Computer Storage and Retrieval of Generic Chemical Structures in Patents. 15. Generation of Topological Fragment Descriptors from Nontopological Representations of Generic Structure Components

J. D. Holliday,* G. M. Downs, V. J. Gillet, and M. F. Lynch

Department of Information Studies, University of Sheffield, Sheffield S10 2TN, U.K.

Received November 6, 1992

The generation of topological fragments from generically expressed components of generic structures is described. Fragments derived wholly from within a component or partial structure (PS) are termed intra-PS fragments, while those which span partial structures are termed inter-PS fragments. The generation of fragments from specific PS's and the linkage of inter-PS part-fragments is described in the preceding paper in this series. Generic partial structures, described in intensional terms by homologous series identifiers (HSIs), are represented in the extended connection table representation (ECTR) as a list of default or explicit parameter values. The method described here compares the numerical parameters of the HSI with similar parameters derived from each topological fragment descriptor in the screen set in order that a subset of fragments can be identified. This subset represents those fragments of the screen set which lie within the scope of the intensional description. These fragments are thus identified by means of a search of the entire fragment set. A tree-structured dictionary of fragments (augmented atoms and linear atom/bond sequences) is used as the basis for generation, using a depth-first trace. The leaves of the dictionary tree point to the parameter descriptions of the fragment being generated. All fragments which are identified in this way are regarded as being optional to the structure and contribute to the optional component (POSS screen) of the two-part bit screen representation of the generic structure.

1. INTRODUCTION

Generic structures comprise certain components which are described in terms of atoms and bonds, and other components which are described intensionally as homologous series identifiers (HSIs), such as "alkyl", "alkenyl", or "carbocyclyl". Homologous series identifiers describe one of a possibly infinite number of structures which are homologous to each other, i.e., exhibit homology variation, or h-variation, as described by Dethlefsen.[2]

The preceding paper in this series[1] describes the generation of fragments, and subsequent assignment of bit screens, for generic structures which exhibit all types of structural variation except h-variation. The generation of inter-PS and intra-PS fragments (PS = partial structure) is described together with the "bubble-up"[3] of fragments through the ECTR.[4]

This paper describes the generation of fragments from h-variant components of generic structures, represented as a list of default or explicit parameter values. These are single values or numerical ranges describing the features attributable to the set of possible structures denoted by the HSI expression.

The tree representation of complete and incomplete fragments, described in the preceding paper in this series, comprises the set of all fragments in the screen dictionary. The generation of all such fragments for a generic structure would result in a full (or "black") bit screen. The function of fragment generation is to identify the subset of this finite set which describes an HSI within its context. The approach taken here is to eliminate those fragments which are not covered by the HSI description, in order to identify the subset which covers a specific partial structure.

A common representation must be established to permit comparison of the topological description of the fragment with the numerical parameter values of the HSI. A possible approach is to represent each fragment by its own list of parameters, the values of which are derived from the topological representation of the fragment. These parameters are called *specific derived parameters*.

Section 2 describes the standard set of specific derived parameters and discusses some of the limitations of this set for representing HSIs. Section 3 describes an alternative form of parameter list which helps to overcome these problems; the derivation of parameter values for fragment descriptors is explained. In order that a comparison be made between the fragment descriptor and the HSI, similar information must be used to represent the HSI. Section 4 describes the format of the HSI representation together with the derivation of respective values. The two sets of parameters, those used to represent the fragment descriptor and those used to represent the HSI, must be compared in order that the fragment may be identified within the scope of the HSI description. The processes used for comparison are described in section 5.

2. SPECIFIC DERIVED PARAMETERS

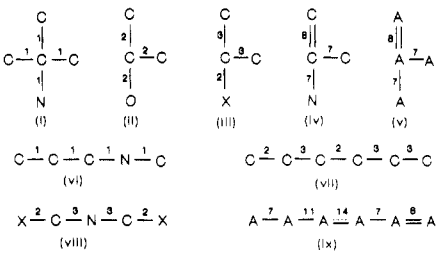
The homologous series identifiers are represented by means of a list of *parameters* which describe certain structural features. Each parameter has an associated value, or range of values, reflecting the status of the structural feature. Thus, in the parameter list of "C4-8 alkadienyl", the obligatory presence of 4–8 carbon atoms (C) and of two double bonds (E), and the possible presence of ternary branching (T), and the obligatory absence of heteroatoms (Z) are reflected in the corresponding parameter values $C\langle 4\text{–}8\rangle$ $E\langle 2\rangle$ $T\langle 0\text{–}\rangle$ $Z\langle 0\rangle$. A set of 13 parameters is used in the Sheffield work, as shown in Table I. The set is tentative and may be altered in a future system; much investigation has been carried out by International Documentation in Chemistry mbH in order to define a new set of parameters.[5]

The fragment types considered here comprise augmented atoms and linear sequences, including multiple types, sizes, and levels of description. The augmented atom fragments

0095-2338/93/1633-0369$04.00/0

Table I. Standard Parameters Used To Represent Generic Partial Structures

A	total non-hydrogen atom count
C	total number of carbon atoms
T	number of acyclic ternary branching atoms
Q	number of acyclic quarternary branching atoms
E	number of localized olefinic unsaturations
Y	number of localized acetylinic unsaturations
RC	number of rings
RN	number of ring atoms
RS	number of ring substitutions
RF	number of ring fusion atoms
RA	number of delocalized, aromatic rings
RZ	number of ring heteroatoms
Z	total number of heteroatoms



