1978

The structural molecular formula: extending the molecular weight - molecular formula series

R. Geoff Dromey
University of Wollongong

THE STRUCTURAL MOLECULAR FORMULA: EXTENDING THE MOLECULAR WEIGHT - MOLECULAR FORMULA SERIES

R. Geoff Dromey

Department of Computing Science, University of Wollongong

P.O. Box 1144, Wollongong, N.S.W. 2500, Australia

ABSTRACT

There is a pressing need for simple, flexible and efficient techniques for substructural searching of large data bases. The structural molecular formula described, represents a logical extension of the molecular weight and molecular formula descriptions of molecular topology. It is shown to come close to structure diagrams in structural resolution and is therefore an ideal tool for screening large data bases.

There are at least three ways of characterizing an atom that are relevant to molecular chemistry. The first and most primitive is in terms of the mass of the atom. The second and somewhat more useful description employs elemental identities (e.g., carbon C, nitrogen N). A third and still more precise description of an atom, is in terms of its structural environment.

Chemistry has evolved a number of conventions for describing molecular structure. The simplest of these representations is the molecular weight. This description contains no obvious information about molecular structure. The molecular formula is more explicit than the molecular weight in that it identifies the number and elemental identities of all atoms in a molecule. However this description of molecular structure, like the molecular weight, only conveys information about molecular structure in a negative sense (e.g., we know the molecule described by the composition C_2H_2 cannot contain a six-membered aromatic ring). A chemically more accurate description of an atom than its elemental identity is its structural identity. (e.g., a carbon atom in an aromatic ring is chemically different from a carbon atom at the end of an acyclic chain.

Examining the conventions for representing molecular structures at the formula level we find that there are representations for atoms at the atomic mass and elemental identity levels but that there is no formalism that accommodates the structural identities of atoms. In the discussion which follows such a formalism is described.

THE ATOM STRUCTURAL IDENTITY

The Structural Identity of an atom can be defined operationally as its structural environment with respect to its location in a ring or chain.

The alphabetic character set provides an adequate set of structural identity descriptors for distinguishing among the most common, and chemic important, atom environments. (e.g., atoms at the ends of chains and other chain environments, atoms in, and directly connected to rings of various sizes, ring positions of structural significance such as fused, perifused, and spiro atoms etc.).

The 26 structural identities and bond types that are used to derive what we shall call the structural molecular formula description of molecular structures are listed in Table I. An effort has been made to have the alphabetic structural identifiers correspond with the first character of simple structural mnemonics or structural templates (e.g., "A" identifies an aromatic ring atom, "X" identifies a tertiary branching environment, and "Y" a secondary branching environment.)

THE STRUCTURAL MOLECULAR FORMULA

The structural molecular formula for a molecule is derived by simply identifying and making a count of the structural identities of all atoms in the molecule. All atoms in the molecule must be structurally identified and assigned to make the representation complete and non-overlapping.

Structural molecular formulas are represented by sets of structural descriptors for each elemental identity. The general format for structural descriptors is

<Structural Identity><Atom Symbol(s)><Atom count>

The atom counts are represented as subscripts in keeping with the standard convention for molecular formulas. For example the structural de

TABLE I

Descriptor	Structural Identity
A	Aromatic ring atom
B	Ring atom at end of bridge
C	Substituent atom directly connected to 3-membered ring
D	Substituent atom directly connected to 4-membered ring
E	Terminal atom - at end of chain
F	Fused ring atom - only for two rings
G	Substituent atom directly connected to 5-membered ring
H	Substituent atom directly connected to ring of more than 6 atoms
I	Atoms in a 3-membered ring
J	Atoms in a 4-membered ring
K	One of a pair of substituent atoms attached to same ring atom
L	Atoms in a 5-membered ring
M	Atoms in a 7-membered ring
N	Atoms in a ring of size greater than 7
O	Co-ordination bond
P	Peri-fused ring atom - involving 3 rings
Q	Spiro ring atom - connected to 4 other ring atoms
R	Six-membered carbocyclic ring atom
S	Substituent atom directly connected to an aromatic ring
T	Substituent atom directly connected to 6-membered carbocyclic ring
U	Double bond - (not in aromatic ring)
V	Substituent atom attached to a fused atom
W	Triple bond
X	Chain atom with 4 non-hydrogens attached
Y	Chain atom with 3 non-hydrogens attached
Z	Atom in a acyclic chain (non-terminal)

FIGURE 1

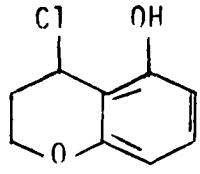
(a) $CH_3CH_2CH_2OH$ — $EC_1EH_4EO_1ZC_2ZH_4$

