hits, their number is always less than five.

It is considered that the strategy developed here can be extended for other more general cases such as patent information by increasing the degree of multiplicity of representations and by incorporating multiple functions of the Markush indicator.

Although the tactics discussed in this paper would be sufficiently powerful even for other types of generic names in the Handbook which are not mentioned in the present paper, it is evident that there are many difficult problems to solve on a wide variety of Markush claims in patent specifications, including indefinite and a theoretically infinite number of structures. Therefore, before we can deal with the patent information in a completely satisfactory way, it would be necessary to give careful consideration at least on the following points: (1) more generic terms such as alkyl and halogen should be taken care of; (2) the Q representation should have a greater degreeof multiplicity; (3) the Markush indicator should be able to discriminate between more than two situations; (4) a more complicated query must be accepted by using logical operations (AND, OR, and so on).

EXPERIMENTAL SECTION

The searching program is written in BASIC PLUS and runs on a Disital Equipment Corp. PDP-11/60 at Japan Association for International Chemical Information (JAICI).

ACKNOWLEDGMENT

We are indebted to T. Oshima of Japan Chemical Industry Ecology-Toxicology & Information Center (JETOC) for discussions related to the Handbook. This work was financially supported in part by a Grant-in-aid for Scientific Research from the Ministry of Education, Science, and Culture to which the authors' gratitude is due.

REFERENCES AND NOTES

(1) (a) Ash, J. E.; Hyde, E., Eds. "Chemical Information Systems"; Ellis Horwood Ltd.: Chichester, 1975. (b) Rush, James E. "Handling Chemical Structure Information". In "Annual Review of Information Science and Technology"; Williams, Martha E. Ed.; Knowledge Industry Publications: 1978; Vol. 13, Chapter 8. (c) Howe, Jeffrey W.; Milne, Margaret M.; Pennell, Ann F., Eds. "Retrieval of Medical Chemical Information". *ACS Symp. Ser.* **1978**, *No. 84*.

(2) "NEW CAS SERVICES". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 117.

(3) Lynch, M. F.; Barnard, J. M.; Welford, S. M. "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 1. Introduction and General Strategy". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 148-150. (b) Barnard, J. M.; Lynch, M. F.; Welford, S. M. "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 2. GENSAL, a Formal Language for the Description of Generic Chemical Structures". *Ibid.* **1981**, *21*, 151-161. (c) Welford, S. M.; Lynch, M. F.; Barnard, J. M. "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 3. Chemical Grammaers and Their Role in the Manipulation of Chemical Structures". *Ibid.* **1981**, *21*, 161-168. (d) Barnard, J. M.; Lynch, M. F.; Welford, S. M. "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 4. An Extended Connection Table Representation for Generic Structures". *Ibid.* **1982**, *22*, 160-164.

(4) The Japanese edition has a name index which helps reduce the time if the substance in question is listed as a specific name.

(5) Chemical Products Safety Division, Basic Industries Bureau, Ministry of International Trade & Industry. "Handbook of Existing Chemical Substances", 2nd Ed.; The Chemical Daily Co. Ltd.: Tokyo, Japan, 1981.

(6) (a) Kudo, Yoshihiro; Yamasaki, Tohru; Sasaki, Shin-ichi. "The Characteristic Polynomial Uniquely Represents the Topology of a Molecule". *J. Chem. Doc.* **1973**, *13*, 224-227. (b) Kudo, Yoshihiro; Sasaki, Shin-ichi. "The Connectivity Stack, A New Format for Representation of Organic Chemical Structures". *Ibid.* **1974**, *14*, 200. (c) Kudo, Yoshihiro; Sasaki, Shin-ichi. "Principle for Exaustive Enumeration of Unique Structures Constituent with Structural Information". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 43. (d) Kudo, Yoshihiro; Aoki, Shotaro; Takada, Yoshito; Taji, Toyoaki; Fujioka, Ichiro; Higashino, Kazuko; Fujishima, Hisayuki; Sasaki, Shin-ichi. "A Structural Isomers Enumeration and Display System (SIEDS)". *Ibid.* **1976**, *16*, 50. (e) Sasaki, Shin-ichi; Abe, Hidethugu; Hirota, Yuji; Kudo, Yoshihiro; Ochiai, Shukichi; Saito, Keiji; Yamasaki, Tohru. "CHEMICS-F: A Computer Program System for Structure Elucidation of Organic Compounds". *Ibid.* **1978**, *18*, 211.

Chemical Inference. 1. Formalization of the Language of Organic Chemistry: Generic Structural Formulas

JOHN E. GORDON*[1] and JOYCE C. BROCKWELL

Chemical Abstracts Service, Columbus, Ohio 43210, and Department of Chemistry, Kent State University, Kent, Ohio 44242

Received February 22, 1983

Categorization, syntax, semantics, and history of generic structural formulas (GSFs) are discussed. Their roles in chemical inference, chemical documentation, and chemistry learning are considered in the context of normalization and formalization of languages of structural formulas, chemical equations, and mechanisms. A formal language (ABSF) of homocomposite GSFs and a heterocomposite language employing normalized structural variables (NVSF) are defined and merged. Useful formal operations involving these languages, their expressive power, and their relationship to Markush SFs and the GENOA and GENSAL languages are considered.

Generic structural formulas, i.e., those that denote whole classes of specific structural formulas (SFs), are heavily used in all domains of the chemical literature. Despite this usage and despite applications in other parts of chemistry to be discussed below, they have not been the subject of a general and fundamental linguistic study, with two partial exceptions: Study of the information-handling aspects of Markush SFs has been motivated by their extensive use in patents.[2-4] Lynch and his students have recently begun to describe a more fundamental treatment of generic structural formulas of variable composition.[5] Our work, which concentrates on fixed-composition generic structures, complements Lynch's rather well. This report describes our initial study and attempts to place it in a broad context of the history of SF-class notation and its applications in chemistry and metachemistry beyond chemical information science.

IMPORTANCE OF GENERAL STRUCTURAL FORMULAS

Maturation of a science is generally accompanied by formalization of its languages. Chemistry is currently in this stage; its linguistic development is proceeding in two directions.

First, new machine-manipulable languages are being developed for use in such systems as the structure-elucidation systems CONGEN[6] and CHEMICS,[7] the synthesis-planning systems LHASA,[8] SECS,[9] etc., and the more general chemical behavior system CICLOPS.[10] Second, we are engaged in formalizing the ordinary "language of organic chemistry" (LOC), the language in which structural formulas and systematic and common names combine with structural operators and other chemical connectives to produce sentences, most commonly in the form of reaction or mechanism equations. This work involves extension of expressive power and removal of ambiguity in addition to extraction of formal morphological, syntactic, and semantic rule systems for generation of grammatical expressions and study of their interpretations, while retaining the form and flavor of established usage as far as possible. Since a great many of the concepts that chemists routinely express and manipulate involve classes of SFs/compounds, the developing LOC rather keenly experiences the need for an expressive and precise language of generic structural formulas, hence the present study.

It is a fundamental characteristic of chemistry that the smallest unit of molecular structure that defines an identifiable and stable set of properties observable in real chemical substances is in most cases a substructure containing 2–20 atoms. Examples are the correspondences between functional groups and their characteristic chemical reactions or between chromophores and their characteristic spectroscopic signals. Thus these substructures (the functional groups, chromophores, etc.) are fundamental chemical concepts. Since a substructure defines a class of structural formulas (those containing the substructure as a subgraph), chemical concepts often involve structural-formula classes. Indeed, one can find in the common inference patterns used by chemists many other examples of manipulation of SF classes (e.g., in making structural analogies and in combining isolated pieces of connectivity information to infer gross molecular structure). Thus one expects search by SF class to be an important tool in chemical information science.

Experience has shown SF-class searching via traditional access tools to be unsatisfactory; the comprehensive chemical data bases are not organized or indexed to facilitate SF-class searching. One alternative lies in mechanized substructure search systems (SSS) based on comparison of library and query screens (substructures). The substantial current interest in SSS confirms the importance of SF-class searching, and SSS will presumably prove to be the method of choice. There is no substitute, however, for a paper-and-pencil notation in the hands of the working chemist. While direct graphic input to SSS is rapidly displacing composition of queries via search dictionaries, a paper-and-pencil language may always intervene between chemist and machine as a medium for defining and conceptually manipulating SF classes.

Thus we visualize three roles for languages of generic structural formulas in chemical information science: (a) as a basis for generic nomenclature and substance-class indexing; (b) as a common interpretive language for discussing (i) input and output SF classes for a mechanized structure-elucidation system and (ii) substructure search—more generic than any SSS query language and independent of their screen organizations and input devices; (c) in chemical patent handling (see ref 5). Concrete examples of items b and a are given later in this paper and in the following paper of this series, respectively.

A consistent and complete language of generic SFs will have basic applications in chemical education as well. A good example of the difficulties produced by lack of accurate means for describing SF classes is provided by the classroom discussion of the elucidation of structure of a naturally occurring compound from chemical and spectroscopic data. Each piece of structural information defines a set of possible SFs for the unknown, and the possible SFs remaining in consideration at any time are the elements of the intersection of these individual sets. The problem is with what kind of SF to represent the incompletely known structure. If one discusses this process of incorporating successive structural facts with the specific SF that represents the final answer visible to the students, they find it very difficult to visualize the intermediate stages where less structural detail is known than is visible in the fully specific SF. Using the molecular formula for the unknown is better, but it underrepresents what is known, so that established structural features must be imagined at intermediate stages. What one needs is a language of generic SFs in which one can express at each stage in structure elucidation precisely what is known about the connectivity of the substance without implying any structural features that have not been established. At present, no language is available that possesses this degree of expressive power in addition to continuity with the standard, fully specific SF on the one hand and with the fully generic molecular formula on the other.

HISTORY OF GENERIC STRUCTURAL FORMULAS (GSFS)

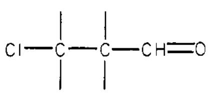
Structural Variables. The use of a *variable name*, symbolized by alphabetic characters and used to represent any element in a *replacement set*, appeared in mathematics no later than the seventh century (numeric variables; geometric variables by 1637).[11] The analogous chemical variables, ranging over replacement sets of atoms, substructures, or complete SFs, date from about 1880,[12] within about a decade of the time when ordinary, specific SFs took their present form. Both mathematical and chemical–structural variables may be either locally defined (e.g., x in the sentence $x^2 + 5x = 32$)[13] or globally (e.g., i in the sentence $x_i = 0$, where by global definition $i \in \mathrm{N}$, and N is the set of natural numbers). Structural formulas containing locally defined structural variables are known as Markush SFs; these are considered in a separate section. We will understand "structural variables" to denote the globally defined variety and consider here the actual extent of "global" and the stability of structural-variable names over time and across schools.

Some of the oldest structural-variable names have had the most stable denotations: e.g., M = metal, R = alkyl, X = halogen. Still, throughout the century of their use, different groups of chemists or the same chemists on different occasions have used these signs to denote narrower, very locally (e.g., within one paper) defined replacement sets (e.g., R = CH_3, CH_3CH_2, $CH_3(CH_2)_{10}CH_2$) or expanded ones (e.g., R = any hydrocarbon group, saturated or unsaturated). A good indication of current usage may be obtained from Morrison and Boyd's organic chemistry textbook (**MB**[14]). **MB** adhere closely to the narrowest denotations for R and X; they use HA and B for generic acid and base [but sometimes use A = any atom (p 132)] and introduce the useful G = any group in the context of a variable substituent [G is used at least once as a local variable (p 632: G = OH, NH_2, NHC_6H_5, $NHCONH_2$)]. W, Y, and Z are used only as locally defined variables (e.g., pp 154, 192, 629, 661).

As one goes progressively further from academic circles, usage becomes less stable; generic SFs in the patent literature very commonly use R and X as completely local variables, whose defined values apply only to a single SF. In informal transfer of information via SFs, particularly that accompanying chiefly oral discourse, local definitions are slow and cumbersome; speakers tend to rely on globally defined values for structural variables and to presume that their audience uses the same global definition. Since such discourse goes unrecorded, it receives little attention in information science, but its volume and impact are very great. Speakers under time constraint and striving to maintain continuity of discourse often press ordinarily stable variables (e.g., R = alkyl) into broadened or distorted, temporary, and unannounced definitions (e.g., R = any group). This makes the information conveyed ambiguous and is particularly frustrating to learners.

In summary, structural variables have never received truly global definitions, and this introduces a certain level of ambiguity wherever their use is not accompanied by a local definition. On the other hand, they are highly valuable. Normalization of global structural-variable definitions is a worthwhile goal. While this may never be achieved across the chemical community, it can be rigorously implemented in formal systems.

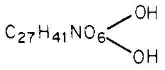
Part Structural Formulas. When one turns to progressively larger classes of SFs, in which the variable portions of the structure dominate the constant portions, structural variables become less useful. Customary usage then denotes the class of all SFs containing a constant substructure simply by drawing the substructure, usually with the connections to the variable portion shown explicitly as open bonds. Thus, I is



I

construed as a generic SF denoting the set of all SF graphs containing this graph as a subgraph, i.e., the class of all β-chloro aldehydes.

Brace Structural Formulas. A molecular formula can be construed as the generic SF denoting all the specific SFs that possess the indicated composition. This continuity of molecular formulas with SFs is illustrated nicely by SFs of the type 1[15] and 2,[16] in which the segment in molecular formula format

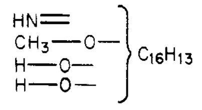


1

$CH_3O.C_{14}H_6O_2.O.COCH_3$

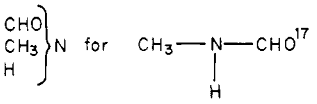
2

denotes constraint of the variable portion of the structure to the indicated elemental composition. To handle molecular formula segments with more than the two open bonds exemplified in 1 and 2, while avoiding implying anything about the points of connection of the fully specific groups to the molecular formula segment, the generic and specific portions were commonly separated by a brace, as in 3, where the partial



3

structures are said to be within the brace, and "$C_{16}H_{13}$" is called the *residue molform*. Such formulas, which we will refer to as *brace SFs* (BSFs), grew out of the earliest method of coding connectivity in *specific* SFs, e.g., in II. When in



II

the period following ca. 1870 connectivity was increasingly indicated by dots or dashes for bonds, the brace SFs continued to be used as *generic* SFs, e.g., structure 3 to denote the class of all SFs possessing one imino group, one methoxyl, two hydroxyls and a total composition of ($C_{16}H_{13}$ + NH + CH_3O + HO + HO =) $C_{17}H_{19}NO_3$. Brace-SF usage peaked in the early 20th century. Structure 3, which represents the alkaloid coclaurin at one stage in its determination of structure, is the last primary-journal instance we have found.[18] In what follows we refer to the traditional language of BSFs as TBSF.