	A	C	T	Q	E	Y	RC	RN	RS	RF	RA	RZ	Z
Augmented Atoms													
(i) C 1C 1C 1N	5	4	0	0	0	0	0	0	0	0	0	0	1
(ii) C 2C 2C 2O	4	3	1	0	0	0	0	0	0	0	0	0	1
(iii) C 3C 3C 2X	4	3	0	0	0	0	1	3	1	0	0	0	1
(iv) C 8C 7C 7N	4	3	1	0	1	0	0	0	0	0	0	0	1
(v) A 8A 7A 7A	4	0	1	0	1	0	0	0	0	0	0	0	0
Sequences													
(vi) C 1C 1C 1N 1C	5	4	0	0	0	0	0	0	0	0	0	0	1
(vii) C 2C 3C 2C 3C 3C	6	6	0	0	0	0	2	5	3	0	0	0	0
(viii) X 2C 3N 3C 2X	5	2	0	0	0	0	1	3	2	0	0	1	3
(ix) A 7A11A14A 7A 8A	6	0	0	0	1	0	2	3	2	1	1	0	0

Figure 1. Specific derived parameter values for example fragments.

Table II. Bond Descriptor Codes

bond definition	code
nonspecific	
any	1
intermediate	
chain	2
ring	3
specific	
chain single	7
chain double	8
chain triple	9
ring single	11
ring double	12
ring triple	13
ring aromatic	14

considered here are of the following type: augmented atoms (AA), i.e., a central atom with its non-hydrogen congener atoms and the bonds to them. Linear sequences comprise the following: (1) atom sequences (AS), being paths of connected atoms of length 4, 5, or 6 with bond information confined to the distinction between ring and chain bonds; (2) bond sequences (BS), being paths of connected bonds, 3, 4, or 5 bonds in length, without definition of the atoms which they connect.

The derivation of values for each of the thirteen parameters from a fragment descriptor is a trivial process. It involves either a traverse of a sequence fragment or an exploration of an augmented atom; the parameter values can be cumulated as this process is carried out. Some example fragments are shown in Figure 1 with their associated parameter values. The codes used to represent the bonds in these descriptors and in later examples are given in Table II. (Atoms are represented by their chemical symbol, by an "A" meaning "any atom" or by an "X" meaning "any halogen").

The values derived are cumulative counts of the features which are deduced from the fragment descriptor. For example, an acyclic bond adjacent to a cyclic bond denotes a ring substitution. Similarly, an aromatic bond adjacent to a nonaromatic, cylic bond denotes a ring fusion, as does the presence of three or more cyclic bonds in an augmented atom.

Little information can be derived from fragments of low specificity. For example, the fragment of Figure 1i gives values for just three parameters (*A*, *C*, and *Z*); no other parameter values can be deduced. Similarly, nonspecific atom definitions give no values for *C*, *Z*, or *RZ*.

The identification of a fragment within the scope of the HSI description requires that none of the parameter values of the fragment exceeds the maximum value of each corresponding parameter of the HSI. For example, the maximum parameter values used to represent the HSI *alkyl* are zero for all parameters except *A*, *C*, *T*, and *Q* whose values are unbounded. None of the fragments in Figure 1 would be identified by this HSI for the following reasons:

Fragments i, ii, and vi exceed the maximum limit of the *Z* parameter.

Fragment iii exceeds the maximum limits of the *Z*, *RC*, *RN*, and *RS* parameters.

Fragment iv exceeds the maximum limits of the *E* and *Z* parameters.

Fragment v exceeds the maximum limit of the *E* parameter.

Fragment vii exceeds the maximum limits of the *RC*, *RN*, and *RS* parameters.

Fragment viii exceeds the maximum limits of the *Z*, *RC*, *RN*, *RS*, and *RZ* parameters.

Fragment ix exceeds the maximum limits of the *E*, *RC*, *RN*, *RS*, *RF*, and *RA* parameters.

The fragment is, in loose terms, a substructure of at least one structure described by the HSI.

The limitations of this particular set of parameters for representing a fragment is that there is no indication of the environment in which the fragment occurs; a simple parameter comparison of this type is therefore incomplete. The values assigned to each feature must reflect the environment in which it is perceived. The fragment in Figure 1vii contains two ring components (the two cyclic bond paths are interrupted by an acyclic bond), each of which must have a minimum atom count of three. The values of the parameters *A*, *C*, and *RN* would therefore be 7, 7, and 6, respectively. The values must, however, be as low as are chemically and topologically allowable in order that every possible structure in which the fragment may occur is accounted for. The lowest values are referred to as *minimum inferred* parameter counts.

These limitations, described below, occur in three distinct areas:

(1) The set of atom count parameters is incomplete.

(2) Separate ring system parameters are grouped together.

(3) The different types of degree (or non-hydrogen connectivity) of atoms (or nodes) in the fragment are not fully reflected.

Atom Counts. The parameters used to quantify atom counts are as follows:

A	total no. of atoms
C	total no. of carbon atoms
Z	total no. of heteroatoms
RN	no. of cyclic atoms
RZ	no. of cyclic heteroatoms

This list is adequate for determining the structure denoted by the HSI since a complete list of parameters, which would include "no. of cyclic carbons", "no. of acyclic atoms", "no.