(b) $CH_3\text{-}O\text{-}CH_2CH_3$ — $EC_2EH_6ZO_1ZC_1ZH_2$

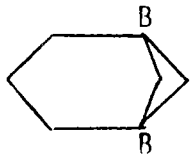
(c) CH_3CH_2CHO — $EC_1EH_3ZC_2ZH_3EO_1EU_1$

(d) $CH_3\text{-}\overset{\overset{O}{\|}}{C}\text{-}CH_3$ — $EC_2EH_6YC_1EO_1EU_1$

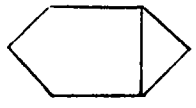
(e) $CH_3CH_2\text{-}\overset{\overset{O}{\|}}{C}\text{-}OH$ — $EC_1EH_4EO_2YC_1EU_1ZC_1ZH_2$

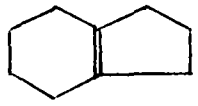
(f) 

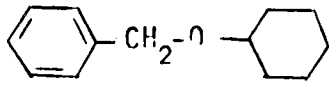
— $AC_4AH_3FC_2RC_3RO_1TCl_1RH_5SO_1SH_1$

(g) 

— $RC_3RH_6BC_2BH_2JC_2JH_4$

(h)  — $IC_1IH_2FC_2FH_2LC_3LH_6$

(i)  — $LC_3LH_6FC_2FU_1RC_4RH_8$

(j) 

— $AC_6AH_5SC_1SH_2RC_6RH_{11}TO_1$

FIGURE 2

C_6H_8 ISOMERS

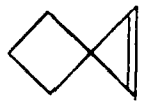
(a) $CH_3CH_2-C\equiv C-CH=CH_2$ — $EC_2EH_5ZC_4ZH_3EU_1ZW_1$

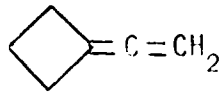
(b) $CH_3CH_2-CH=CH-C\equiv CH$ — $EC_2EH_4ZC_4ZH_4ZU_1EW_1$

(c) $CH_2=CH-CH=CH-CH=CH_2$ — $EC_2EH_4ZC_4ZH_4EU_2ZU_1$

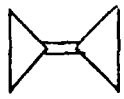
(d) $CH_2=C(-CH=CH_2)-CH=CH_2$ — $EC_3EH_6ZC_2ZH_2YC_1EU_3$

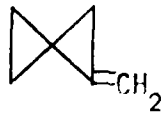
(e) $CH_2=CH-CH(-CH_3)-C\equiv CH$ — $EC_3EH_6ZC_2ZH_1YH_1YC_1EU_1EW_1$

(f)  — $IC_2IH_2IU_1QC_1JC_3JH_6$

(g)  — $JC_4JH_6DU_1DC_1EC_1EH_2EU_1$

(h)  — $IC_2IH_4FC_4FH_4$

(i)  — $IC_6IH_8CU_1$

(j)  — $IC_4QC_1IH_6CU_1CC_1CH_2$

AC_6 indicates that there are a total of six aromatic carbons in a molecule while FC_2 identifies the presence of two fused carbons.

A sample set of structural molecular formulas is given in Figures 1 and 2. Studying the structural molecular formula of figure (1f) we observe that it preserves almost all the structural information about the molecule except for the relative positions (in the respective rings) of the two substituents and the heteroatom. The $AC_4AH_3FC_2$ part of the formula tells us that there is one aromatic ring present which is fused to another ring. The AH_3 rather than AH_4 tells us that there is one hydrogen missing and so the aromatic ring must be substituted. The SO_1SH_1 tells us that there is one oxygen substituted on the aromatic ring and that it has a hydrogen attached. The RC_3RO_1 tells us that the ring fused to the aromatic ring is a six-membered carboxyclic ring with an oxygen heteroatom. The RH_5TCl_1 tells us that there is a chlorine substituted on the carbocyclic ring and the RH_5 rather than RH_6 confirms this fact.

In figure 2 structural molecular formulas for five acyclic and five cyclic C_6H_8 isomers are given. The fact that each of the structures has a different formula suggests that the formalism has considerable power for structural differentiation among isomers. This example is by no means meant to imply that the structural molecula formula provides a unique representation for any complete set of isomers. For example the structural formula for the compound

$$CH_3\text{-}CH{=}CH\text{-}CH_2\text{-}C{\equiv}CH$$

is identical with that for structure 2(b). That is the relative position of bonds within the same structural environment are not distinguished. However whenever the structural molecular formulas for two compounds

are the same the compounds are strikingly similar structurally.