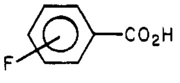
Brace SFs are equally well suited to represent operations of SF specification, in which discovery of additional specific structural details corresponds to moving atomic symbols from the residue molform to inside the brace, and SF generalization, which does the opposite.

Markush Structural Formulas. A Markush SF is a structural formula containing one or more structural variables (atoms, groups, substructures, or subscripts) together with an annexed definition of the replacement set for each variable; the definitions are strictly local and generally apply to one Markush SF only.

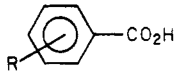
CLASSIFICATION AND FREQUENCY OF OCCURRENCE OF GSFS

GSFs fall into two major categories, depending on whether the elemental composition is fixed or variable. Table I labels these the *homocomposite* and *heterocomposite* types, respectively, and subcategorizes the latter according to the type of compositional variation present.

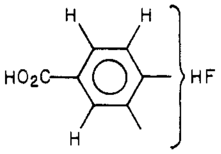
Brace SFs handle variable connectivity well, but not variable composition. Markush and (permanently defined) structural-variable SFs do the opposite. One useful and unambiguous combination suggests itself and is shown (labeled "Brace–Markush") in Table I; we consider various combined notations in a later section. Structural formulas with ambiguous bond placements (e.g., 4 or 5) are often called Markush SFs, even



4



5



6

when the composition is constant, as in 4, but the connectivity variation is advantageously conveyed in brace notation, as in 6.

Table II summarizes the distribution of a sample of 691 GSFs among the categories and graphic representations identified above. In each entry the upper number comes from analysis of the 376 GSFs appearing in one issue of *The Journal of Organic Chemistry*;[37] the number beneath refers to the 315-item sample of patent-abstract GSFs from *Chemical Abstracts*.[38] Brace SFs are entirely absent. While *The Journal of Organic Chemistry* sample displays a sizeable fraction of both Markush and ostensibly globally defined structural-variable GSFs (those with no accompanying definition of structural variables), the patent sample uses essentially all Markush SFs. The 204 homologous instances from *The Journal of Organic Chemistry* were examined in context to attempt to infer the SF class that the author(s) wished to represent. In 133 instances (65%) the usage does not appear to conform to strict, traditional definitions (R used to denote groups beyond alkyl and X for those beyond halogen).

The actual denotations of R (again, inferred from context) in both samples of GSFs are shown in Table III. In the Markush SFs of the patent sample, R, X, Y, etc. are used quite interchangeably in the explicit, local definition of any replacement set. However, it is also clear that in the ostensibly predefined structural-variable GSFs (i.e., those lacking an explicit local definition) of *The Journal of Organic Chemistry* sample, R ranges over essentially any replacement set the author pleases. This is probably true of the primary and abstract literature generally. It is not necessarily true of academic and textbook usage, where "permanent" definitions

Table I. Generic Structural Formulas

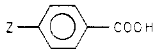
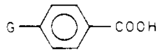
	class-member attributes			
type	have same molform?	have common substrs?	how class members vary	graphic representation available
homocomposite	+	+	connectivity, hybridization, formal electron/charge distribution	brace, e.g. $\left.\begin{array}{l}\text{-OH}\\ \text{-CH=O}\end{array}\right\}C_6H_{12}$
heterocomposite				
natural	–	+	elemental identity within a periodic family	Markush, e.g., $N{\equiv}C\text{-}CH_2\text{-}CO_2{}^-M^+$ for M = Li, Na, or structural-variable, e.g., CH_3CH_2X
arbitrary	–	–	"free" variation of substructure	Markush, e.g.  $Z = NO_2$, F, Cl, H, CH_3O or structural variable, e.g. 
repeating				
structurally homologous	–	+	number of CH_2s in structural formula	Markush, e.g., $(CH_3)_3C(CH_2)_nCOOH$ for $n = 8$–16
compositionally homologous	–	+	multiples of CH_2 in molform	brace–Markush, e.g. $\left.\begin{array}{l}\text{-OH}\\ \text{-CH=O}\end{array}\right\}C_nH_{2n}$ $1 \leqslant n \leqslant 10$, or structural variable, e.g., R-Br
non-C/H	–	+	multiples of any substructure	Markush, e.g., $C(CH_2X)_nH_{4-n}$, $1 \leqslant n \leqslant 4$

Table II. Occurrence of GSF Types

type	Markush	globally defined structural variable	free-standing substructure	ambiguous bond placement
homocomposite				0
				19
heterocomposite				
natural	1	19		
	<10	0		
arbitrary	118		16	
	94		0	
repeating				
structurally homologous				
compositionally	11	204		
homologous	189	0		
non-C/H	7			
	3			

Table III. Replacement Sets for the Structural Variable "R" Used in the 506-Item Sample[a] Studied[b]

	R			
source	alkyl	alkyl or other hydrocarbon group	any group	unclear
J. Org. Chem.	27	7	63	3
patent abstracts	9	12	79	

[a] See text. [b] Values given are percents.

of structural variables may be in force within a locale or volume, and it will not be true in any formal language of GSFs.

This paper aims primarily at description of one such language. Subsequent papers in this series illustrate the utility of the language and develop an isomorphic language of systematic generic names and augmented connection-table languages for chemical reactions and reaction mechanisms, together with a parallel formalization of the traditional graphic language of organic reactions as SFs connected by arrows (and perhaps other connectives).

DESIDERATA FOR GENERIC SFS

A language of generic structural formulas should (a) incorporate as much of stable, existing notation as possible, (b) be morphologically continuous with the set of specific structural formulas, (c) have sufficient expressive power to denote any desired class of structural formulas, and (d) be conductive to creation of generic nomenclature in 1:1 correspondence.

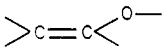
We will develop the one existing starting point for homocomposite SFs: the brace SF notation. As an adjunct to Markush SFs and GENSAL (GENeric Structural LAnguage of Lynch et al.[5]) as descriptive languages for heterocomposite SF classes, we explore the development of a formalized version of globally defined structural-variable SFs under the name *normalized-variable* SFs (NVSFs). The principal developmental tasks are then to assure the expressive power (completeness) and consistency of these languages, to provide inference rules for manipulation of SFs in these languages, and to explore the extent to which they can be merged into a single language. It proves possible to accomplish a good bit of this while generally satisfying the above desiderata.

DEFINITIONS AND SYMBOLIZATION

In general, upper-case Roman, italic, and boldface characters are used to symbolize structural variables, functions, and sets, respectively. The *i*th element of **Z** is z_i. In discussing formal properties of SFs we refer frequently to the (meaningful) combinations of atomic symbols with various juxtaposed unshared pairs, formal charges, and open bonds (e.g., $>N^+{=}$, $-\overset{..}{\underset{..}{O}}-$, etc.). These have been referred to as single atom fragments (SAFs) and as atomic bonding units (ABUs). We use the latter description and symbolize the set of ABUs as U. The following definitions hold:

S is the set of well-formed individual structural formulas (formation rules given in the Appendix).

P is the set of part structures (Appendix), e.g.



V = {R, X, Ar, G} is the set of normalized structural variables used in the NVSF language described in a later section.

The replacement sets for these variables are named **R**, **X**, **A**, and **P**, respectively. The formation rules for elements of **P** are given in the Appendix. **R** is then that subset of **P** in which only the ABUs >C< and –H appear and the number of open bonds/free valences is exactly one.

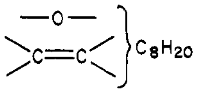
A is the set of aryl radicals, defined as that subset of **P** whose elements are composed solely of ABU tokens of type H—, =C<, =N̈—, =N⁺<, or =Ö⁺—, so arranged that each ABU bearing an open bond is part of at least one cycle of 6, 10, 14, ... ABUs whose mutual connections are of type "—" and type "=" in alternation around the cycle.

M is the set of well-formed molecular formulas (Appendix).

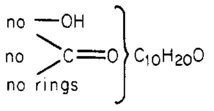
B is the set of well-formed brace SFs (formation rules in the Appendix). The constitution of an element of **B** is then given by eq 1, in which $p_{i,j} \in \mathbf{P}$ and $m_i \in \mathbf{M}$.

$$b_i = \left.\begin{matrix} p_{i,1} \\ p_{i,2} \\ \vdots \\ p_{i,n} \end{matrix}\right\} m_i \qquad (1)$$

For various chemical reasons we construe GSFs *intensionally*.[19a,20] Suppose two laboratories report what they have learned about the structure of a newly discovered compound as **7** and **8**, respectively (using, for the moment, a very intuitive



7



8

BSF syntax). Structures **7** and **8** describe the same set of SFs (are coextensional), and on an extensional criterion of class identity they would have identical denotations. Yet the chemist wishes them to represent something more than two different names for the same thing. We find good reasons to keep both and treat them as distinct GSFs. In a classroom context we need to demonstrate how one implies the other, and our formal system should contain inference rules for deriving each from the other. In the context of chemical documents, we might want to map structural conclusions, such as those in the above example, onto experimental subsections of the respective articles; this would require both the intensional and extensional statement of structural conclusions.

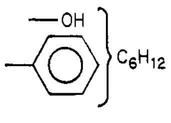
A formal treatment of SF classes as intensional expressions requires various *kind* functions[19b] for forming the *extensions* of such expressions (i.e., functions from GSFs to sets of SFs). The only kind function used in this article is K_B, which maps BSFs into isocomposite sets of SFs, as defined in eq 2. Here

$$K_B(b_i) = \hat{x}(x \in S \;\&\; p_{i,1} \,\Delta\, x \;\&\; p_{i,2} \,\Delta\, x \;\&\; \ldots \;\&\; p_{i,n} \,\Delta\, x \;\&$$
$$C(x) = C(m_i) + \sum_{l=1}^{n} C(p_{i,l})) \qquad (2)$$

$\hat{x}(\ldots x \ldots)$ is *Principia* notation for the class of entities x for which $(\ldots x \ldots)$ is true. The function C: $\mathbf{S} \cup \mathbf{B} \rightarrow \mathbf{M}$ returns the elemental composition of a SF/BSF. The sign Δ denotes "is a proper subgraph of".

BRACE SFS (BSFS) AS A LANGUAGE OF VARIABLE STRUCTURAL SPECIFICITY

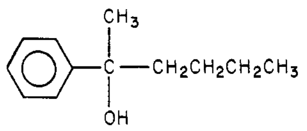
Qualitative Variables. A brace structural formula is a molecular formula from which a subset of the atomic symbols has been removed, taken under the brace, and structurally *specified*. Thus BSF **9** is a partial structural specification of **10**, and SF **11** is in turn a further structural specification of **9**. But there are generic structures intermediate between **9** and **11** (e.g., **12**). Thus BSFs exist in hierarchically ordered sets. We will consider the precise nature of the ordering relation, the uses of the hierarchy in inference, and strategies for its construction in later articles. Here we need to consider the qualitative varieties of structural specification and (in the next section) their interaction.



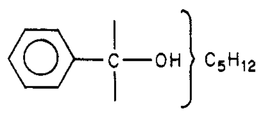
9

$C_{12}H_{18}O$

10

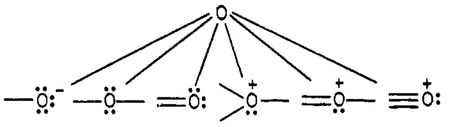


11



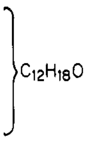
12

The information specified in taking selected atoms under the brace consists of valence, hybridization, charge, and connectivity. Since the first three properties are not independently variable, we lump them and speak of two variables only: a generalized *valence* and (local) *connectivity*. Specification of valence takes atomic symbols into ABUs, as in **13**.

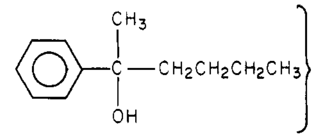


13

Desiderata a and b above motivate the following formal convention. Specific structural formulas and molecular formulas are defined as possessing the brace SF format, and their usual forms are defined as abbreviations of this morphology. Thus **10** is an abbreviation of **14**, **11** is an abbreviation of **15** (with null residue molform), and **9** and **12** are not abbreviations.



14



15

Qualitative Types. In principle both valence and connectivity information can be specified independently for each part of the molecule. To generate a wide variety of BSF types and investigate the expressive power of TBSF and the interaction between valence and connectivity information, we distinguish three values (no information, partial information, complete information) on the valence and on the connectivity dimension. Nine qualitative types of SF should result. However, when partial information is specified on each dimension, the molecular parts (substructures) specified for the two variables may be disjoint, identical, partially overlapping, or included (in two possible ways), thus increasing the number of such qualitative types of homocomposite generic SF to 13. These are displayed schematically and hierarchically in Figure 1.

Moving to actual BSFs, we consider the amounts of valence and connectivity information on an atom-by-atom basis and measure them by the fractions of all the atoms whose valence/connections are specified. We limit hydrogen to the ABU –H (no free protons in solution) and exclude it from the accounting. The two variables are not quite independent: there is nothing to connect until some ABU specification has been done. Consequently, we adopt a minimal criterion for connectedness of a non-H atom (any of its connections fixed) and a more stringent one for judging an atom to be valence specified (valence, hybridization, and charge fixed) in order to populate all of the above qualitative types with reasonable examples, which are shown in Figure 2.[21] Those types poor

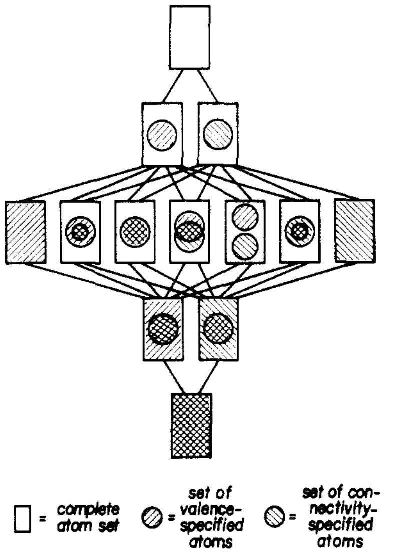
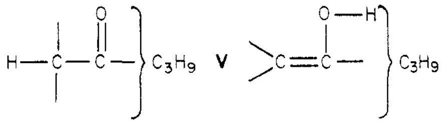
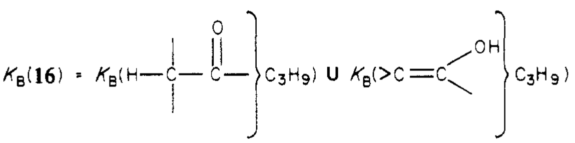

□ = complete atom set
▨ = set of valence-specified atoms
▧ = set of connectivity-specified atoms

Figure 1. Qualitative types of generic SFs as a function of the relative extensions of the sets of valence-specified and connectivity-specified atomic symbols.

in valence specification (right/center of Figure 1) tend to be cases in which tradeoffs between ABU combinations of equal composition, total connections, and net charge are possible but left unspecified. These often represent sets of chemically related SFs: mesomers (example 3), tautomers (example 12), acid–base species (example 8), valence tautomers (example 10), or the somewhat broader classes in examples 7 and 9. It is precisely these cases that prove difficult to express in TBSF, and Figure 2 resorts to representing these classes as the union of the extensions of two brace structures,[22] which is equivalent to a disjunction of BSFs. For example, consider the type-7 isomer subset[23] consisting of the enolizable carbonyl compounds of $C_5H_{10}O$ and their corresponding enols. This proves impossible to express in TBSF. In this case (and in the Figure 2 examples of types 3, 7, 8–10, and 12) one needs some means of expressing disjunction, specifically that in **16** and eq 3 for the present problem.