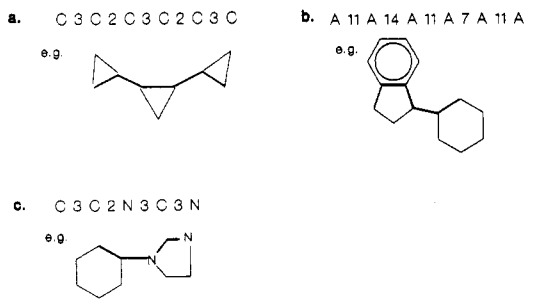


Figure 2. Fragments and possible structures.

of acyclic carbons", and "no. of acyclic heteroatoms", is not necessary. These new parameters are redundant since their values may be derived by calculation. This is because all atoms in the structure denoted by the HSI expression are either cyclic or acyclic. However, consideration of the sequence fragment

C 2 C 3 C 2 C 2 X

where 2 and 3 are the bond codes given in Table II, shows ambiguity in the representation. The cyclic nature of the two carbon atoms adjacent to the cyclic bond (3) is clear, the acyclic nature of the halogen is clear, but the remaining carbon atoms may or may not be acyclic. The lack of adjacent cyclic bonds does not imply that an atom is acyclic in nature. Such atoms are thus undecided with regard to their cyclic nature.

Ring System Parameters. The ring parameters are

RC	total no. of rings
RN	total no. of ring atoms
RF	total no. of ring fusions
RA	no. of delocalized, aromatic rings
RZ	no. of ring heteroatoms

These parameters describe the total counts for the corresponding features and in no way reflect the individual counts for separate cyclic systems. The sequence fragment shown in Figure 2a, for example, contains three separate cyclic components in which each pair of cyclic atoms is a member of a ring containing at least three atoms. The standard parameter values for the above fragment would be $RC = 3$ and $RN = 9$. Using the notation

$$(RC, RN, RF, RA, RZ)$$

for each component, the three cyclic components are more correctly represented by

$$(1, 3, 0, 0, 0)(1, 3, 0, 0, 0)(1, 3, 0, 0, 0)$$

Similar problems occur in more complex sequence fragments, such as those of Figure 2b,c. In fragment b, the left hand cyclic component is a fused aromatic system and the right hand component may or may not belong to an aromatic or a fused system. The standard parameter values $RC = 3$ and $RA = 1$ do not fully reflect this. In fragment c, the left hand cyclic component contains no heteroatoms, while the right hand component does. Again, this is not reflected in the standard parameters $RC = 2$, $RZ = 2$, where all values are grouped together.

Node Degree Counts. The parameters used to represent nodes of differing degree are incomplete. For nodes of degree greater than 2, the standard parameters are *T* and *Q* for acyclic branches of degree 3 and 4, respectively, *RF* for ring fusions, and *RS* for substitutions on a ring. The nature of the ring fusions is not clear and must be fully represented. The parameters for an HSI description might be limited to, for example, $RC = 2$ and $RF = 2$. This clearly indicates a fused ring system containing two fusion nodes with a cyclic degree

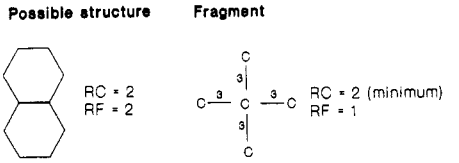


Figure 3. Incorrect numerical match due to incomplete node information.

of 3 each, as shown in Figure 3. An augmented atom containing four cyclic bonds ($RF = 1$) would, however, qualify since the degree of the two HSI fusions is not explicit. Another example is shown in the fragments given in Figures 1i,vi, where the standard parameter values are identical and yet the fragments are very different. A more complete form of representation is required for node degree counts in order that the fragments are described more fully.

The standard parameters clearly show limitations in representing fully the fragment descriptors and the environments in which they occur. An extension of the standard parameter list, which describes numerically the minimum inferred environments in which the fragement may be contained, is now described together with the derivation of this information. These extended parameters are called *extended enumerative parameters.*

3. EXTENDED ENUMERATIVE PARAMETERS

A common representation for both HSI expression and for each fragment descriptor is required for comparison purposes. Due to the differences between the HSI expression and the fragment descriptors, in terms of type and amount of variation, the extended enumerative parameters (EEPs) differ in format for each. They are more desirable for comparison purposes than the original representations but reflect the differences in the degree of variability in the structures they represent.

The format of the EEPs for the fragment descriptors is described here together with the derivation of their values. This is a "once-only" operation on the tree representation of the screen dictionary. The format of the EEPs for HSI expressions and the derivation of their values is described in section 4. Section 5 describes the comparison techniques used to identify the required subset of fragment. The processes described in sections 4 and 5 are carried out each time a generic PS is encountered during fragment generation.

The fragment descriptors EEPs (or FEEPs) are divided into four groups: *unsaturation counts*, *atom counts*, *ring system parameters*, and *node degree counts*. These are described here.

Unsaturation Counts. The unsaturation parameters *E* and *Y* are the two parameters which are not dependent on their environment and whose values can therefore be assigned directly from the fragment descriptor. All other parameters are dependent on further qualifying criteria in order to determine their values.

Atom Counts. Each atom of the fragment must be determined as being cyclic, acyclic, or undecided with regard to cyclic definition. In order to do this, a value, termed the *free valency*, is assigned to the atom. This value is the difference between the valency of the atom and the number of adjacent bonds represented explicitly in the fragment. The maximum possible valency is used for elements such as sulfur. Thus, a carbon atom in a fragment which is connected to two atoms by single, acyclic bonds has a free valency of 2. The free valency of the atom determines whether further cyclic bonds may or may not be present. The counts of the three atom types (*A*, *C*, and *Z*) are divided into three categories: "cyclic", "acyclic", and "undecided".