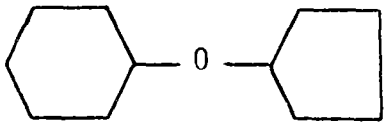
ENCODING RULES FOR STRUCTURAL MOLECULAR FORMULAS

With the aid of Table I assigning structural identities to atoms is self evident in almost all cases. The following precedence rule resolves possible conflicts of assignment.

Rule 1.

Where there is more than one possible structural identity assignment for an atom, choose the identity occurring earliest in the alphabet.

For example in the structure below the oxygen is assigned as though it were a substituent



directly attached to the 5-membered ring rather than to the six-membered ring because G comes before T.

Other encoding rules are:-

Rule 2.

Multiple bonds and hydrogen atoms take on the structural identities of the atoms to which they are attached in accordance with rule 1.

The descriptor E is used to identify terminal atoms. A terminal atom of a chain is an atom other than hydrogen that is attached to only one other non-hydrogen atom in the chain.

There are two other encoding rules that apply to only specialized and rather unusual molecules with bridging substructures. A bridging substructure is said to exist in a molecule when a pair of rings of smallest size share more than one atom-pair bond [see figure (1g)]. A detailed

formalism for interpreting bridging systems is referenced elsewhere[1].

The bridging encoding rules are:

Rule 3.

Atoms or multiple bonds at the ends of bridges are given a bridging structural identity (see figure 1g).

Rule 4.

In bridging systems the ring type identity of an atom or multiple bond is established by assigning it to the ring of smallest size of which it is a member (see figure 1g). This is consistent with rule 1.

APPLICATIONS OF THE STRUCTURAL MOLECULAR FORMULA

The structural molecular formula contains two types of information, the molecular composition and the accompanying atom structural identities. With respect to structural resolution this representation is a considerable refinement over the traditional molecular composition formula. In fact, one could go so far as to say that the structural molecular formula provides a structurally close approximation to the two dimensional structure diagrams (Markush formulas) frequently used by chemists.

As a representation of molecular topology the structural molecular formula fits in the series of descriptions after structure diagrams, connection tables[2,3] and line formula notations[4] (e.g., Wiswesser Line Notation) but before molecular composition and molecular weight.

The structural molecular formula formalism is in no way an attempt to compete with or replace structure diagrams, connection tables or line notations. However there are important applications where the structural molecular formula can serve as a powerful and effective tool

A task that is becoming increasingly important in chemistry is the

searching of very large chemical data bases for both specific substructures and complete molecular structures. It is here that the structural molecular formula formalism can play an important role as a powerful screening tool. In the computer environment two-dimensional structural diagrams are impracticle, connection tables are unwieldy and slow to search, and line formula notations are not structurally consistent enough for general substructure searching. It is true that molecular compositions are easy to search and order however they lack the structural precision for working with very large data bases. In this context the structural molecular formulas stand as potentially highly effective screening tools. They can also be used to screen a set of molecules for their maximal substructural commonality and to set up a dictionary order for molecular structures. These and other applications of the structural molecular formula are discussed in detail elsewhere[4].

CONCLUSIONS

The structural molecular formula concept represents a logical extension of the molecular weight, and molecular formula descriptions of molecular structure. Its simplicity, structural resolution, and linearity make it an ideal screening tool for computer searching and screening of large chemical structure data bases. On the basis of its simplicity and structural resolution it should be very competitive with line formula notations and other computer representations of molecular structure.

Another important advantage of the structural molecular formula is that it can readily be derived automatically from connection tables and other molecular structure representations used by computers. Garfield's work[5] on the derivation of molecular formulas from nomenclature suggests that it may also be practical to derive structural molecular formulas from nomenclature by a computer algorithm.

ACKNOWLEDGEMENTS

The author wishes to thank Ann Titus for typing the manuscript.

REFERENCES

1. R.G. Dromey, "A Structural Molecular Formula for Flexible and Efficient Substructure Searching of Large Data Bases", *J.Chem. Inf.Comp.Sci.*, (submitted for publication).

2. D.J. Gluck, "A Chemical Structure Storage and Search System Developed at Du Pout", *J.Chem.Doc.* 1, 43-51 (1965).

3. L.C. Ray, R.A. Kirsch, "Finding Chemical Records by Digital Computers" Science 126, 814-819 (1957).

4. E.J. Smith, "Wiswesser's Line Formula Chemical Notations", McGraw-Hill, N.Y. (1968).

5. E. Garfield, "An Algorithm for Translating Chemical Names to Molecular Formulas", Nature, 192, 192 (1961).