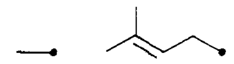
16


κ_B(16) = κ_B(H—C—C—}C3H9) ∪ κ_B(>C=C<OH }C3H9)

(3)

The next two sections explore means of expressing disjunction within the BSF language.

Open-Bond and Free-Valence Convention. In reviewing potential disjunctive operations for appropriateness to an augmented TBSF, we examine first a device borrowed from the mechanized formal system CONGEN,[24] namely, the free valence. We distinguish between *open bonds*, symbolized by "—" (or by A (below), in order to distinguish them from



A B

methyl groups, when using line-segment SFs, as in B), and *free valences* (symbolized by an asterisk) to which only multivalent atomic-symbol tokens may be attached. Addition of the free valence to TBSF does not solve the problem posed by **16**, but it solves the third lower homologue of that problem, namely, to express **17** with a single brace SF. The solution

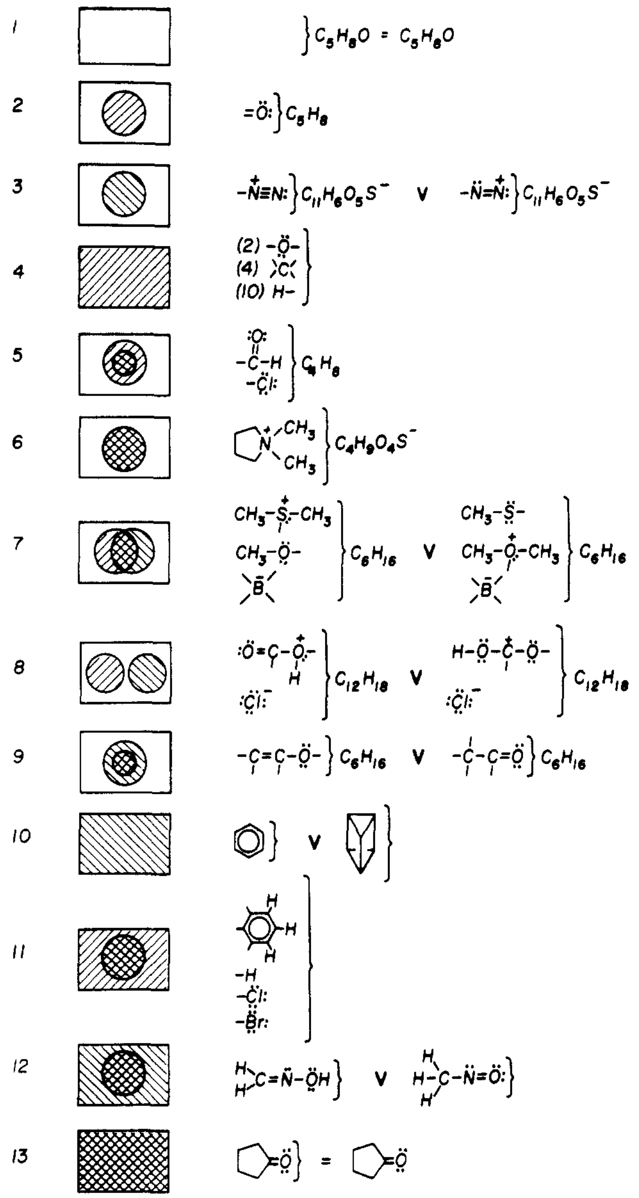
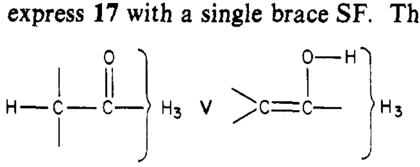
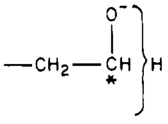


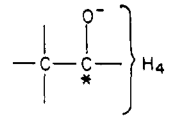
Figure 2. Examples of the various qualitative types of homocomposite generic structural formulas, illustrated by using traditional brace-SF syntax. The type symbolization is defined in Figure 1. "V" is logical OR.



17



18



19

may be written as **18** or **19**, which both represent the isomer subset $\{CH_3—CH=O, CH_2=CH—OH\}$ of the isomer set $\{CH_3—CH=O, CH_2=CH—OH, \overline{CH_2—CH_2—O}\}$. They do so by using the free valence to rule out the substructure $\overline{—CH_2—O—}$ (and hence the structure $\overline{CH_2—CH_2—O}$). Thus the asterisk introduces a selective, partial negating capability that (in combination with conjunction) is logically equivalent to a limited disjunctive capability.[25]

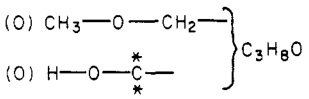
Further Negating Capability via Quantification. We extend flexibility and expressive power by adding a provision for integer coefficients on substructures within the brace and construing a zero coefficient as "none of these substructures present". The further convention is adopted that negated substructures have no effect on the atomic-symbol counts. Thus, to revise **20** to incorporate the information that no, one, or two hydroxyl groups are present, we write **21**, **22**, or **23**, respectively. We refer to the intrabrace coefficients as *quantifiers*.

$\}C_{10}H_{20}O_2$ **20** (0) $HO\text{-}\}C_{10}H_{20}O_2$ **21** (1) $HO\text{-}\}C_{10}H_{19}O$ **22**
(2) $HO\text{-}\}C_{10}H_{18}$ **23**

The isomer subset **24** rather arbitrarily selects some of the alcohols and some of the ethers from the seven-SF $C_4H_{10}O$ isomer set, and it is not expressible in TBSF. Addition to TBSF of the quantifiers discussed above allows **24** to be expressed as **25**.

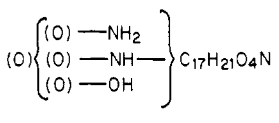
$\}CH_3(CH_2)_3OH,\ (CH_3)_2CHCH_2OH,\ CH_3CH_2OCH_2CH_3,\ (CH_3)_2CHOCH_3\{$

24

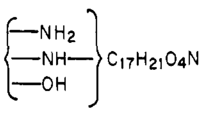


25

One further addition completes the negating capability of the BSF language, namely, negation of selected conjunctions of substructures. We realize this by binding the substructures in a reversed, internal brace and prefixing a zero coefficient. Thus, if the IR spectrum of $C_{17}H_{21}O_4N$, of otherwise unknown structure, reveals N–H or O–H stretching vibrations but fails to distinguish the two possibilities, this information can be expressed by **26**, which captures the continued disjunction

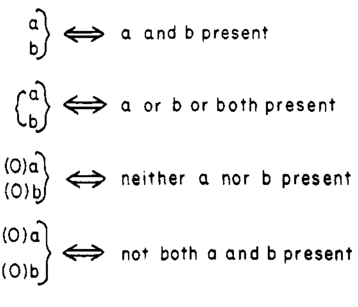


26



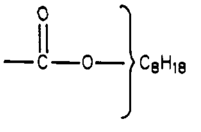
27

"–NH_2 or –NH– or –OH" as "not (no –NH_2 & no –NH– & no –OH)". In such constructions it is convenient to accept the BSF with all (0)s omitted as an abbreviation of the original; thus **27** abbreviates **26**. If a and b symbolize substructures, we then have the equivalences

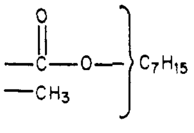


Overlapping Substructures. In the course of representing incomplete structural observations on an unknown SF, a given substructure may be identified in two experiments and be recorded in the BSF twice, a situation we shall refer to as overlapped substructures. Such overlapping is related to negative integers as quantifiers (intrabrace) and as subscripts in the residue molform. The main principles governing overlap and examples of its handling in the BSF language are enumerated here.

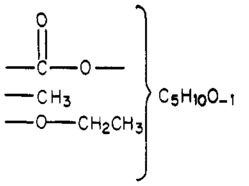
(1) The existence of overlap may or may not be revealed by negative subscripts in the residue molform. For example, suppose chemical and/or spectroscopic data on an unknown of composition $C_9H_{18}O_2$ identify the substructures –O–CO–, CH_3–, and CH_3–CH_2–O–, allowing CO_2, CH_3 and C_2H_5O chunks to be successively removed from the residue molform and taken under the brace to given in turn **28**–**30**. The



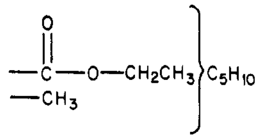
28



29



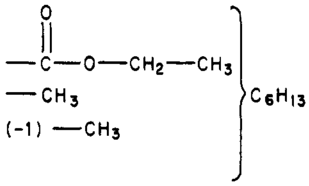
30



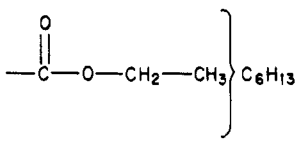
31

negative subscript on oxygen in **30** betrays the double counting of one oxygen, and examination of the substructures identifies this unequivocally as the etherial oxygen, allowing revision to **31**. However, it may also be the case that the two instances of CH_3– under the brace in **31** actually result from two different experimental observations of the same CH_3– substructure. In this event there is still overlap inside the brace, but nowhere in the BSF is there a clue to its presence.

(2) Overlap, once discovered, may be recorded in the BSF by (a) adding, under the brace, a copy of the overlapped substructure and prefixing to it the quantifier "(–1)" and (b) adding the equivalent complement of atomic-symbol tokens to the residue molform. Thus, if further experiments show that the intrabrace CH_3s in **31** are duplicates, this information is recorded as in **32**. A substructure carrying a (–1) quantifier is called an *overlap marker*.



32



33

(3) If, as in **32**, the existence of overlapped substructures has been recorded and the identity of the overlapped substructures is unambiguous, then they may be coalesced into a single instance with cancellation of the overlap marker. In the case of **32**, for example, this produces **33**.

(4) For the same reasons expressed in footnote 21, the inferences that may be made under principles 1 and 3, for instance **30** ⇒ **31** and **32** ⇒ **33**, should not be automatic or obligatory.

(5) Discovery of hidden overlap is additional structural information, the more so the more exactly the overlap locus can be identified. Assume that **34** (Figure 3), for example, summarizes observations to date on the structure of compound X. The class of possible SFs for X, i.e., the extension of **34**, has eight members. If the existence of double counting of a –CH_2– part structure is now discovered, but not the identity of the overlapping CH_2s, then **34** may be revised to **35**.[26] Figure 3 shows how narrowing the possible overlap locus narrows the class of SFs possible for X. To code this information into the BSF, we use the position of the quantifier to

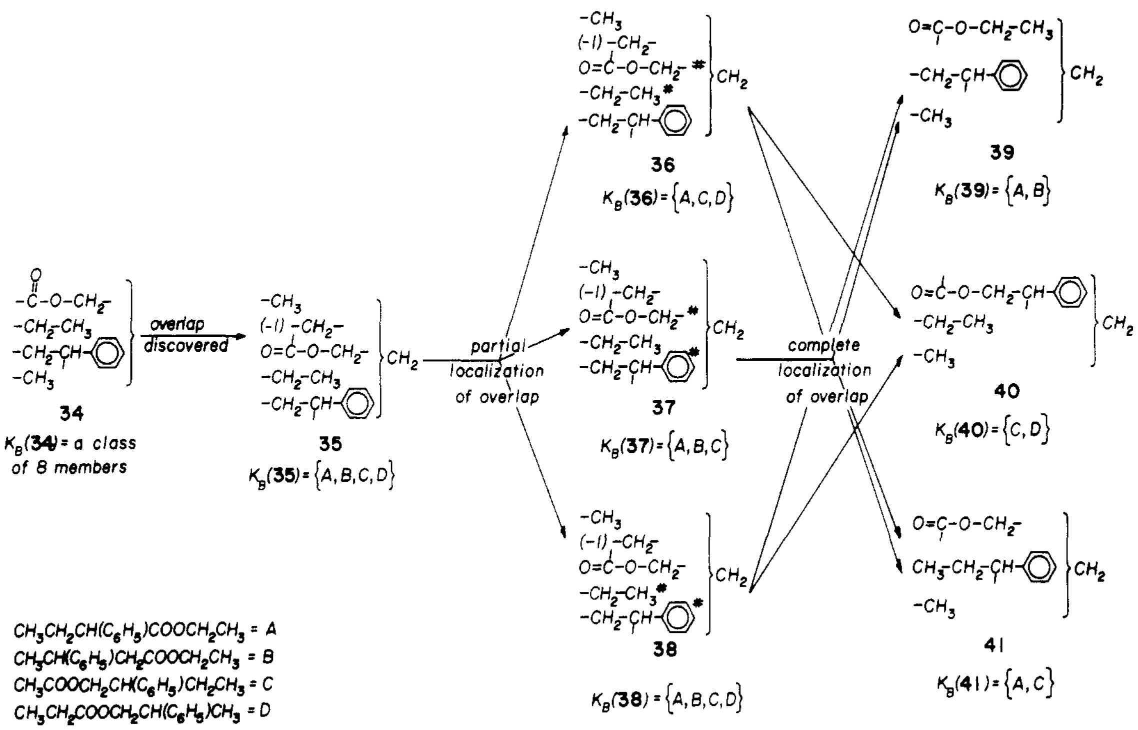
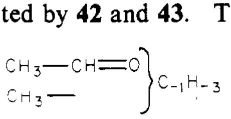

34
K_B(34) = a class of 8 members
overlap discovered
35
K_B(35) = {A,B,C,D}
partial localization of overlap
36
K_B(36) = {A,C,D}
37
K_B(37) = {A,B,C}
38
K_B(38) = {A,B,C,D}
complete localization of overlap
39
K_B(39) = {A,B}
40
K_B(40) = {C,D}
41
K_B(41) = {A,C}
CH3CH2CH(C6H5)COOCH2CH3 = A
CH3CH(C6H5)CH2COOCH2CH3 = B
CH3COOCH2CH(C6H5)CH2CH3 = C
CH3CH2COOCH2CH(C6H5)CH3 = D

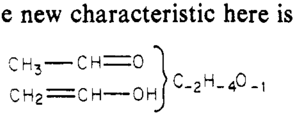
Figure 3. Effect of overlapping part structures on the extension of a brace structural formula.

indicate its scope: "(-1)" applies to (is cancellable with pairs of) part structures occurring below itself in the intrabrace list. When a possible overlap scheme can be eliminated by further data, the disallowed pairing is marked by a pair of #s ($s and %s as necessary).