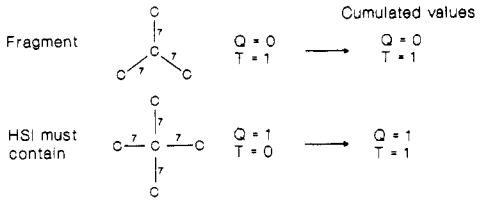


Figure 4. Cumulation of node information.

Ring System Parameters. The ring parameters also have a different format from that of the standard parameter list. Five of the parameters are used: *RC*, *RN*, *RF*, *RA*, and *RZ*. (The cyclic carbon count can be calculated from *RN* and *RZ*.) These five parameters are repeated for each cyclic component of the fragment; furthermore, they represent the minimum inferred counts for each component.

A further parameter, *Xtra*, is used for each component; its use is described later.

Node Degree Counts. The information describing node counts is of four types:

Fusion Degree Counts. All fusions are of degree 3 or more with the arbitrary maximum of 6. All fusions contained within a fragment are therefore counted according to cyclic degree, i.e., the number of adjacent cyclic bonds.

Acyclic Degree Counts. These represent directly the values of T and Q.

Substitution Degree Counts. A cyclic atom may be incident with, and therefore substituted by, more than one acyclic bond; the maximum number of substitutions on one acyclic atom in the screen dictionary is 3. (This is due to the maximum degree of augmented atoms being 4.) For this purpose three node counts are used for single, double, and triple substitution on the focal atom.

Total Degree Counts. The overall degree of the atoms can be determined in the range 3–6. These parameters are useful for fragments defined nonspecifically.

The notation used is

[1 2 3 4; 5 6; 7 8 9 10; 11 12 13]

where elements 1–4 represent fusion degrees 3–6, elements 5 and 6 represent acyclic counts T and Q, elements 7–10 represent total degrees 3–6, and elements 11–13 represent substitutions of type single, double, and triple.

This form of topological representation is not useful for direct comparison purposes. A fragment which should be found in an HSI such as "alkyl" which is further qualified by parameters $T = 0$ and $Q = 1$ is a three-connected carbon which would have the specific derived parameters $Q = 0$ and $T = 1$; see Figure 4. Straight numerical comparison would not allow this fragment to be contained within any structure described by the HSI. An alternative method is to define each count as "the no. of nodes of given degree or less". A result of this definition is the cumulation of counts for each of the four node count types. Values for the above HSI parameters then become $Q = 1$ and $T = 1$, allowing the inclusion of the fragment.

In Figure 5 the nodes shown contribute the following values to the node counts:

Node 1 contributes a 3 degree fusion and a total degree of 3.

Node 2 contributes a 3 degree fusion, a single ring substitution, and a total degree of 4.

Node 3 contributes an acyclic branch of degree 3 and a total degree of 3.

Node 4 contributes an acyclic branch of degree 4 and a total degree of 4.

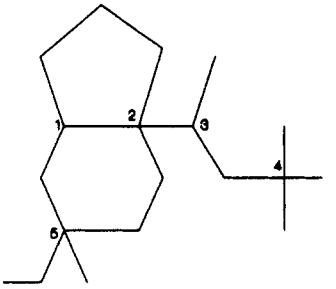


Figure 5. Structure containing several node types.

Node 5 contributes a double ring substitution and a total degree of 4.

For the whole structure the node counts are then

[2 0 0 0; 1 1; 2 3 0 0; 1 1 0]

which, in cumulative form, become

[2 0 0 0; 2 1; 5 3 0 0; 2 1 0]

A further Boolean value, used in later processing, is assigned for augmented atoms. This value, *FusnSubs*, is true where a substitution occurs on a fusion node.

Values for FEEPs are derived by processes described in sections 3.1 and 3.2. The screen dictionary subtree representations are traced by a recursive depth-first algorithm. At each leaf node the derivation is carried out, and that leaf node points to the record containing the parameters. If the fragment represented resides in the screen dictionary (i.e., is a complete fragment), then the leaf node also points to the bit screen vector.

3.1. Derivation of Extended Enumerative Parameters for Augmented Atoms. Values for FEEPs are readily derived for augmented atoms. Each bond is examined to deduce the cyclic or noncyclic (acyclic or undecided) nature of the congener atoms and the focal atom. Noncyclic atoms are assigned as acyclic or undecided by determining their free valency. A free valency of less than 2 indicates an acyclic atom. Values are then assigned for *atom counts*, *unsaturations*, and *FusnSubs*.

Node degree counts are derived by direct assignment. A count is made of the number of cyclic bonds and the number of aromatic bonds. Only one cyclic component may exist in an augmented atom; therefore only one *ring system parameter* record is assigned with parameter values calculated for the minimum inferred counts. Table III shows the directly-assigned values for these parameters according to the number of congeners and their connecting bond types. *RN* and *RZ* are assigned from the atom count values, and *RF* is from the node degree counts.

3.2. Derivation of Extended Enumerative Parameters for Sequence Fragments. Values of the FEEPs for sequence fragments are cumulated during a traverse of, or walk through, the fragment. The cyclic nature of each atom is determined by examination of adjacent bonds, and if it is found to be noncyclic, then, as for augmented atoms, it is assigned as acyclic or undecided by determination of its free valency. Values for *atom counts* and *unsaturations* are cumulated during the traverse.

