.

2. GOPPELT CODES

The coding method, implemented in the WRAIR system since *ca.* 1962, avoids the need for artificial bonds in resonating structures. The method, developed by Richard Goppelt, is based on the use of a connection table which is insensitive to differences among optional representations of alternating single and double bonds.

Conventional connection tables for chemical structures normally consist of three lists: one for the atoms, one for the bonds, and one for the connections. Figure 1 shows a connection table obtained according to Morgan's method.[3] The connection table developed by Goppelt differs in that only two lists are present. The information on atoms and bonds is combined. For each atom, the element code is followed by an inventory of all the bonds, charges, hydrogen atoms, etc., that are attached to it (Figure 2). In this inventory, only classical valence bonds are recognized.

From Figure 3, it is seen that this method produces identical codes for the two resonating isomers of benzene. Furthermore, the fragments shown in Figure 4 will match these benzene isomers. This is because, in a substructure search, the categories DBN, SBN, HCT, etc., of the atom codes of the query are matched inclusively with the corresponding codes of the file compounds. That is, the C on line 2 in Figure 3 is a match for the C on line 2 in Figure 4a, as the former possesses at least two SBN. It will also be a match for the C on line 2 in Figure 4b, because it possesses at least one DBN and one SBN.

The original Goppelt method could not take odd ring atoms into the resonating system (Figure 5). Further, it could not distinguish certain cases of known isomerism, e.g., the cyclooctatetraenes. The extension of Goppelt codes to other classes of shifting structures, described below, suggested, however, a solution to these deficiencies.

3. ADAPTATION OF GOPPELT CODES TO TAUTOMERS

In the redesigned WRAIR system,[4] Goppelt codes have been adapted to tautomers, avoiding the need for artificial tautomer bonds in the connection table.

Prior to applying Goppelt codes to tautomers, the tautomers must be detected. This is accomplished by means of appropriate algorithms (Figure 6). An effort was made to keep these algorithms similar to those used by CAS.

The tautomer detecting algorithms each define a central atom (Q) and two terminal atoms (M and Y for algorithm A; N and Z for algorithm B). Codes are made insensitive to tautomeric variations in structure by overcoding these terminal atoms. The necessary rules are shown in Figure 7.

For example, in the structure shown in Figure 8, the terminal atom on line 1 has its categories SBN and HCT incremented each by one. The terminal atoms on lines 3 and 7 are similarly treated. The terminal atom on line 9 has its category DBN incremented by one. The central atoms, located on lines 2, 4, and 8, are not altered.

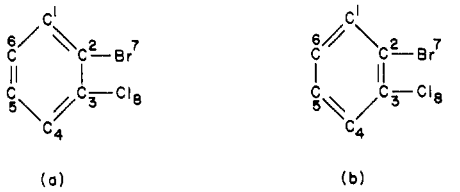
As a result of the application of the rules of Figure 7, different tautomers of a compound will have identical codes.

Rank number (not coded)	Node Value (Atoms)	Line Value (Bonds)	From Attachment (Connections)
1	R	blank	blank
2	C	1	1
3	N	1	2
4	C	2	2
5	C	1	4
6	N	1	3
		2	5-6 (ring closure)

Figure 1. Example of conventional connection table. The compound coded is the one shown in Figure 12a.

Symbol	Interpretation	Number of bits required
ELM	Element code	7
VAL	Valence	3
TBN	Number of triple bonds	1
DBN	Number of double bonds	2
SBN	Number of single bonds	3
HCT	Hydrogen count	3
CHG	Number of ionic charges	3
NEG	Nature of ionic charges (+ or -)	1
RNG	Member of a ring	1
Etc.		

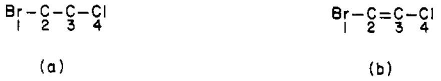
Figure 2. Categories represented in Goppelt atom codes.



NO.	ELM	DBN	SBN	HCT
1	C	1	1	1
2	C	1	2	
3	C	1	2	
4	C	1	1	1
5	C	1	1	1
6	C	1	1	1
7	Br		1	
8	Cl		1	

NO.	ELM	DBN	SBN	HCT
1	C	1	1	1
2	C	1	2	
3	C	1	2	
4	C	1	1	1
5	C	1	1	1
6	C	1	1	1
7	Br		1	
8	Cl		1	

Figure 3. Goppelt codes for the two resonance isomers (a and b) of benzene. Only the atom codes are shown here. The connections are omitted.