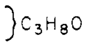
Although overlap in a BSF may or may not be signaled by negative residue-molform subscripts, if the process of logging redundant structural information into the BSF is carried far enough, residue-molform subscripts will eventually turn negative. Consider now what happens if this borrowing power against the residue molform is allowed to go unchecked. At and beyond the point where *all* residue molform subscripts have turned negative, there are sufficient atomic symbols under the brace to fashion one of the complete, specific SFs of the isomer set in question. Suppose there is sufficient structural information under the brace to make this inference; then there is nothing to prevent us from writing BSFs of the types illustrated by **42** and **43**. The new characteristic here is the presence of a "finished" SF (one with no open bonds or free valences) under the brace. If the extension of a BSF is the set of specific SFs each of which possesses each of the structures/part structures under the brace as a subgraph, then we have eq 4 and 5. As long as the supernumerary substructures, such as CH_3– in **42**, are, in fact, substructures of the "finished" SF under the brace, the brace contents remain consistent and read out as the finished SF. As soon as a structure/substructure (such as CH_2=CH—OH) is introduced that is not a substructure of the finished SF, logical inconsistency ensues, and the extension degenerates to null.



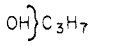
42



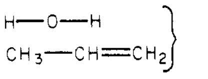
43



44



45



46

$$K_B(\mathbf{42}) = \{CH_3—CH=O\} \cap K_B(CH_3—\}C_1H_1O_1) = \{CH_3—CH=O\} \cap \{CH_3—CH=O\} = \{CH_3—CH=O\} \quad (4)$$

$$K_B(\mathbf{43}) = \{CH_3—CH=O\} \cap \{CH_2=CH—OH\} = \phi \quad (5)$$

Despite their apparent interpretability, it seems preferable to eliminate **42**, **43**, etc. on other grounds. Consider **44**–**46**. Since **45** is a structural specification of **44**, its extension is a subset of **44**'s. **46** appears to be a structural specification of **45**; that this cannot be true in the same sense as our previous cases, however, is shown by the extension of **46**. Presumably, $K_B(\mathbf{46}) = \{H—O—H + CH_3—CH=CH_2\}$, in which the "+" operator is that used in chemical-reaction equations such as A + B → C + D; i.e., $s_i + s_{j'}$ denotes juxtaposition of s_i and s_j. Thus $K_B(\mathbf{46})$ is not a subset of $K_B(\mathbf{45}) = \{CH_3CH_2CH_2OH, CH_3CHOH–CH_3\}$. Surely blocking of such anomalies as **46** within the BSF grammar is well motivated, and we accomplish this by adding the following rule: A BSF must possess either (a) at least one bond path, covalent and/or ionic, between all pairs of intrabrace atoms or (b) sufficient multivalent atoms in the residue molform to create such connectedness in the intrabrace graphs via some structural specification scheme. According to this rule, the only well-formed BSFs containing "finished" intrabrace structures are those with exactly one such finished structure under the brace and no further atomic symbols either under the brace or in the residue molform. Hence **46** is ungrammatical.

Completeness of the BSF Language. At this point we incorporate all of the expressive devices described above (the open-bond/free-valence convention, integers as meaningful coefficients on intrabrace part structures, and prohibition of unconnectable intrabrace part structures) and refer to the result as the augmented BSF (ABSF) language. We take desideratum c, above, as our criterion of expressive power; it is then easy to establish that ABSF satisfies this criterion.

Let the occurrence of substructures in the SFs of the isomer set of composition *j* be represented by a matrix whose columns correspond to the isomeric SFs. The (*k*,*l*)th element of this distinctive-feature matrix (DFM) is either 1, denoting at least one occurrence of the *k*th substructure in the *l*th isomer's SF, or 0, denoting no such occurrence. We define the *null feature* as that empty substructure that is a proper subgraph of every SF, and we make it the first feature of the DFM, so that the first row is the vector ⟨1, 1, 1, ..., 1⟩. Similarly, we produce

Table IV. Operations Producing Secondary from Primary Rows

operation	on corresponding BSF	on substructure feature	on element vector
1 (unary)	$b_{k_3} = (0)p_{k_1}\}\ m_j$	$f_{k_3} = \sim f_{k_1}$	change each 1 to 0 and each 0 to 1 in the row k vector
2 (binary)	$b_{k_3} = \left.\begin{matrix}p_{k_1}\\ p_{k_2}\end{matrix}\right\}\ m_j - C(p_{k_1}) - C(p_{k_2})$	$f_{k_3} = f_{k_1}\ \&\ f_{k_2}$	write 1 in any column in which the k_1 and k_2 entries are both 1; write 0 otherwise.
3 (binary)	$b_{k_3} = \left.\begin{matrix}p_{k_1}\\ p_{k_2}\end{matrix}\right\}\ m_j$	$f_{k_3} = f_{k_1}\ \mathrm{V}\ f_{k_2}$	write 0 in any column in which the k_1 and k_2 entries are both 0; otherwise write 1.

a last row, ⟨0, 0, 0, ..., 0⟩, by choosing a feature possessed by no member of the isomer set, such as an atom not present in the molform or an impossible feature, e.g., a quadruple bond to carbon. We associate with each row, in addition to its substructure feature, a *corresponding* BSF, which, for the first and last rows, is defined as the BSF of composition j with one intrabrace part structure, namely, the empty part structure (first row) or the impossible part structure (last row). Thus, for the first and last rows, the element vector (⟨1, 1, 1, ..., 1⟩ or ⟨0, 0, 0, ..., 0⟩) represents the denotation (extension) of the corresponding BSF according to the following code: 1 (or 0) in column l means isomer l is present in (absent from) the isomer subset denoted by the corresponding BSF.

The first and last rows are called *primary* rows because they involve a single substructure feature. Further primary rows are added by the process of *specification* of single substructures. For specification of the corresponding BSF of such a row, q, remove atomic symbols from the molform of an "empty" BSF, }m, assemble them to make a part structure, p_q, and place p_q under the brace. The feature for row q is then just p_q, and the element vector is produced by entering 1 in the column under every SF that contains p_q as a substructure and 0 in every other column.

Secondary feature rows are produced by unary and binary operations on primary rows performed in parallel on the corresponding BSF, the substructure-feature and the element-vector columns. In the following definitions k_1 and k_2 are argument rows, and k_3 indexes the secondary row that results from the operation; the primary BSF corresponding to row k_1 is represented by $b_{k_1} = p_{k_1}\}m_j - C(p_{k_1})$, where m_j is the molform of composition j, and $C(p_{k_1})$ is the composition of p_{k_1} (see Table IV).

A sample DFM for the $C_4H_{10}O$ isomer space is shown in Table V. Only a few sample secondary features have been included.

We now assert the following.

(1) Repeated applications of the above operations, starting from the primary rows, suffice to complete the DFM such that it contains element vectors corresponding to every binary number between 111...1 and 000...0. If the number of isomers (columns) is n_j, then the complete DFM thus contains 2^{n_j} rows.

(2) It follows from the above definitions that the 1:1:1 correspondence of the linked operations preserves 1:1 correspondence between the BSFs and the features and preserves either 1:1 or many-to-1 correspondence between BSFs or features and element vectors in each row of the completed DFM. Thus, for each row, the extension of the corresponding BSF is exactly that subset of the isomer set that would be obtained by including each isomer whose column contains a 1 and excluding each isomer marked 0.

The first assertion is justified by displaying an effective method for producing a BSF and a substructure feature that correspond to any needed element vector: From the SF corresponding to column 1 delete one H (or other monovalent) atomic symbol. The resulting part structure is a substructure of no other SF in the isomer set; name it the *indicative* feature of column 1, and symbolize it by Z_1. Similarly, visualize the indicative features of columns 2, 3, ..., n_j. Create Z_1, Z_2, ..., Z_{n_j} by specification, and add the following *indicative rows* to the DFM:

corresp BSF	feature	isomers
$Z_1\}\mathrm{H}$	Z_1	⟨1, 0, 0, . . ., 0⟩
$Z_2\}\mathrm{H}$	Z_2	⟨0, 1, 0, . . ., 0⟩
.	.	.
.	.	.
$Z_{n_j}\}\mathrm{H}$	Z_{n_j}	⟨0, 0, 0, . . ., 1⟩

Table V. Distinctive Features for Description of the $C_4H_{10}O$ Isomer Power Set

feature no.	feature	isomers[a] ABCDEFG	corresponding BSF
1	null	1111111	$\}\mathrm{C_4H_{10}}$
2	$*\mathrm{OH}$	1111000	$*\mathrm{OH}\}\mathrm{C_4H_9}$
3	$*\mathrm{CH_2}*$	1110110	$*\mathrm{CH_2}*\}\mathrm{C_3H_8O}$
4	$*\mathrm{CH_2OH}$	1100000	$*\mathrm{CH_2OH}\}\mathrm{C_3H_7}$
5	$*\mathrm{CH_2CH_3}$	1010110	$*\mathrm{CH_2CH_3}\}\mathrm{C_2H_5O}$
6	$*\mathrm{CH_2CH_2}*$	1000010	$*\mathrm{CH_2CH_2}*\}\mathrm{C_2H_6O}$
7	$*\mathrm{CH_2CH_2CH_2}*$	1000000	$*\mathrm{CH_2CH_2CH_2}*\}\mathrm{CH_4O}$
8	$*\underset{*}{\mathrm{C}}\mathrm{H}*$	0110001	$*\underset{*}{\mathrm{C}}\mathrm{H}*\}\mathrm{C_3H_9O}$
9	$*\mathrm{CH_2}\underset{*}{\mathrm{C}}\mathrm{H}*$	0110000	$*\mathrm{CH_2}\overset{*}{\mathrm{C}}\mathrm{H}*\}\mathrm{C_2H_7O}$
10	$\mathrm{CH_3}\underset{*}{\mathrm{C}}\mathrm{HCH_3}$	0100001	$\mathrm{CH_3}\underset{*}{\mathrm{C}}\mathrm{HCH_3}\}\mathrm{CH_3}$
11	$*\underset{*}{\mathrm{C}}\mathrm{HCH_2OH}$	0100000	$*\underset{*}{\mathrm{C}}\mathrm{HCH_2OH}\}\mathrm{C_2H_6}$
12	$*\underset{*}{\mathrm{C}}\mathrm{HOH}$	0010000	$*\underset{*}{\mathrm{C}}\mathrm{HOH}\}\mathrm{C_3H_8}$
13	$*\overset{*}{\underset{*}{\mathrm{C}}}*$	0001000	$*\overset{*}{\underset{*}{\mathrm{C}}}*\}\mathrm{C_3H_{10}O}$
14	$*\mathrm{O}*$	0000111	$*\mathrm{O}*\}\mathrm{C_4H_{10}}$
15	$*\mathrm{OCH_2}*$	0000110	$*\mathrm{OCH_2}*\}\mathrm{C_3H_8}$
16	$*\mathrm{CH_2OCH_2}*$	0000100	$*\mathrm{CH_2OCH_2}*\}\mathrm{C_2H_6}$
17	$*\mathrm{OCH_3}$	0000011	$*\mathrm{OCH_3}\}\mathrm{C_3H_7}$
18	$*\mathrm{CH_2CH_2O}*$	0000010	$*\mathrm{CH_2CH_2O}*\}\mathrm{C_2H_6}$
19	$*\mathrm{O}\underset{*}{\mathrm{C}}\mathrm{H}*$	0000001	$*\underset{*}{\mathrm{C}}\mathrm{HO}*\}\mathrm{C_3H_9}$
20	$*\overset{**}{\underset{**}{\mathrm{C}}}$	0000000	$*\overset{**}{\underset{**}{\mathrm{C}}}\}\mathrm{C_3H_{10}O}$
i	2 & 3	1110000	$\left.\begin{matrix}*\mathrm{OH}\\ *\mathrm{CH_2}*\end{matrix}\right\}\mathrm{C_3H_7}$
j	4 V 15	1100110	$\left.\begin{matrix}*\mathrm{CH_2OH}\\ *\mathrm{CH_2O}*\end{matrix}\right\}\mathrm{C_4H_{10}O}$
k	¬10	1011110	$(0)\mathrm{CH_3}\underset{*}{\mathrm{C}}\mathrm{HCH_3}\}\mathrm{C_4H_{10}O}$
l	¬(6 V 17) = ¬6 & ¬17	0111100	$\left.\begin{matrix}(0)*\mathrm{CH_2CH_2}*\\ (0)*\mathrm{OCH_3}\end{matrix}\right\}\mathrm{C_4H_{10}O}$
⋮			

[a] A, $CH_3(CH_2)_3OH$; B, $(CH_3)_2CHCH_2OH$; C, $CH_3CH_2CHOHCH_3$; D, $(CH_3)_3COH$; E, $(CH_3CH_2)_2O$; F, $CH_3(CH_2)_2OCH_3$; G, $(CH_3)_2CHOCH_3$.

Now the row corresponding to any arbitrary element vector, $\langle x_1, x_2, ..., x_{n_j}\rangle$, can be produced by starting with the last row $(\langle 0, 0, 0, ..., 0\rangle)$ and applying operation 3 successively to each indicative row whose 1 element matches a 1 in the $\langle x_1, x_2, ..., x_{n_j}\rangle$ vector. The resulting (secondary) feature will be a continued disjunction of indicative features.