Node degree counts are cumulated for single substitutions and for fusions of degree 3. Values for these two node types only may be derived from the information given in the fragment. Substitutions are incremented where an atom is adjacent to a cyclic and an acyclic bond, and fusions are incremented where an atom is adjacent to an aromatic and a nonaromatic bond.

Table III. Directly Assigned Values for Augmented Atoms

no. of con-geners	no. of cyclic bonds	no. of aromatic bonds	node counts [13..1]	*RC*	*RA*	FusnSubs
0	0	0	[000;0000;00;0000]	0	0	
1	0	0	[000;0000;00;0000]	0	0	
1	1	0	[000;0000;00;0000]	1	0	
1	1	1	[000;0000;00;0000]	1	1	
2	0	0	[000;0000;00;0000]	0	0	
2	1	0	[001;0001;00;0000]	1	0	
2	1	1	[001;0001;00;0000]	1	1	
2	2	0	[000;0000;00;0000]	1	0	
2	2	1	[000;0001;00;0001]	2	1	
2	2	2	[000;0000;00;0000]	1	1	
3	0	0	[000;0001;01;0000]	0	0	
3	1	0	[010;0010;00;0000]	1	0	
3	1	1	[010;0010;00;0000]	1	1	
3	2	0	[001;0001;00;0000]	1	0	
3	2	1	[001;0010;00;0001]	2	1	T
3	2	2	[001;0001;00;0000]	1	1	
3	3	0	[000;0001;00;0001]	2	0	
3	3	1	[000;0010;00;0010]	2	1	
3	3	2	[000;0001;00;0001]	2	1	
3	3	3	[000;0001;00;0001]	2	2	
4	0	0	[000;0010;10;0000]	0	0	
4	1	0	[100;0100;00;0000]	1	0	
4	1	1	[100;0100;00;0000]	1	1	
4	2	0	[010;0010;00;0000]	1	0	
4	2	1	[010;0100;00;0001]	2	1	T
4	2	2	[010;0010;00;0000]	1	1	
4	3	0	[001;0010;00;0001]	2	0	T
4	3	1	[001;0100;00;0010]	2	1	T
4	3	2	[001;0010;00;0001]	2	1	T
4	3	3	[001;0010;00;0001]	2	2	T
4	4	0	[000;0010;00;0010]	2	0	
4	4	1	[000;0100;00;0100]	3	1	
4	4	2	[000;0010;00;0010]	2	1	
4	4	3	[000;0100;00;0100]	3	2	
4	4	4	[000;0010;00;0010]	2	2	

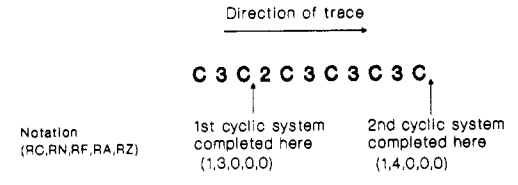


Figure 6. Derivation of cyclic system information from a fragment.

Ring system parameters are not calculated until the completion of a cyclic component; this occurs when a cyclic bond precedes an acyclic bond or when a cyclic bond completes the fragment. The fragment of Figure 6 contains two cyclic components and therefore two ring system information records. Derivation of the minimum inferred parameter counts is performed on each ring system. The figure indicates the points at which each set of cumulated information is written to a new ring system record.

The calculation of ring system parameter values for fragments whose bonds are described at an intermediate level (see Table II) is trivial; the end of a cyclic system indicates a count of at least one ring, and those atoms which are connected by the cyclic bonds make up the minimum ring atom counts (*RN* and *RZ*).

Consider, however, a fragment with specific bond definitions such as

A 14 A 11 A 14 A 11 A 14 A

Derivation of the minimum inferred parameter counts for such a fragment is less straightforward. Simply accumulating counts as each atom and bond is encountered in a path trace of this fragment would give unrepresentative minimum values for the parameters as no account would be given for the types of environment in which the fragment might occur. It therefore

Table IV. Priority Rules and Associated Parameter Values

priority rule	replacement rule	count *RC*	*RF*	*RA*	*RN*	*Xtra*
1	000 → 00	0	0	0	1	0
2	11 → 1	0	0	0	0	0
3	1001 → 11	1	2	0	1	0
4	010 → 00	1	2	1	3	0
5	101 → 11	2	2	1	4	2
6	00 → 0	0	0	0	0	0
7	01 → R	2	1	1	5	1
8	10 → R	2	1	1	5	1
9	0 → R	1	0	0	3	0
10	1 → R	1	0	1	4	0

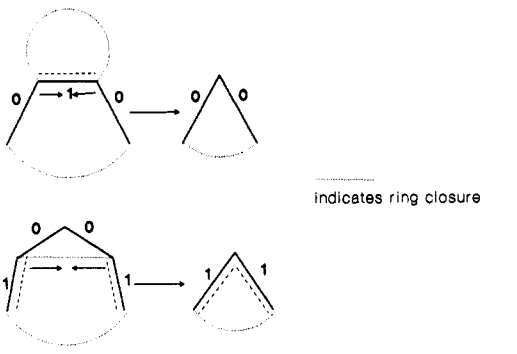


Figure 7. Examples of structure reduction using a string.