NO.	ELM	DBN	SBN	HCT
1	Br		1	
2	C		2	
3	C		2	
4	Cl		1	

NO.	ELM	DBN	SBN	HCT
1	Br		1	
2	C	1	1	
3	C	1	1	
4	Cl		1	

Figure 4. Goppelt codes for two fragments that match the resonance isomers in Figure 3. Only the atom codes are shown.

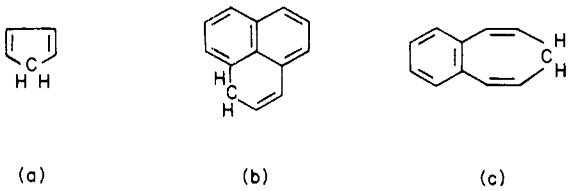
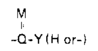


Figure 5. Nonresonating atoms in the original Goppelt codes.

Algorithm A:



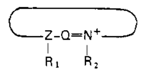
where:

M = O, S, Se or Te

Y = O, S, Se or Te

Q = As, Br, Cl, I, N, P, S, Sb, Se, Te

Algorithm B:



where:

N = Nitrogen atom with a positive charge and a valency = 5

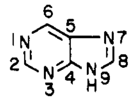
Z = Nitrogen atom with normal valency

Q = N or C

Figure 6. Algorithms for recognizing tautomers.

```
                    Rules for coding tautomer atoms

Algorithm A:

1.  In the connection table, identify the atoms corresponding to M and Y.

2.  For the atom corresponding to Y:
      - increment category DBN by 1
      - enter the value 2 (binary) into category TAU.

3.  For the atom corresponding to M:
      - increment category SBN by 1
      - enter the value 1 into category TAU.
      - if atom Y had a hydrogen attached:
        increment category HCT by 1
      - if atom Z had a negative charge:
        increment categories CHG and NEG each by 1.

Algorithm B:

1.  In the connection table, identify the atoms corresponding to N and Z.

2.  For the atom corresponding to N:
      - increment category SBN by 1
      - enter the value 1 into category TAU.

3.  For the atom corresponding to Z:
      - increment categories DBN and CHG by 1
      - enter the value 2 (binary) into category TAU.
       - raise valency to 5.
```

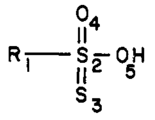
Figure 7. Rules for coding tautomer atoms.



NO.	ELM	DBN	SBN	HCT	TAU	
1	N	1	2	1	1	
2	C	1	1	1		(Q atom)
3	N	1	2	1	1	
4	C	1	2			(Q atom)
5	C	1	2			
6	C	1	1	1		
7	N	1	2	1	1	
8	C	1	1	1		(Q atom)
9	N	1	2	1	2	

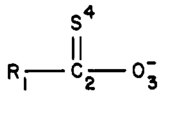
Figure 8. Multiple tautomer groups.

In an identity search, such tautomers will match, and in a substructure search, the condition of inclusion will be satisfied as it was in the example of Figures 3 and 4. The examples shown in Figures 9–12 indicate that the present method of coding is of fairly general applicability.



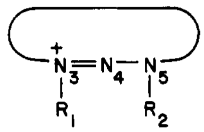
NO.	ELM	DBN	SBN	HCT	TAU	
1	R		1			
2	S	2	2			(Q atom)
3	S	1	1	1	1	
4	O	1	1	1	1	
5	O	1	1	1	2	

Figure 9. Multiple tautomer groups.



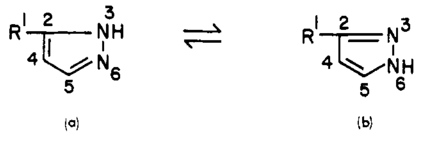
NO.	ELM	DBN	SBN	CHG	NEG	TAU	
1	R		1				
2	C	1	2				(Q atom)
3	O	1	1	1	1	2	
4	S	1	1	1	1	1	

Figure 10. Migrating negative charge.