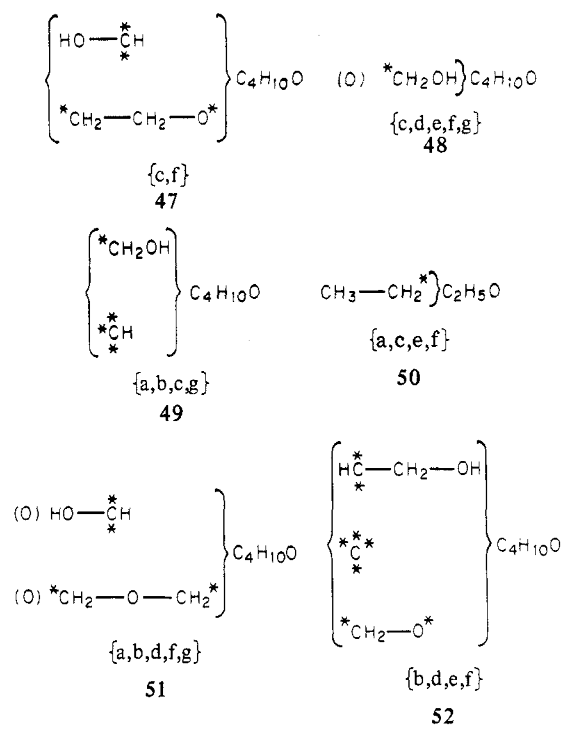
The proof of completeness of the ABSF language may then be stated in these terms: There are precisely 2^{n_j} subsets of the isomer set of composition j (the so-called power set of the isomer set). The 2^{n_j} distinct element vectors, $\langle 1, 1, 1, ..., 1\rangle$, ..., $\langle 0, 0, 0, ..., 0\rangle$, unambiguously denote these 2^{n_j} subsets via the code defined above. The completed DFM affords at least one distinct BSF corresponding to each element vector. Each of these BSFs is a well-formed sentence of ABSF; therefore, ABSF provides distinct representations of any (structural) isomer power set.

Minimal and Canonical ABSF Representations and Various Properties of Features. The number of primary features required to complete the DFM and thus represent the isomer power set via ABSF depends upon the nature of the substructures used as primary features. Clearly, since the whole DFM may be constructed by using the n_j indicative rows as the only primaries, precisely n_j features are required if they are all indicative features. However, the BSFs produced solely from indicative part structures, though legitimate, are unwieldy. The indicative part structures are question-begging features whose use in BSFs for isomer subset representation approaches enumeration of the set. Of greater interest are "more intensional" BSFs employing more broadly distributed substructure features. Thus we must consider briefly the problem of constructing "better" or "minimal" ABSF representations of arbitrary isomer subsets. This problem is essentially *isomorphic* with that of finding (disjunctive) normal forms of the formulas of sentential logic[27] or finding the simplest switching or logic circuit representing an arbitrary truth function.[28] BSF representation shares with these problems two general characteristics: several variously motivated criteria of simplicity exist, and many construction strategies are possible. The BSF problem *differs* from the others in, first, requiring a means of extracting appropriate substructure features and, second, requiring for many applications a *canonical* BSF representation.

The feature-selection problem is closely related to that of screen definition in a substructure search, which has received considerable attention;[29] it will not be discussed here. The substructure features displayed in Table V for the $C_4H_{10}O$ isomer power-set problem were extracted by a simple algorithm described in the Appendix. Several randomly selected isomer subset–BSF pairs are shown as **47–52** (Chart I). (The a, b, ..., g codes for these isomers are defined in Table V). These BSFs were produced by the construction algorithm given in the Appendix. The strategy embodied in this particular algorithm uses a minimization criterion that gives preference to (a) smaller (vs. larger) substructure features, (b) primary (vs. secondary) features and (c) conjunctions > disjunctions > negations > continued disjunctions of primary features. The incomplete compatibility of these principles was bridged rather arbitrarily (see Appendix).[30]

We believe the above to represent canonical representations by virtue of the following characteristics of the generation procedure: (a) Primary feature extraction is exhaustive, irredundant, and independent of the representation or order of presentation of the elements of the isomer set. (b) Elaboration of secondary features is exhaustive, irredundant, and produces a canonical, complete ordering of the extended feature set. (c) The BSF construction algorithm embodies unambiguous precedence rules for the selection of logical forms for the BSF and the allocation of feature values to their variable slots.

Chart I



{c,f}
47

{c,d,e,f,g}
48

{a,b,c,g}
49

{a,c,e,f}
50

{a,b,d,f,g}
51

{b,d,e,f}
52

The primary feature-selection algorithm used in Table V excludes indicative part structures while producing all of the part structures obtainable by breaking single bonds between polyvalent atoms in all members of the isomer set. There are 48 such features for the $C_4H_{10}O$ isomer space, each of which occurs, on the average, in 1.62 of the seven isomeric SFs of this isomer set. However, the 48 represent only 18 element vectors that are both mutually distinct and distinct from the $\langle 1, 1, 1, ..., 1\rangle$ and $\langle 0, 0, 0, ..., 0\rangle$ rows (62% redundancy). The irredundant set chosen by the algorithm (features 2–19 of Table V) has an occurrence rate in the isomer set of 2.11 isomers per feature, as compared to an optimum of 3.5 and to 1.00 for the indicative features.

What is the minimum number of features of this type needed to describe the isomer power set? Fixing the number of features used at some value d amounts to choosing to work in a description space of d (binary) dimensions. This choice determines the resolving power of our vision: we can perceive any individual only as one of 2^d types.[31] Thus in a description space of d substructure features we implicitly agree to describe the universe of SFs in terms of just 2^d types.[31] For example, in a description space containing two substructure features, a and b, all specific SFs reduce to $2^2 = 4$ types: (I) a present, b present; (II) a present, b absent; (III) a absent, b present; (IV) a absent, b absent. In terms of these four types, at most 2^4 classes of SFs may be described, as shown in Table VI, in which a plus in a column indicates that that SF is included in the class description (isomer subset), and a minus indicates that it is excluded. Thus, in general, the number of SF types is 2^d, the length of a class description (description of an isomer set) is 2^d bits, and 2^{2^d} distinct class descriptions exist. Now let n_j be the number of isomers in the isomer set of composition j. The isomer power set then consists of 2^{n_j} isomer subsets, and description of these via the 2^d 2^d-bit class descriptors requires that eq 6 hold. Combination with the upper limit from

$$2^{2^d} \geq 2^{n_j} \tag{6a}$$

$$d \geq \log_2 n_j \tag{6b}$$

use of indicative part structures gives eq 7. Formally, eq 6

$$n_j \geq d \geq \log_2 n_j \tag{7}$$

Table VI. Classification of the Universe of Structural Formulas in a Two-Dimensional Description Space

	class membership				
class	type I[a]	type II[b]	type III[c]	type IV[d]	class description
1	+	+	–	–	a present
2	–	–	+	+	a absent
3	+	–	+	–	b present
4	–	+	–	+	b absent
5	+	–	–	–	a and b present
6	+	+	+	–	a or b present
i	+	–	–	+	b present if and only if a present
⋮					
j	–	+	+	+	not both a and b present
15	–	–	–	–	a present and absent
16	+	+	+	+	a present or absent

[a] a present, b present. [b] a present, b absent. [c] a absent, b present. [d] a absent, b absent.

Table VII. Properties of Numerically and Morphologically Sufficient Feature Sets

d	n_j	$s^d{}_{n_j}$	total d-fold combinations in DFM	fraction of sufficient combinations
1	2	2	4	0.5
2	2	6	6	1.0
2	3	12	28	0.429
2	4	12	120	0.100
3	5	1120	4960	0.226
3	6	3360	41664	0.081
3	7	6720	341376	0.020
3	8	6720	2763520	0.002

says that given d rows of the DFM and operations 1–3, the remainder of the 2^{2^d}-row DFM can be derived: it defines *numerical* sufficiency. However, not just any combination of d rows suffices. Only those combinations of d rows succeed which, when stacked in any order, produce vertical d-bit numbers that are distinct for all 2^d columns.[32] In other words, use of a d-dimensional binary description space to describe an isomer power set succeeds only when d descriptors can be found that suffice to distinguish the n_j isomers (condition of *morphological* sufficiency).

For determination of the number of morphologically sufficient combinations of d rows in an n_j-column DFM, start with that unique, ordered stack of d 2^d-bit rows whose columns are both distinct and in increasing numerical order

```
0011 for d = 2        00001111
0101                  00110011 for d = 3, . . .
                      01010101
```

From this (sufficient) combination select n_j columns; there are $2^d!/[n_j!(2^d - n_j)!]$ such selections. Each of these can be multiplied $n_j!$-fold by permutation of columns, but the resulting list of combinations is $d!$-fold redundant. The number of numerically and morphologically sufficient combinations, $s^d{}_{n_j}$, is thus given by eq 8. Table VII summarizes some $s^d{}_{n_j}$ values,

$$s^d{}_{n_j} = \frac{2^d!}{(2^d - n_j)!d!} \tag{8}$$

together with other morphological properties of the sufficient row combinations.

Thus, there is a statistical bias against the success of a random choice of descriptors. Exhaustive examination of Table V reveals no numerically minimal combination of (three of) the primary features that is morphologically sufficient. In the larger sense, however, such failures reflect not only the statistical bias but also chemical–structural and conceptual/perceptual factors; there is a competing criterion of *semantic* acceptability. Consider a near-miss three-feature combination from Table V. With columns ordered GFEDCBA, rows 1 and 4 provide two members of a numerically and morphologically sufficient row set: {0001111, 0110101, 0010011}. The missing member, 0010011, is one of 16 candidate rows that form morphologically complete combinations with the first two members. It corresponds to a feature possessed by $CH_3CH_2CH_2CH_2OH$, $(CH_3)_2CHCH_2OH$, and $(CH_3CH_2)_2O$ but not by isomers c, d, f, or g. Can such a feature be found? In the form of a discrete part structure, no; among more arcane forms (see e.g., ref 29), perhaps it can, but such a solution is not obvious to any visually oriented discovery procedure. The remaining 15 candidate rows behave similarly. Choosing to categorize a universe of individuals (e.g., SFs) by means of a description space with a minimal number of dimensions, d, constrains the permissible relations between the descriptors severely enough that there may be no overlap between sets of d descriptors that are morphologically sufficient and sets that we are willing on semantic grounds to accept as "primary" features. Thus, Table V lacks feature sets that are numerically minimal and morphologically sufficient because of our underlying definition of semantic acceptability of features as *visually constructable and manipulable structural concepts*. On the other hand, in the SF-class problem, one need not back off very far on the dimension of numerical sufficiency in order to gain considerable flexibility of expression using acceptably rational and cognitively practical features. For example, many morphologically sufficient and semantically acceptable sets of *four* of the primary features in Table V suffice to express the $C_4H_{10}O$ isomer power set, namely, features 1, 2, 5, and 11 or 1, 5, 7, and 9, etc.

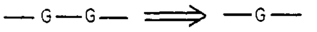
Table VIII. Structural Variables

sign	denotation	valence	atomic bonding units
R	alkyl (saturated)	1	R–
X	halogen	1–4	$:\underset{\cdot\cdot}{\ddot{X}}:^{-}$, $:\underset{\cdot\cdot}{\ddot{X}}-$, $-\underset{\cdot\cdot}{\ddot{X}}-^{+}$, $-\underset{\cdot\cdot}{X}<^{2+}$, $>X<^{3+}$
Ar	aryl	variable	Ar–, –Ar–, –Ar<, . . .
G	any part structure (radical)	variable	G–, –G–, –G<, . . .

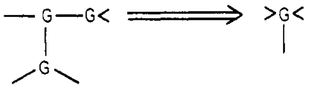
NORMALIZED-VARIABLE SFS (NVSFS)

Here we describe what appear to be the minimal extensions and restrictions on intuitive specific-SF notation required for a fully hierarchic description of the substantives of the language of organic chemistry (LOC) while realizing the desiderata set out earlier. While it is not possible to bring out here all of the motivations for the scheme proposed, which stem from hierarchy construction (part 4 of this series), LOC syntax (part 5), and generic nomenclature (part 2), enough will be said to indicate its scope and consistency. We admit the four structural variables listed in Table VIII to the alphabet, U, of atomic bonding units from which well-formed specific SFs may be built (Appendix).

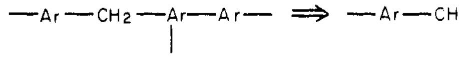
Functional groups directly attached to one another do not retain their individual chemical properties but modify one another and amalgamate to produce a new functional group. Thus we adopt the convention that directly connected occurrences of G (and, by extension, of Ar) are obligatorily and recursively redrawn to replace all instances of G–G or Ar–Ar by G and Ar, respectively, as in the examples of eq 9 and 10.



(9)

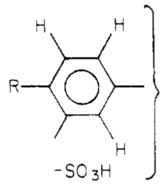


(10a)

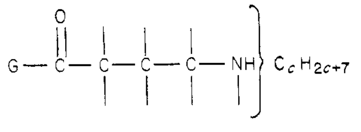


(10b)

Normalized variables may occur inside the brace of an ordinary, homocomposite BSF without restriction (e.g., **53**),

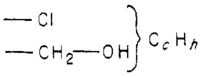


53

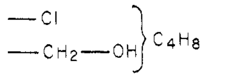


54

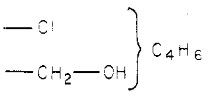
the BSF thus becoming heterocomposite. In order to complete the spectrum of GSFs from the most specific (whose extensions are single, specific SFs) to the most general (whose extension is the class of all specific SFs) it proves desirable to provide also for BSFs with *variable residue molforms*. This is accomplished, and continuity with the homocomposite BSFs maintained, by permitting independent or dependent variation of the residue molform subscripts. The latter case is illustrated by **54**, which constrains the denotation to saturated skeletons. In the former case (e.g., **55**) we must adopt also the convention



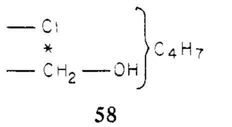
55



56



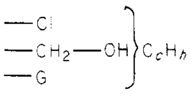
57



58

that only those combinations of variable subscripts are implied that produce well-formed SFs. Thus **55** includes **56** and **57** but not **58**, which is ill-formed.

One further restriction upon BSFs having structural variables within the brace may be well motivated. When the variable-valence signs G or Ar appear under the brace of a BSF whose residue molform is partly or fully constrained, their valences must be made explicit by open bonds, for only in this case is it possible to compute the unsaturation indices (see below) or determine whether the BSF implies a nonempty extension. Under this convention the extension of **54** includes only monovalent values for G, but that of **59** includes all SFs



59

containing one chloro group, one primary alcohol function, and one additional radical of any valence attached to a well-formed hydrocarbon skeleton, e.g., *c* and *h*, and the valence of G may take on any compatible set of values. We refer to the variable-connectivity, variable-composition language resulting from accommodation of normalized variables and variable molform subscripts in ABSF, under the guidelines just described, as "NVBSF".

INFERENCE RULES FOR MANIPULATION OF GSFS

Various types of rules for transforming GSFs and incorporating them in higher order linguistic structures will be described in subsequent papers. Here we consider a housekeeping rule (for preserving the grammaticalness of BSFs under transformation by the other rules) and a utility assisting in such transformations.