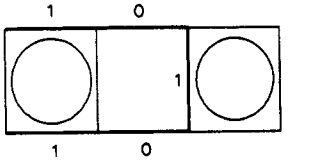


Figure 8. Example ring system.

requires an involved parsing mechanism, similar to those used by *Turing machines*[6] in the treatment of strings.

The derivation of the parameter values for this fragment is performed by first representing the fragment by a string and then reducing the string by means of a series of replacement rules which have been specifically devised for this purpose. The string representation reflects the aromatic or nonaromatic nature of the sequence of bonds contained within the cyclic system, and the reduction of the string representation simulates the removal of atoms and bonds from the cyclic system. Initially, each bond of the fragment contained within the cyclic system is represented by a Boolean value denoting its aromatic nature (1 for aromatic, 0 for nonaromatic). The fragment above then becomes the string 1 0 1 0 1. A series of priority replacement rules is applied to this string, each rule producing a new, reduced string until the final recognition state ("R") occurs. Each time a rule is applied, an associated set of parameter values is accumulated, which represents the minimum inferred parameter values for the structure whose removal has been simulated. The list of replacement rules and the associated minimum inferred parameter values are given in Table IV. The rule 0 1 0 → 0 0, for example, represents the removal of the two atoms associated with a central aromatic ring and its replacement by a single atom; see Figure 7a.

If a minimum aromatic ring size of 4 is used (this figure is tentative and dependent on system definition), then three atoms of the aromatic ring will be removed, together with two fusions.

This rule, rule number 4, is the first rule applied to the string 1 0 1 0 1 since it is the highest priority substring. The string then becomes 1 0 0 1. The next rule, rule 3 (1 0 0 1 →

Table V. Alterations to Ring Parameters for Given XTRA Values

Xtra	*RZ* alteration	*RF* alteration	*RN* alteration
0	none	none	none
1	none	increased by 1	none
	increased by 1	none	increased by 1
2	none	increased by 2	none
	increased by 1	increased by 1	decreased by 1
	increased by 2	none	increased by 1

1 1), removes the central nonaromatic ring and replaces it with a single atom; see Figure 7b.

Parameter values are again cumulated. The last two rules applied are rule 2 (1 1 → 1) and then rule 10 (1 → R), the "recognition state". The total cumulation of parameter values is *RC* = 3, *RF* = 4, *RA* = 2, and *RN* = 8. These values refer to the structure given in Figure 8.

Examples of the derivation of parameter values from similar Boolean strings are given below (rule numbers are shown in parantheses).

	RC	*RF*	*RA*	*RN*	*Xtra*
0.01101 → (2) 0101	0	0	0	0	0
→ (4) 001	1	2	1	3	0
→ (6) 01	1	2	1	3	0
→ (7) R	3	3	2	8	1
2.10001 → (1) 1001	0	0	0	1	0
→ (3) 11	1	2	0	2	0
→ (2) 1	1	2	0	2	0
→ (10) R	2	2	1	6	0
3.01001 → (3) 001	1	2	0	1	0
→ (2) 01	1	2	0	1	0
→ (7) R	3	3	1	6	1

The *Xtra* parameter is a special parameter which enables the program to select from a choice of total values as follows:

The sequence fragment descriptors containing specifically-defined bonds are all "bond sequences"; i.e., the atom definition is always nonspecific and therefore gives no indication of the maximum degree of each atom. A carbon atom has a maximum degree of 4, and a heteroatom has a maximum degree of 6, the largest connectivity dealt with here; a nonspecified atom must therefore be given a choice of maximum degree, either 4 or 6.

This choice is necessary since, when qualifying a fragment in an HSI with no heteroatoms, such as "cycloalkyl", the maximum degree of 4 must be used for all atoms; when qualifying a fragment in an HSI with one heteroatom, the maximum degree of 6 may be used for one atom (4 for the rest); with two heteroatoms, two atoms of the fragment may be degree 6, and so on.

The Boolean string 1 0 describes a fusion node between two rings, one aroamtic, the other nonaromatic. The single fusion node may be a spirofusion if the fusion atom is a heteroatom, in which case the *RZ* count would be incremented by 1, or it may be one of a pair of orthofusions, in which case the *RF* value must be counted as 2. The *Xtra* parameter gives a choice when processing the HSI.

The *Xtra* value alters the values of the parameters *RZ*, *RF*, and *RN* dependent on its own value. There are two choices for a value of 1 and three for a value of 2; Table V gives these choices.

Examples 1 and 3 above may therefore give values (notation (*RC*, *RF*, *RA*, *RN*, *RZ*)) of

(3, 3, 2, 9, 1) or (3, 4, 2, 8, 0)

and

(3, 3, 1, 7, 1) or (3, 4, 1, 6, 0)

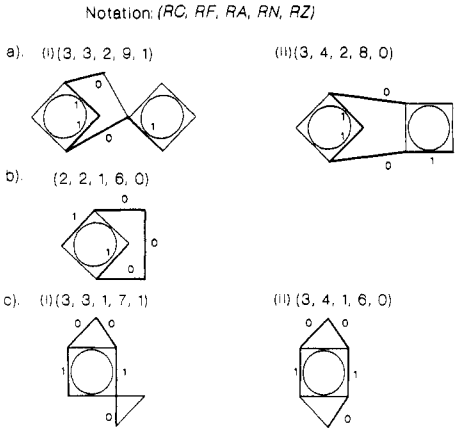


Figure 9. Structures generated for derivation examples 1, 2, and 3.

respectively. The minimum inferred structures represented by these values are shown in Figure 9a(i), 9a(ii), 9c(i), and 9c(ii), respectively. The minimum inferred structure represented by example 2 is shown in Figure 9b.