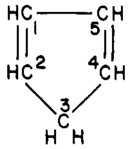
NO.	ELM	VAL	DBN	SBN	CHG	TAU	
1	R	1		1			
2	R	1		1			
3	N	5	1	3	1	1	
4	N	3	1	1			(Q atom)
5	N	5	1	3	1	2	

Figure 11. Migrating positive charge (onium).



NO.	ELM	DBN	SBN	HCT	TAU
1	R		1		
2	C	1	2		
3	N	1	1	1	2
4	C	1	1	1	
5	C	1	1	1	
6	N	1	2	1	1

Figure 12. "Extended" tautomerism. Although no algorithm is available at present to detect this type of tautomerism, the coding method can handle it, assuming that atoms 3 and 6 correspond respectively to atoms Y and M.



NO.	ELM	DBN	SBN	HCT	TAU
1	C	1	1	1	
2	C	1	1	1	
3	C	1	1	1	
4	C	1	1	1	
5	C	1	1	1	

Figure 13. Code for resonance in uneven membered rings.

The above method of overcoding can be applied again to resonance, to code the odd atom (Figure 5) that did not participate in resonance under the earlier procedure. The participation of this atom can be obtained by incrementing, by one, its category DBN, and HCT (Figure 13). Thereafter, a substructure search for the fragment$=C-CH_2-C=$ will no longer match an odd element ring system, which is as it should be.

Organic acids represent another area where overcoded Goppelt codes might be used. It is often desired to retrieve acids, regardless of whether they were originally coded in their free form, or as salts. These acids could be retrieved on a substructure search, but this would bring back their esters as well. An alternative (not implemented in the WRAIR system) is to determine acid salts by a detection algorithm, then to overcode the acidic oxygen by incrementing its HCT by one.

Other classes of compounds may lend themselves to similar techniques.

4. RETENTION OF ORIGINAL STRUCTURES

Insensitivity to structural variation, as obtained by the above method, may not always be desired. Tautomers defined by algorithm are to some extent arbitrary, as were the resonance isomers. For example, the above method is unable to handle ring-chain tautomerism, nor can it manage anionotropy, where an atom other than hydrogen shifts position.

In attempting to deal with complex characteristics such as resonance and tautomerism, one could strive to improve the accuracy of the method, steadily decreasing the number of compounds falsely interpreted. Alternatively, and provided that the number of retrieved structures remains manageable, the goal might be to keep the definitions, by means of which the machine recognizes tautomers and resonating structures, as simple as possible. The rationale, with the latter philosophy, is to allow the average user to predict the arbitrary behavior of the mechanical system. By judiciously phrasing his queries and by properly interpreting the retrieved data, he can then deal intelligently with the system's limitations.

The latter philosophy being subscribed to, the following two considerations have governed the design:

a. The definitions of resonance and tautomerism should be kept simple and readily understandable to the casual user, even though the definitions may lead occasionally to arbitrary results.

b. Where the definitions of the system are objectionable, the user should have the alternative of basing his search on structures used originally for input.

It was possible to implement these considerations economically. The recovery of original structures was made possible through the use of flags. A category TAU was added to the inventory table carried by each element (Figures 8–12). Two bits were allocated to this flag. The rules in Figure 7 indicate when and how this flag is set. The recovery of a tautomeric input structure is obtained by using these rules in reverse. (Where the original structure is no longer known, e.g., if the input was obtained from a system in which structures are normalized, this can be indicated by placing the number 3 into the TAU category).

The original structures of other types of compounds can be recovered in a similar manner. For example, cyclic dienes can be specified by retaining the location of one original double bond for each ring system. The odd ring atom (Figure 5) can be specified by a special flag, comparable to TAU. But because there was little need for these additional recovery capabilities, they are not available in the WRAIR system.

For retrieving compounds with shifting structure, a user of the WRAIR system thus has the option of searching either under relaxed or under stringent identity rules. There is thus less need for caution in the application of tautomer codes than there is with the use of normalized bonds. Extension of the above method to keto–enol tautomers may consequently be considered in the future.