Rule 1. When a quantified part structure, $(n_i)p_i$, is inserted (removed from) beneath the brace, its composition, $C((n_i)p_i)$, is subtracted, element by element, from (added to) the residue molform. Negative residue-molform coefficients are valid, but operations under this rule are subject to two constraints: (a) If a complete SF (no open bonds or free valences) occurs under the brace, it must be the only item under the brace. (b) If charged substructures are inserted under the brace, equal and opposite charges must be posted to the residue molform, displaying their algebraic sum as a suffixed superscript.

Rule 2. The unsaturation index (hydrogen deficit)[33] of a BSF may be represented as partitioned between internal (under the brace) and external contributions, as in eq 11. The

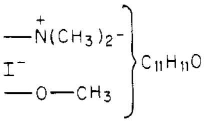
$$\Delta = \Delta_i + \Delta_e \qquad (11)$$

computation of Δ_i is done in normal fashion (+1 for each ring or double bond, +2 for each triple bond, −1 for each ionic bond) for the substructures under the brace, ignoring open bonds and free valences. Δ_e is computed from eq 12, in which

$$\Delta_e = c - h/2 - x/2 + n/2 + e/2 - q \qquad (12)$$

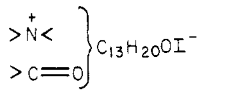
c, *h*, *x*, and *n* are the residue-molform subscripts on carbon, hydrogen, halogen (total), and nitrogen, respectively, *e* is the excess of open bonds/free valences under the brace, and *q* is the absolute value of the net charge on the residue molform. *e* is the number of open bonds, free valences, and charges remaining under the brace after the maximum number of free-valence/free-valence, open-bond/open-bond, free-valence/open-bond and +/− pairings between substructures under the brace have been made. Pairings may not (a) make rings or (b) pair two valences of the same atom.

Examples (all $C_{14}H_{20}O_2NI$, $\Delta = 5$) are given in **60**–**62**.



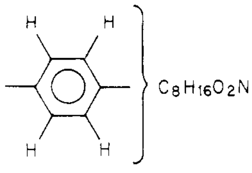
$e = 1$
$\Delta_i = -1$
$\Delta_e = 11 - 11/2 + 1/2 = 6$

60



$e = 5$
$\Delta_i = 1$
$\Delta_e = 13 - 20/2 - 1/2 + 5/2 - 1 = 4$

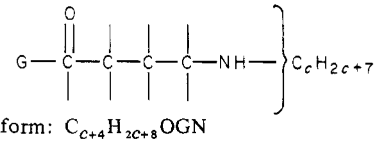
61



$e = 3$
$\Delta_i = 4$
$\Delta_e = 8 - 16/2 + 1/2 + 3/2 - 1 = 1$

62

If, by the convention proposed above, all valences of G and Ar are made manifest as bonds, charges, open bonds, or free valences, then eq 11 remains valid for composition-variable BSFs. In this case, however, the term $g(v - 2)/2$ must be added to the analogous expression for computation of Δ for the complete molform.[33] Here *g* is the number of G atoms, and *v* is the valence of G. In many cases, such as **55**, numerical evaluation of Δ_i and Δ_e is possible.



molform: $C_{c+4}H_{2c+8}OGN$
$\Delta = c + 4 + 1 - (2c + 8)/2 - 1/2 + 1/2 = 1$
$e = 7$
$q = 0$
$\Delta_i = 1$
$\Delta_e = c - (2c + 7)/2 + 7/2 = 0$

55

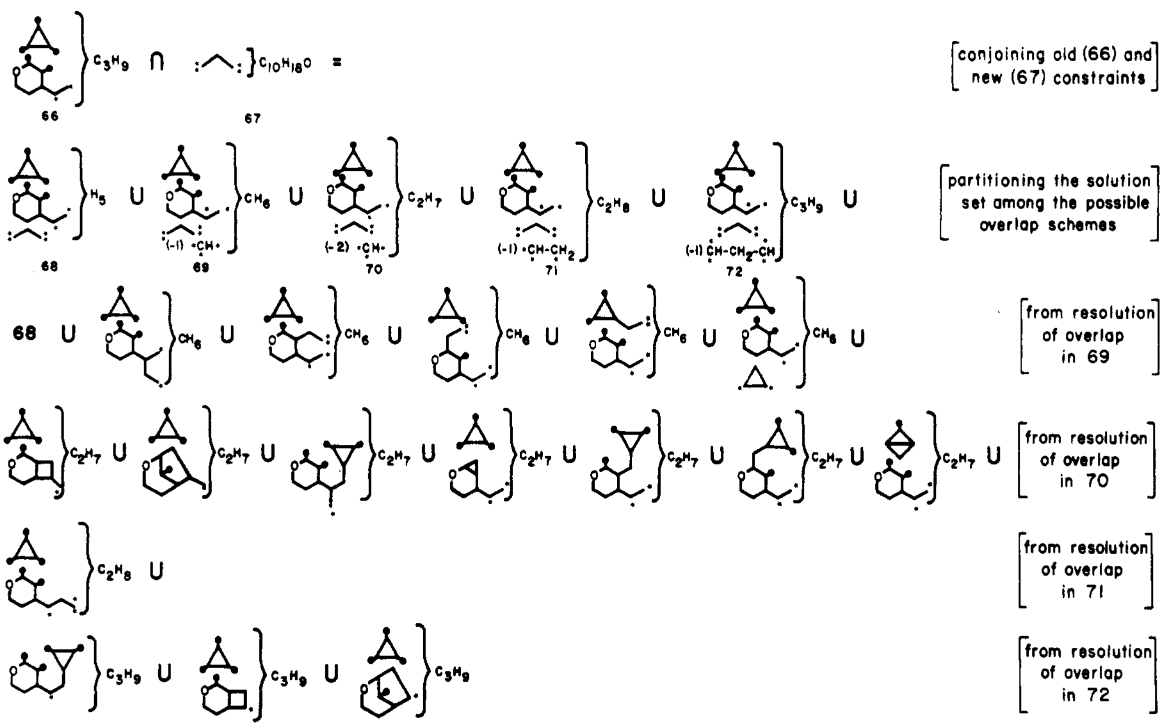

66
67
C3H9 ∩ C10H18O =
conjoining old (66) and new (67) constraints
68
69
70
71
72
partitioning the solution set among the possible overlap schemes
68 U
from resolution of overlap in 69
from resolution of overlap in 70
from resolution of overlap in 71
from resolution of overlap in 72

Figure 4. Fragment, expressed in ABSF, of a constructive substructure search in the $C_{13}H_{22}O$ isomer power set, after Carhart, Smith, Gray, Nourse, and Djerassi.[35]

RELATIONSHIP OF ABSF TO SSS, GENOA AND GENSAL

Any GSF language is obviously closely related to substructure search. The latter asks "What are the known members of the SF class defined by the following common substructures?" The former asks "With what set of attributes can I give an intensional definition of some particular set of SFs?" and generally finds that the most likely attributes are substructures. Some qualitative differences are worth noting. In defining search queries, substructure search systems (SSS) face a minimization problem analogous to those of finding simplest switching circuits or BSF representations alluded to in an earlier section. However, the definition of simplicity changes in at least two ways from that for ABSF: (a) The criteria of conciseness and readability particularize to numerous questions governing lucidity of query composition, ease of input, compactness of storage, screening speeds, etc. (b) Complete accuracy in SF-class definition may be sacrificed for search efficiency if diminished precision in the retrievals can be tolerated.

Substructure search systems' technical *competence* has progressed rapidly to direct graphic input/output capability,[34] eliminating the need for any mediating language of screen codes. It does remain to be seen at what rate behavioral and economic factors permit the *performance* of large numbers of chemists in manipulation of SF classes (for data base searching, structure elucidation, or other problems) to be upgraded by computer support. At least until the time when our daily chemical scribbling is done at a terminal, well-made paper-and-pencil SF languages will have a role to play.

GENOA,[35] an extension of/adjunct to CONGEN,[6,24] is a minicomputer-based system for inferential manipulation of descriptions of SF classes. It is most commonly used for reducing inputs of chemical and spectroscopic data to exhaustive and irredundant SF-set representations that constitute (partial) solutions to structure-elucidation problems. GENOA is the first mechanized system to have the capability of handling the combinatoric elaboration of overlapping substructures automatically. Bearing in mind that GENOA is a fully automated formal *system*, not just a formal language, it is of interest to compare its means of describing isomer subsets with that of ABSF. Both representations are intensional; GENOA produces the extension of its representation, as a set of complete SFs, only when given the GENERATE command. As the examples presented below illustrate, a considerable isomorphism exists between the representation of SF classes in GENOA and ABSF. In several respects, ABSF can serve as graphic realization of GENOA representations. This suggests a simple-minded but not necessarily trivial use of ABSF in facilitating descriptions of the use/operation of GENOA and other mechanized systems that manipulate SF-class information, in a system-independent language.

As an example, Figure 4[36] depicts (with some amplification) the content of Figure 3 from the published description of GENOA;[35] it appears to have three advantages: (a) discursive elements of the individual class representations have been fully iconized (the BSFs are unitary presentational signs); (b) an additional layer of the overlap-reduction logic is readily portrayed; (c) fully formalized representations replace ad hoc ones.

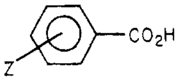
A second set of examples shows how GENOA describes substructure features and their logical relations. The GENOA commands producing constructs equivalent to BSFs **48** and **47** are shown as **63** and **64**.

```
DEFINE MOLFORM C 4 H 10 O  DEFINE MOLFORM C 4 H 10 O
DEFINE SUBSTRUCTURE Y      DEFINE SUBSTRUCTURE A
      CHAIN 2                    CHAIN 3
      ATNAME 2 O                 ATNAME 3 O
      HRANGE 122 211             HRANGE 122 222 300
CONSTRAINT Y NONE          DEFINE SUBSTRUCTURE B
             63                  CHAIN 2
     (equivalent to 48)          ATNAME 1 O
                                 HRANGE 111 211
                              ALTERNATIVE A 0* B 0*
                                          64
                                  (equivalent to 47)
```

*⟨score⟩ = relative plausibility rating

Through the courtesy of the DENDRAL group at Stanford University we have been able to check a substantial number of ABSF representations against GENOA as a benchmark system. Thus we have confirmed for **63**/**48** and **64**/**47**, and all of the other examples in this article, that the isomer subsets inferred by GENOA from (translated) BSF inputs are, in fact, identical with the extensions of the BSFs obtained by applying the definition of the ABSF language and the feature-extraction and construction algorithms given in the Appendix to those isomer subsets.

Markush notation was designed to provide maximum flexibility in denoting compositional variation of (generally peripheral) substructures in otherwise standard SFs. Its ability to portray variable connectivity (e.g., in **65**) is slight and essentially accidental. Thus there is little overlap between the Markush SF (MSF) and ABSF languages, but it is a simple extension, already anticipated in Table I, to add locally defined variables to ABSF. We do not give a formal description of this merged graphic language, but part 3 of this series will contain a formal definition of the connection tables capable of representing the fully extended (ABSF + NV + MSF) language (abbreviated BNMSF).



65, Z = CH_3O, CH_3, Cl, N≡C, O_2N

The first computer-manipulable language to capture the full expressive power of Markush SF notation is GENSAL, recently devised by Lynch and his students.[5] GENSAL encompasses, in addition to Markush variables, all global-variable notations, and hence the above NVSF language. GENSAL is designed primarily to handle the combinatorics of multiple (including nested) Markush-variable substituents on an essentially fixed skeletal core. Though not designed for the purpose, GENSAL also has some expressive power for homocomposite, connectivity-variable SFs. Variable substituent locations on the core and minor connectivity variations within the core must be handled extensionally, as lists of alternatives.

GENSAL is chemist and machine readable, whereas BNMSF is chemist readable but relies on connection tables for computer representation. On the one hand, BNMSF is intended primarily for chemist–chemist communication and paper-and-pencil inference (in connection with other languages), and it hews as closely as possible to traditional, graphic SF notation. GENSAL, on the other hand, is intended primarily for chemist–data base communication and for its unity of representation willingly sacrifices some chemist readability. It provides an elegant solution to the problem of formalizing and mechanizing Markush notation. While in principle GENSAL could subsume all SF notation using structural variables, global as well as local, the common globally defined structural variables (those normalized in NVSF) are so well entrenched in chemists' usage that no one is likely to suggest giving them up in favor of GENSAL's locally defined variables.

Finally, certain features that are not directly expressible as substructures are useful in extending flexibility of expression in languages of SF classes. "Number of rings" is an example that CONGEN and GENSAL incorporate but ABSF does not.

ACKNOWLEDGMENT

This work was supported, in part, by Chemical Abstracts Service, a division of the American Chemical Society, by National Science Foundation Grant SED 8020159, and by Kent State University. We gratefully acknowledge both this support and the appointment of J.E.G. as a Chemical Abstracts Service Visiting Research Scientist during 1980–1981 that made much of this work possible. We are grateful to the Stanford University DENDRAL project for providing access to CONGEN, GENOA, and STRCHK on Stanford's SUMEX facility.

APPENDIX. FORMAL DEFINITIONS

Well-Formed Brace Structural Formula (Structural Formula, Molecular Formula). Definition 1. A molform is a string of the form

$$A_1\alpha_1A_2\alpha_2 \ldots A_i\alpha_i \ldots A_n\alpha_n$$

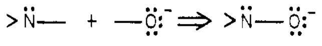
where $\{A_1, A_2, ..., A_n\}$ is the set of atomic symbols, and the α_i are nonnegative integers. The α_i are constrained to such values that the sums, over the molform, of (a) the valences of the polyvalent A_i and (b) the α_i for monovalent A_i have the same parity. Either or both of the following abbreviations may replace a molform in any context without alteration of meaning: (a) Any $\alpha_j = 1$ may be deleted. (b) For any $\alpha_j = 0$, both α_j and A_j may be deleted.

Definition 2. The signs in Chart II together constitute the set, **U**, of *atomic bonding units* (ABUs).