Structures such as those of Figures 9a(i),(ii) exhibit unusual chemical features; the intensional expression might not determine such extensions depending on how the intension is defined. The choice of 4 for a minimum aromatic ring size is also dependent on the definition of the intension. A more appropriate approach would be to use a minimum aromatic ring size of 6 and to ignore the possibility of a spirofusion in an aromatic ring system. Such limitations produce a narrower extension to the inferred structure types which reflects the ideas discussed in Dethlefsen et al.[2,7] A result of this treatment is the production of two alternative sets of replacement rules, the *broadest* and *narrowest conceivable extension* rules (BCE and NCE). Dethlefsen has described the concept of user-defined levels of search whereby the desired extension can be chosen by the searcher. The two sets of rules can be used in conjunction with a similar choice (BCE and NCE) of standard default parameter values used to represent the HSI.

The replacement rules and their associated parameter values are tentative and may be altered in a future system.

4. EXTENSION OF THE NONTOPOLOGICAL REPRESENTATION OF HOMOLOGOUS SERIES IDENTIFIERS

Due to the nature of HSIs, the fact that they are represented in the ECTR by a list of parameters and not real atoms, and that they describe one of a number of possible structures, the EEPs used to represent HSIs (HEEPs) have a different format to FEEPs. FEEPs are divided into three groups: *standard parameter ranges*, *further parameter values*, and *node degree count lists*. These values, together with their derivation from the original parameter representation of the HSI, are described here.

Standard Parameter Ranges. The parameters of the homologous series identifier have associated ranges of values which describe each of the structural features they represent. The processing of these values requires only their maximum and minimum values; these are read into two parameter lists: *HighPars* and *LowPars*, respectively. In the case of an unbounded parameter range, where the maximum value is infinity, the HighPars value is set to *NotSet*, a constant with the value −1.

Further Parameter Values. Four more values are calculated (where *max* and *min* are the HighPars and LowPars values of the respective parameter):

AcyAts, the maximum number of acyclic atoms which may exist in any structure covered by the HSI, $A_{max} - RN_{min}$

Figure 10. Basic graphs for node combinations 0, 2, 0, 0 and 1, 0, 1, 0.

AcyHets, the maximum number of acyclic heteroatoms which may exist in any structure covered by the HSI, $Z_{max} - RZ_{min}$

KMax, the maximum total degree possible from the *RC* and *RF* values and given by the formula KMax = $2(RC_{max} + RF_{max} - 1)$

AcyclicLink, the maximum number of acyclic components which join two or more cyclic components in the structure, (i.e., the maximum number of cyclic components less one), given by the formula AcyclicLink = $RC_{max} - 1 - ((RF_{min} + 1)$ div 2).

These first two groups, standard and further parameter values, are used in qualifying calculations, described later, which process the ring system parameters and the unsaturation and atom counts of the fragments.

Node Degree Count Lists. The node degree counts, which describe the numbers of different types of node of degree 3 or more, have the same format in HEEPs as they do in FEEPs. The fact that the HSI describes one of many possible structures means, however, that there may be many possible combinations of nodes of differing degrees. This is reflected by using a list of alternative node degree records rather than a single record. Alternative combinations of node degrees are enumerated using all possible values in the HSI ranges for the parameters which affect ring fusions (*RC* and *RF*), ring substitutions (*RS*), and acyclic branching (*T* and *Q*).

Fusion nodes are enumerated for all possible combinations of *RC* and *RF* as follows:

The sum of the degrees in a homeomorphically reduced graph is shared between the *RF* vertices in that graph, each having a minimum degree of 3 (a homeomorphically reduced graph (or basic graph) is a reduced graph in which nodes of degree 2 of the original graph have been systematically removed). Consider a cyclic system containing four rings and two fusions. The total connectivity, *K*, of the fusion nodes is given by the formula $K = 2(RC + RF - 1)$

$$K = 2(4 + 2 - 1)$$

$$K = 10$$

The sum of degree values must be shared between the two fusion nodes, the two possibilities being either two nodes of degree 5 or one node of degree 6 and one of degree 4 (note that the maximum degree for any node is 6, the maximum number of congeners in the ECTR). Using the notation

$$[M_6, M_5, M_4, M_3]$$

where M_x = no. of nodes of degree x, then the two combinations can be denoted as

[0, 2, 0, 0] or [1, 0, 1, 0]

These graphs are shown in Figure 10.

Enumeration of such combinations is carried out in the following way: First set all nodes to degree 3; the initial sum of degrees, denoted by K_i, is then $3RF$. ($RC = 4$, $RF = 2$ is used as an example).

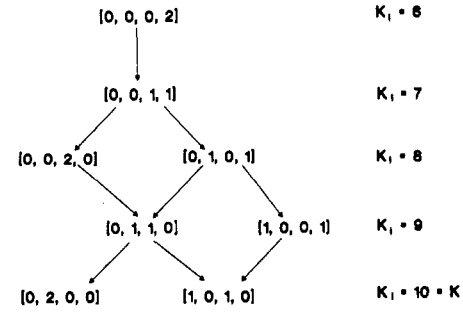


Figure 11. Generation of node combinations.