5. ACKNOWLEDGMENT

I am indebted to Ted N. Fujimoto for programming support and to Ms. Margaret Milne, of the University of Pennsylvania, for her invaluable comments and suggestions.

REFERENCES AND NOTES

(1) L. Hodes and A. Feldman, "An Efficient Design for Chemical Structure Searching. II. A Solution to the Large Data Base File Structure Dilemma", in preparation.

(2) G. W. Adamson, M. F. Lynch, and W. G. Town, "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. II. Atom-Centred Fragments", *J. Chem. Soc. C*, 3702 (1971). D. J. Gluck, "A Chemical Structure Storage and Search System Developed at Du Pont", *J. Chem. Doc.*, **5**, 43 (1965); M. Milne, D. Lefkovitz, M. Plotkin, H. Hill, and R. Powers, "A Study of the Potential for Useful Exchange in Chemical Structure Handling Areas of Four Chemical Information Systems", and references cited therein, available from NTIS, AD 782, 239.

(3) H. L. Morgan, "The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service", *J. Chem. Doc.*, **5**, 107 (1965).

(4) A. Feldman and L. Hodes, "An Efficient Design for Chemical Structure Searching. I. The Screens", *J. Chem. Inf. Comput. Sci.*, **15**, 147 (1975).

An Interactive Computer Graphics System for Processing Chemical Structure Diagrams[†]

JAMES E. BLAKE, NICK A. FARMER,* and REGINALD C. HAINES

Chemical Abstracts Service, P.O. Box 3012, Columbus, Ohio 43210

Received February 3, 1977

An interactive computer graphics system has been developed at Chemical Abstracts Service (CAS) which allows creation of chemical structure diagrams of a quality suitable for publication in *Chemical Abstracts* (CA). The system is based on a Digital Equipment Corporation (DEC) PDP-15 graphics system connected to an IBM 370/168 host computer. Employing a light pen, the user at a terminal creates a structure diagram by selecting individual atoms or chemical rings from a menu list and by indicating how the various pieces are to be connected. The computer program then calculates the preferred placement of the structural pieces according to a set of formatting rules. The user at the terminal can, however, override these rules by moving items on the screen with the light pen. Once input, structure diagrams are stored on disk (or tape) at the IBM 370 until required for insertion into a specific publication. Then text and graphic data are selected from the IBM 370 database and photocomposed on an Autologic APS-4, producing full-page publication quality output. This automated system has been in daily use for over two years. Two reference files have been built in this manner. One file includes basic chemical ring shapes and currently contains over 19 000 entries. The second file, which currently contains over 42 000 structure diagrams, is the source for the structure diagrams that routinely appear in the CA Volume and Collective Chemical Substance Indexes and in the Parent Compound Handbook. The system is also used to produce the structure diagrams that appear in the weekly issues of *Chemical Abstracts*.

INTRODUCTION

Each year Chemical Abstracts Service (CAS) publishes over 59 000 pages of chemical information, consisting mainly of the weekly issues of CA and its volume indexes, which are published every six months. (These page number figures do not include the Collective Indexes to CA which are published every five years.) These publications represent over 800 000 000 characters of textual information and more than 62 000 chemical structure diagrams. Although the first steps toward the computer-controlled composition of CAS publications began in 1965, it was not until 1971 that attention was directed toward the computer processing of graphical data. The computer-composed publications had "windows" in the text where the graphics were to appear. Graphics were drawn by illustrators, photoreduced, and then stripped (pasted) into the page containing the text that had been photocomposed. Over the past five years CAS has developed a system that provides the capability for computer composition of graphical as well as textual data. This paper describes a part of that system. The first section describes the input sub-system, its hardware and software structure, and some of the basic system characteristics. The next section describes three applications which process chemical structure diagrams that have been implemented using this system. The last section summarizes our experience gained from using this system and describes our plans for future work in this area.

There have been several computer graphics systems developed over the last few years that process chemical structure diagrams. The system developed by Corey and Wipke is used for computer-assisted synthesis of organic compounds.[1] The PROPHET system, developed by Bolt, Beranek, and Newman, Inc., is used for information retrieval in a medical environ-

[†] Presented in part at the 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 8, 1975.

* Author to whom correspondence should be addressed.