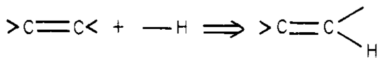
Definition 3. The signs "—", "═", "≡", ":", "·", "+", and "-" are called *electronic symbols*. The signs "—" (or —●), "═" (or ═●), and "≡" (or ≡●), when juxtaposed to a single atomic symbol, are called *open bonds* of types 1, 2, and 3. A pair of distinct ABUs may be aggregated if and only if the members of the pair (a) each have an open bond of the same type and (b) have not previously aggregated with one another. *Aggregation* is defined as the formal operation of coalescing or overlapping two open bonds of like type, each of which is juxtaposed to a distinct atomic symbol, as in

$$>C= + =C< \Rightarrow >C=C<$$

or

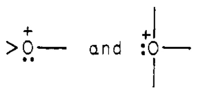


or



An atomic symbol together with its immediately juxtaposed electronic symbols is still identified as an ABU of some particular variety after aggregation with (an)other ABU(s). The single, double, and triple bond signs, when juxtaposed to *two* atomic symbols, are no longer open and are called single, double, and triple *bonds*, respectively. The compound signs resulting from aggregation are called *part structures* if they contain one or more residual open bonds or *structural formulas* (SFs) if they do not. A part structure with exactly one open bond is a *group*. If SFs are construed as labeled graphs, then any proper subgraph of a well-formed SF is a *substructure*. *Radical* is a term used informally to mention a part structure or a substructure.

Definition 4. ABUs are types whose tokens may have their electronic symbols juxtaposed in any order or orientation. Thus



are well-formed tokens of the same type and have identical denotations.

Definition 5. A well-formed brace structural formula (BSF) is the result of applying one of the sequences of steps permitted by the morphological diagram shown in Chart III.

Step 1. To the left of a molform juxtapose a right-hand brace, }, and to the right juxtapose a superscript zero. The portion of this and subsequent formulas that lies to the right of the brace is called a *residue molform*. Signs placed to the left of the brace are said to be under the brace. A residue molform with no terminal superscript is taken by convention as an abbreviation of the same residue molform with superscript 0.

Chart II

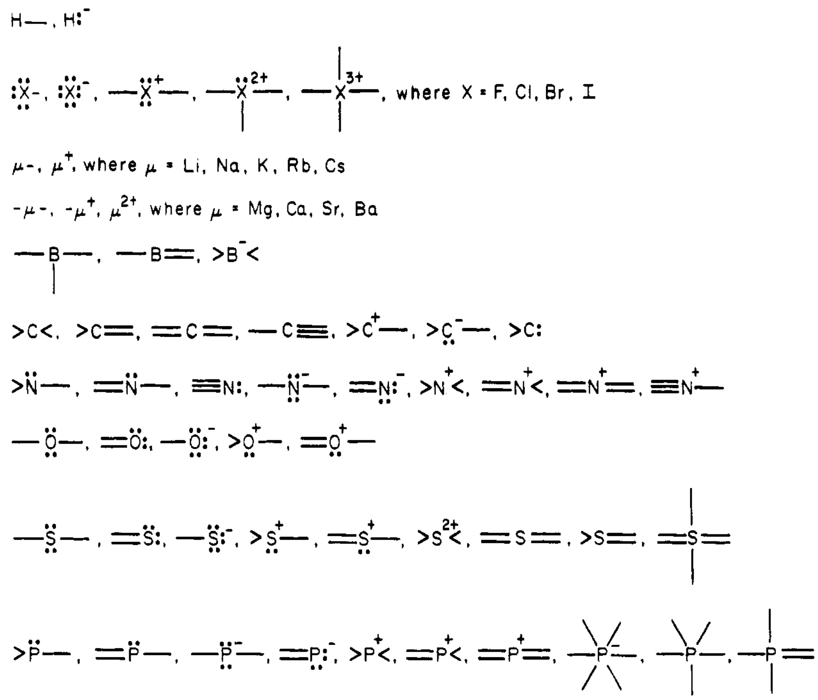
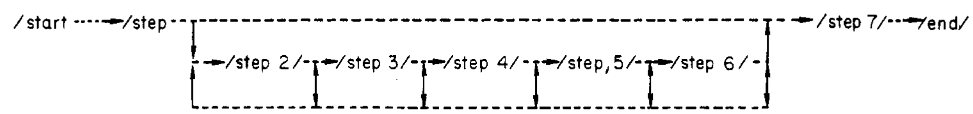

where X = F, Cl, Br, I
μ-, μ+, where μ = Li, Na, K, Rb, Cs
-μ-, -μ+, μ2+, where μ = Mg, Ca, Sr, Ba

Chart III



Step 2. Select any atomic symbol represented in the residue molform, decrement the integer to its immediate right by one unit, place a copy of the atomic symbol under the brace, and add the necessary electronic symbols to expand the copy to an element of the set of ABUs. For each plus (minus) sign placed under the brace in this step, add −1 (+1) to the superscripted suffix of the residue molform. Arrange all ABUs in a single column under the brace.

Step 3. Aggregate a pair of open bonds under the brace [subject to the restriction that the resulting aggregate may be an individual structural formula only if (a) no open bonds remain under the brace and if (b) all integers in the residue molform are zero]. Arrange all ABUs/ABU aggregates in a single column under the brace.

Step 4. Optionally replace replicate ABU or substructure tokens under the brace by a single token suffixed to a *quantifier* consisting of an integer in curves, for example

$$\left.\begin{array}{l}\text{—Cl}\\ \text{—Cl}\\ \text{—OH}\end{array}\right\} \Rightarrow \left.\begin{array}{l}(2)\ \text{—Cl}\\ \quad\ \ \text{—OH}\end{array}\right\}$$

Optionally add to the intrabrace column (a) any desired ABU or substructure to which is prefixed "(0)". Optionally reorder the elements of the intrabrace column. Optionally bind any contiguous subcolumn of ABUs/substructures bearing (0) prefixes by juxtaposing to their immediate left an enclosing left-hand brace, {. Optionally prefix (0) to any left-hand brace.

Step 5. Replace each element of any subset of the open-bond tokens of type 1 under the brace by an asterisk, token for token. The asterisk is called a *free valence.*

Step 6. If m instances of a part structure, a, occur under the brace such that (a) each occurrence is as a subgraph of some other intrabrace part structure and (b) neither occurrence lies within the scope of a { or (0) token, then one $(-n)$p token may be optionally added to the intrabrace column (where n is a positive integer and $n < m - 1$) provided that for each atomic symbol, A_i, in p, the corresponding α_i in the residue molform is increased by $n \times x$, where x is the number of occurrences of A_i in p. Obligatorily reorder any elements of the intrabrace column having the form $(-n)$p such that at least $n + 1$ occurrences of p appear as substructures of part structures lying below $(-n)$p in the intrabrace column; such part structures are said to be in the scope of the $(-n)$p token. Optionally prefix pairs of # (%, &, etc., as necessary so that no such type is reused) tokens to pairs of part structures containing a as a subgraph and lying in the scope of a $(-n)$p token, provided that at least one such part structure remains unmodified.

Step 7. If there are no open bonds or free valences under the brace, optionally delete the brace. The result is either a molecular formula or an individual structural formula, either of which is by convention an abbreviation of a well-formed BSF. Optionally rewrite any instance of

$$(0)\left\{\begin{array}{l}(0)a\\ \vdots\\ (0)x\end{array}\right.$$

where a, ..., x are part structures, as

$$\left\{\begin{array}{l}a\\ \vdots\\ x\end{array}\right.$$

Well-formed Normalized-Variable Structural Formula. A well-formed NVSF results from application of one or more of the following transformations to any well-formed individual SF, s.

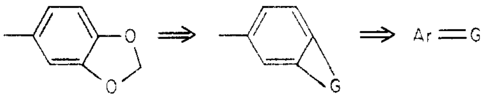
(1) Any atomic-symbol token of type F, Cl, Br, or I in s may be replaced with one of type X.

(2) Construe s as a labeled graph or multigraph whose nodes are atomic symbols or atomic symbols to which electronic symbols have been juxtaposed. Identify a subgraph, g, of s,

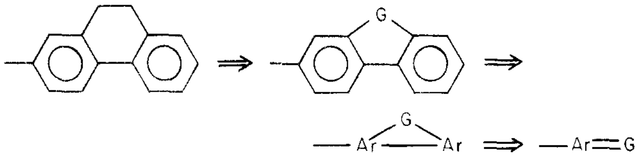
and conceptually delete all of the nodes of s that are not nodes of g. Retain as an open bond each edge that formerly connected a node of g to a now-deleted node. Call the resulting part structure p, and let the number of open bonds on p be *o*. The permissible substitutions for g then depend in part upon *o*. (a) If $o = 1$ and the nodes of g are all tokens of the type-class {>C<, –H}, then g may be replaced with the formative R–. (b) If (i) every node of p is a token of the type class {H—, >C=, —$\ddot{N}$=, >N$^+$=, —$\ddot{O}^+$=} and (ii) every node of p that bears an open bond is a node of at least one cyclic subgraph, c, of p such that the cycle in c consists of 6 + 4*n* nodes ($n = 0, 1, ...$) and of alternating single and double edges, then g may be replaced with the formative Ar. However, if g contains positively charged nodes, then Ar replaces g *and* its counterionic part structure(s). (c) In any case, g may be replaced with the formative G.

This transformation obligatorily applies (recursively) if prior transformations have produced two adjacent nodes of type G or of type Ar: one of the two tokens and the edge(s) joining the two are deleted, and the deleted token's other edges, if any, are transferred to the retained token.

More than one connection between G and Ar can arise from the above transformations in several ways, and these come to be written as multiple bonds. For example, see eq 13 and 14.



(13)



(14)

Since the definition of Ar excludes the possibility of multiple connections to the same site in Ar, such multiple bonds are obligatorily construed as several independent connections to different sites in Ar/G.

Algorithm for Constructing a Distinctive Substructure-Feature Matrix on the Isomer Space of Composition *j*. (1) Let *j* index compositions, *i* index SFs (and hence columns in the distinctive-feature matrix), and *k* index substructures (and hence rows in the feature matrix). The set of all SFs of composition *j* is $\mathbf{S}_j$; n_j is the cardinality of $\mathbf{S}_j$.

(2) Let $\mathbf{Q}_{i,j}$ be the set of single bonds between polyvalent atoms in structure $s_{i,j}$, and let $q_{i,j}$ be its cardinality. Form the power set, $2^{\mathbf{Q}_{i,j}}$, of $\mathbf{A}_{i,j}$. Carry out $2^{q_{i,j}} - 1$ distinct bond-breaking operations on $s_{i,j}$, each operation corresponding to breaking all the bonds in one element of the power set of $\mathbf{Q}_{i,j}$, i.e., all the bonds in one of the nonempty subsets of $\mathbf{Q}_{i,j}$. Discard duplicate fragments. Repeat the $2^{q_{i,j}}$ operations for each $s_{i,j} \in \mathbf{S}_j$ and collect all of the resulting unique fragments as elements, $f_{k,j}$, of the set $\mathbf{F}_j$. The $f_{k,j}$ are part structures in which the original location of each operand bond is marked as a free valence with an asterisk.

(3) Construct an empty matrix with n_j columns representing the SFs of the isomer set $\mathbf{S}_j$. For each $f_{k,j} \in \mathbf{F}_j$, generate a row of binary values in the feature matrix: successive elements are equal to 1 if and only if $f_{k,j}$ is a subgraph of isomer $s_{i,j}$ or are equal to 0 otherwise. Let each $f_{k,j}$ label its respective row.

(4) Discard rows whose entries duplicate an existing row by using the following rules, in order, to decide which row will be retained: (a) Retain the $f_{k,j}$ of lowest molecular weight; (b) among $f_{k,j}$s of equal molecular weight, retain the one (i) whose asterisked atoms have the higher atomic weight or (ii) whose substituents have the higher Cahn–Ingold–Prelog priorities. For this purpose construe the asterisk as an atom of highest priority.

Chart IV

form of A'_j	replace A'_j by
$f_{i,j}$	$d(f_{i,j})$
$\urcorner f_{i,j}$	$(O)\ d(f_{i,j})$
$f_{i,j}\ \&\ f_{i',j}$	$d(f_{i,j})$ $d(f_{i',j})$
$\urcorner(f_{i,j}\ \&\ f_{i',j})$	$\begin{cases}(O)\ d(f_{i,j}) \\ (O)\ d(f_{i',j})\end{cases}$
$f_{i,j} \vee f_{i',j}$	$\begin{cases}d(f_{i,j}) \\ d(f_{i',j})\end{cases}$
$\urcorner(f_{i,j} \vee f_{i',j})$	$(O)\ d(f_{i,j})$ $(O)\ d(f_{i',j})$

(5) If the row whose elements are all 1 is missing, add it and label it with the null feature considered to be a substructure of every SF. In any case, add the row whose elements are all 0, labeled with some impossible structure-feature such as a quadruple bond to carbon.

(6) Construing the rows as binary integers, arrange them in decreasing numerical order and renumber them 1 to *m*. Call this subset, $\{f_{1,j}, f_{2,j}, ..., f_{m,j}\}$, of $\mathbf{F}_j$ the *primary feature set.*

(7) Expand the feature set to include the features $f_{a,j}$ & $f_{b,j}$ and $f_{a,j} \vee f_{b,j}$, etc., for all $f_{k,j} \in \mathbf{F}_j$ such that $a \neq b$; call these additional features *secondary features.*

(8) Repeat step 3 for the secondary features, and discard duplicate rows by using the following orders of preference in retention: (a) retain a primary in preference to a secondary feature; (b) retain secondary conjunctions in preference to secondary disjunctions; (c) retain the secondary feature composed from the lowest numbered primary feature(s) (via 1:1 comparison of features arranged in increasing numerical order).

(9) For each of the possible n_j-digit binary numbers containing exactly one instance of 1 that does not yet appear as a row in the feature matrix, generate such a row by adding to the list of features the substructure obtained by deleting an instance of H from the appropriate $s_{i,j} \in \mathbf{S}_j$.

(10) Construing each row of the feature matrix as an n_j-digit binary number, rearrange the rows in decreasing numerical order. Call the result the *canonical feature matrix* (CFM).

(11) For convenience in referring to the CFM, label the features/rows, top to bottom, as $A_1, A_2, ..., A_k, ..., A_r$. Let *l* index the columns of CFM ($1 < l < n_j$). Represent an arbitrary element of CFM by $B_{k,l} \in \{0,1\}$ and an arbitrary row by R_k. Define the indicative substructure of isomer *x*, z_x, as the one corresponding to that row of the CFM in which only the element $B_{k,x}$ has the value 1 and all other elements, $B_{k,l}$ ($l \neq x$), have the value 0.

Algorithm for Constructing a BSF Representing Any Arbitrary Isomer Subset of $\mathbf{S}_j$. (1) Produce the canonical feature matrix (CFM) for composition *j*. Represent the cardinality of $\mathbf{S}_j$ by n_j.

(2) Represent the desired set as an n_j-digit binary number obtained by considering the $s_{i,j}$ in the order of occurrence of columns in the CFM and for each digit/column/isomer, writing 1 if that isomer is present and 0 if it is absent from the desired set. Call the resulting number *w* and its digits D_l ($1 < l < n_j$).

(3) Execute the subalgorithm shown in Figure 5.

(4) Construe the value of *string* as a Boolean polynomial of the type

$$A'_i \vee A'_j \vee ... \vee A'_p$$

in which A'_i is either a name of a feature, A_i (substructure or conjunction/disjunction of substructures), or a negation of a feature

$$\urcorner A_i$$

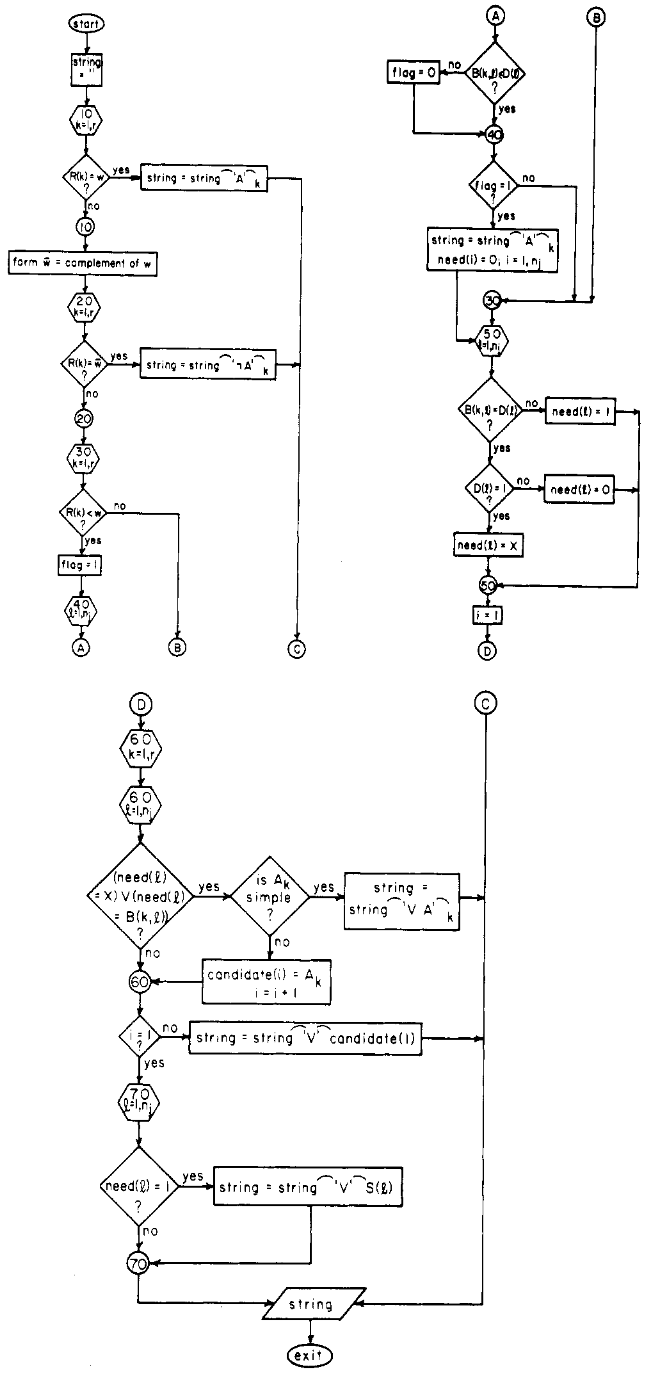


Figure 5. Subalgorithm for completion of the distinctive feature matrix and/or construction of the corresponding ABSF representations for the power set of composition j. Hexagons begin loops that end with the label number contained in the hexagon.

and $A'_j \dots A'_p$ have the same form or are nulls. Translate *string* into a BSF by the following steps: (a) Place under the main brace either A'_i or

$$\left\{\begin{array}{l}A'_i\\ \vdots\\ A'_j\\ \vdots\\ A'_p\end{array}\right.$$

as the case may be, omitting any empty A'_j. (b) Represent each A'_j by substructures read from the CFM according to Chart IV. Let d(s) represent the drawing of the graph of s. (c) To the right of the main brace add a molform computed by subtracting from composition j the composition of each intrabrace $s_{i,j}$ that is not in the scope of any instance of (0).

REFERENCES AND NOTES

(1) Correspondence should be directed to the Kent State University address.

(2) Valence, E. H. "Understanding the Markush Claim in Chemical Patents". *J. Chem. Doc.* **1961**, *1*, 87–92.

(3) Sneed, H. M. S.; Turnipseed, J. H.; Turpin, R. A., Jr. "A Line-Formula Notation System for Markush Structures". *J. Chem. Doc.* **1968**, *8*, 173–178.

(4) Deforeit, H.; Caric, A.; Combe, H.; Leveque, S.; Malka, A.; Vals, J. "CORA. Semiautomatic Coding System. Application to the Coding of Markush Formulas". *J. Chem. Doc.* **1972**, *12*, 230–233.

(5) Barnard, J. M.; Lynch, M. F.; Welford, S. M. "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 2. GENSAL, a Formal Language for the Description of Generic Chemical Structures". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 151–161.

(6) (a) Carhart, R. E.; Varkony, T. H.; Smith, D. H. "Computer Assistance for the Structural Chemist". In "Computer-Assisted Structure Elucidation". Smith, D. H., Ed.; American Chemical Society: Washington, DC, 1977; Chapter 9. (b) Carhart, R. E.; Smith, D. H. "Applications of Artificial Intelligence for Chemical Inference. XX. Intelligent Use of Constraints in Computer-Assisted Structure Elucidation". *Comput. Chem.* **1976**, *1*, 79–84. (c) Carhart, R. E.; Smith, D. H.; Brown, H.; Djerassi, C. "Applications of Artificial Intelligence for Chemical Inference. XVII. An Approach to Computer-Assisted Elucidation of Molecular Structure". *J. Am. Chem. Soc.* **1975**, *97*, 5755–5762.

(7) Yamasaki, T.; Abe, H.; Kudo, Y.; Sasaki, S.-I. "CHEMICS: A Computer Program System for Structure Elucidation of Organic Compounds". In "Computer-Assisted Structure Elucidation". Smith, D. H., Ed.; American Chemical Society: Washington, DC, 1977; Chapter 8.

(8) (a) Pensak, D. A.; Corey, E. J. "LHASA—Logic and Heuristics Applied to Synthetic Analysis". In "Computer-Assisted Organic Synthesis"; Wipke, W. T., Howe, W. J., Eds.; American Chemical Society: Washington, DC, 1977; Chapter 1. (b) Orf, H. W. "Computer-Assisted Synthetic Analysis". Dissertation, Harvard University, 1976.

(9) Wipke, W. T. "SECS—Simulation and Evaluation of Chemical Synthesis: Strategy and Planning". In ref 8a, chapter 5.

(10) Dugundji, J.; Ugi, I. "An Algebraic Model of Constitutive Chemistry as a Basis for Chemical Computer Programs". In "Computers in Chemistry"; Veal, D. C., Ed.; Springer-Verlag: West Berlin, 1973; p 19.

(11) Cajori, F. "A History of Mathematical Notations"; Open Court: La-Salle, IL, 1928; Vol. 1, pp 75, 381.

(12) Examples for this period showing a variety of replacement sets may be found in: *Justus Liebigs Ann. Chem.* (e.g.: **1876**, *181*, 391; **1878**, *193*, 35; **1880**, *205*, 355; **1881**, *206*, 309).

(13) Locally defined variables range over the solution set for the sentence(s) in which they occur; the solution set is that subset of the replacement set each of whose elements makes the sentence(s) true when it replaces the variable. This truth-functional connotation of local is not attached to chemical-structural variables, which may correspond to quite arbitrarily assigned replacement sets.

(14) Morrison, R. T.; Boyd, R. H. "Organic Chemistry"; 3rd ed.; Allyn and Bacon: Boston, 1973.

(15) Freund, M. "Beitrag zur Kenntniss des Cevadins". *Chem. Ber.* **1904**, *37*, 1946–1957.

(16) Vongerichten, E. "Ueber die stickstofffreien Spaltungsproducte des Morphins". *Chem. Ber.* **1898**, *31*, 51–56.

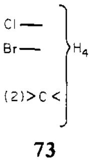
(17) Gautier, A. "Ueber die Einwirkung der Saüren auf die Carbylamine". *Justus Liebigs Ann. Chem.* **1869**, *151*, 239–244.

(18) Kondo, H.; Kondo, T. "Über das Alkaloid 'Coclaurin' von Cocculus laurifolius, D. C.". *J. Prakt. Chem.* **1930**, *126*, 24–52.

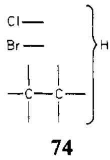
(19) Bunge, M. "The Furniture of the World"; D. Reidel: Dordrecht, 1977; (a) p 57, (b) 144.

(20) Feys, R.; Fitch, F. B. "Dictionary of Symbols of Mathematical Logic"; North-Holland Publishing Co.: Amsterdam, 1969; p 95.

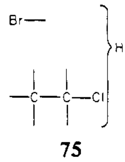
(21) At first glance one is tempted to object that the examples offered in Figure 2 for types 2 and 4 do not, as maintained, portray zero connectivity information. Indeed, since carbon is the only polyvalent atomic symbol in the residue molform of example 2, the double bond to oxygen can only terminate in carbon, and the structure can immediately be rewritten as >C═O} C_4H_8. Similarly, the two >C< part structures in the following example must be connected, so that **73** can be immediately rewritten as **74** or, indeed, as **75** without altering its extension, {CH_2ClCH_2Br, $CH_3CHBrCl$}. These *rewritten* BSFs do indeed contain connectivity information. However, there is strong motivation for keeping the syntax of the BSF language free of such inferences. We wish the language to be able to portray all steps in any lines of argument involving SF classes, without having some inferential moves preempted by autoinferencing within the BSF language.



73



74



75

(22) In the cases like example 3 one could adapt the ↔ connective to reduce the representation to a single BSF, say

$$-^{+}N{\equiv}N: \leftrightarrow\ -N{=}N:^{+}\}C_{11}H_6O_5S^{-}$$

but similar solutions do not exist for the other cases.

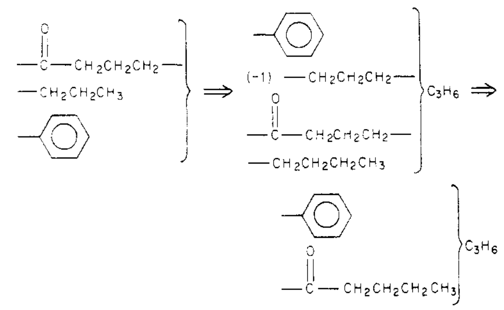
(23) We refer to the set of all SFs dominated by a single molform as the *isomer set* of the molform. Any motivated or arbitrary collection of isomeric SFs is then an *isomer subset.*

(24) CONGEN[6] appears to have complete expressive power for all homocomposite SF classes. It thus provides a convenient benchmark for other class–SF languages.

(25) TBSF expresses conjunction of substructures by juxtaposition under the brace; thus **9** denotes the class of SFs of $C_{12}H_{18}O$ possessing an –OH AND a phenyl group as subgraphs. Any disjunction is expressible by using only conjunction and negation:

$$a \vee b <\!\!=\!\!=\!\!> \neg(\neg a \mathbin{\&} \neg b)$$

(26) Discovery of overlap does not always reduce class cardinality, as this example shows:



(27) Thomas, J. A. "Symbolic Logic"; Merrill Publishing Co.: Columbus, OH, 1977; Chapter 15.

(28) Mendelson, E. "Boolean Algebra and Switching Circuits"; McGraw-Hill: New York, 1970; Section 4.7 ff.

(29) (a) Sussenguth, E. H., Jr. "A Graph-Theoretic Algorithm for Matching Chemical Structures". *J. Chem. Doc.* **1965**, *5*, 36–43. (b) Figueras, J. "Substructure Search by Set Reduction". *J. Chem. Doc.* **1972**, *12*, 237–244.

(30) One piece of arbitrariness involved limiting the search for conjunctive/disjunctive secondary features to two-way combinations before switching to the last-resort strategy embodied in the construction algorithm in the Appendix. In the $C_4H_{10}O$ example, 14 of the 128 rows remain to be described after primary features, their negations, and their two-way combinations, or negations of these, have been exhausted. To seek to complete these rows by systematic investigation of three-way combinations would potentially require testing 293 760 such combinations. Of course, a certain number of expressions higher than two way result from two-way combination of two-way features.

(31) For example, see: Hunt, E. B. "Concept Learning"; Wiley: New York, 1962.

(32) Proof takes this form: Let the m_j rows of the minimal set be numbered $x_1, x_2, ..., x_m$. Produce a new row, d_1, by conjoining rows x_1 through x_m, but in so doing use the negation of feature x_i, rather than the feature itself, for just those rows x_i whose element in column 1 is 0. Since the column-1 elements conjoined are then all 1s, the new row d_1 has the value 1 in column 1. Since the entries in at least one of the rows x_1 to x_m are distinct for every column, the column-2, ..., column-n elements conjoined will contain at least one 0, and the row d_1 element for all of these columns will be 0. Row d_1 is thus the indicative row for isomer 1. Repetition provides the indicative rows for the remaining columns, and (continued) disjunction of these produces any desired row as previously discussed.

(33) The unsaturation index is computed for complete molforms (no net charge or open bonds) or molecular ions (complete molforms plus or minus electrons) from $\Delta = c + 1 - h/2 - x/2 + n/2$, where c, h, x, and n are the molform subscripts on C, H, halogen, and N. To compute Δ for an isolated ion, an additional term, $-q/2$, must be added, where q is the absolute value of the (net) charge on the ionic molform.

(34) "Structure Diagrams Can Be Used for Substructure Searches". *Chem. Eng. News* **1981**, Dec 14, 30.

(35) Carhart, R. E.; Smith, D. H.; Gray, N. A. B.; Nourse, J. G.; Djerassi, C. "GENOA: A Computer Program for Structure Elucidation Utilizing Overlapping and Alternative Substructures". *J. Org. Chem.* **1981**, *46*, 1708–1718.

(36) Figure 4 depicts a precise, meaningful sentence in the language of set theory, except for one liberty taken in its presentation: operands for the set operations are actually the extensions of the BSFs, $K_B(b_i)$, rather than the b_i themselves. The first row represents merging of intensional GSF descriptions of the same substance based on two separate sets of experimental data, using a rule of type d (above). The nominal result, **68**, is expanded in the second row to include the possibility of overlap (**69**–**72**). Succeeding rows are expansions of these individual overlap schemes according to the principles in the fourth section of this article.

(37) *J. Org. Chem.* **1980**, *45*, 3545–3730.

(38) *Chem. Abstr.* **1980**, *93*, No. 7 and 10.