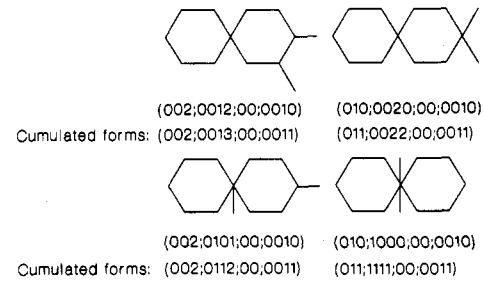


Figure 12. Substitution types on a fused ring system.

$$[0, 0, 0, 2] \quad K_i = 6$$

Then, systematically increment each node by degree 1, increasing K_i appropriately until $K_i = K$. An example of this enumeration is given in Figure 11.

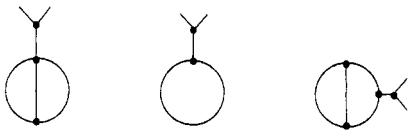
The resulting combinations are written to the respective *node degree* count elements.

Values of the acyclic node counts are directly assigned from the maximum *T* and *Q* parameter values.

Values of possible substition types are also enumerated from the maximum *RS* parameter value. An *RS* value of 2 results in two possible combinations, either two single substitutions or one double substitution. The variety of possible combinations of substitution types are generated and each is held in a three-element array.

The total degree node information must be calculated firstly by linking the substitution and fusion information and then by adding the acyclic information. The substitution types generated may occur either on a nonfusion node or on fusion nodes generated using the method described above, any permutation of which will result in nodes of differing total degree. Consider, for example, two substitutions and one spirofusion; the substitutions may be geminal on one nonfusion node, or geminal on the spirofusion (valency dependent on the *RZ* value), may be single substitutions on nonfusion nodes, or one on the spirofusion and one on a nonfusion node. Examples of real structure possibilities are shown in Figure 12, each of which results in the different total degree combinations shown, and must therefore be enumerated. Enumeration is carried out for each substitution combination and each fusion combination in all possible positions of substitution. The acyclic node values are then added to each permutation to give the total degree information. Each of the four types of node degree counts is cumulated in the same way as those describing fragments (see section 3), as shown in Figure 12.

The possible number of permutations resulting from this enumeration of node values may be very large and need not be fully represented. Instead, a record is physically represented only if it has a unique maximum value; otherwise it will be covered by all previously generated records. For example, a combination representation (0 0 1;0 0 1 3;0 1;0 0 0 2), which



(001;0013;01;0002) (001;0002;01;0000) (001;0004;01;0002)

Figure 13. Homeomorphically reduced graphs of node combinations.

may have been generated previously, covers the newly-generated combination (0 0 1; 0 0 0 2; 0 1; 0 0 0 0) as no new node value exceeds those of the previously generated representation. However, the new combination (0 0 1; 0 0 0 4; 0 1; 0 0 0 2) must produce a new physical representation as it has a unique maximum node value. Homeomorphically reduced graphs of these three node combinations are shown in Figure 13. A record is held for each unique representation.

5. FRAGMENT QUALIFICATION

We now have two comparable representations (extended enumerative parameters) for the HSI and for the fragment descriptors.

For the HSI, we have the HEEPs

standard parameter ranges	HighPars and LowPars
further parameter values	Acyts, AcyHets, KMax, and Acyclic Link
node degree count list	a list of alternative node degree counts

For the fragment descriptors, we have the FEEPs

unsaturation counts	i.e., the *E* and *Y* parameters
atom counts	*A*, *C*, and *Z* in cyclic, acyclic, or undecided categories
ring system parameters	*RN*, *RC*, *RF*, *RA*, and *RZ* for each cyclic component
node degree counts	one single record of node degree counts

Identification of those fragments which may exist within the scope of an HSI description, or elimination of those which may not, is a numerical process using the representations shown above. The values of EEPs for each fragment are compared with values of EEPs for the HSI using the processes described here. The reuslt of this processing identifies whether or not a fragment might exist within the environment described by the HSI representation.

Firstly, the fragments are generated by a depth first trace of each of the screen dictionary trees. At each leaf node the extended enumerative parameters for each fragment (FEEPs) are compared against those generated for the HSI (HEEPs). Comparison is carried out as follows.

Unsaturations are tested by a straightforward comparison of the two fragment parameters against the respective HighPars values of the HSI.

Node degree counts for the fragment are tested against each of the possible node degree records of the HSI. If all counts for the fragment do not exceed those of the HSI record in at least one of the HSI records, then the fragment is tested further. If the counts for the fragment exceed those of the HSI in all cases, then the fragment is eliminated.

Acyclic *atom counts* are tested in order to deduce whether the noncyclic counts for atoms and heteroatoms of the fragment lie within the counts AcyAts and AcyHets for the HSI. If this is not so, then as many undecided atoms or heteroatoms as are necessary may be deemed cyclic in order that the noncyclic counts equal those of the HSI. If these atoms are deemed cyclic, then each extra cyclic atom will add the following to the parameters of the fragment: (1) an extra cyclic component, (2) an extra ring, (3) at least three extra cyclic atoms, and (4) any extra heteroatoms so deemed adds one cyclic heteroatom.

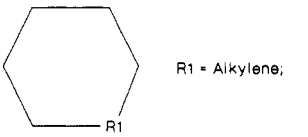


Figure 14. Doubly-connected generic group.

These extra counts are held temporarily in the variables *MinExtraRings* for the extra rings to be added, *MinExtraAts* for the extra cyclic atoms, and *MinExtraHets* for the extra cyclic heteroatoms. Consider the HSI parameters AcyAts = 2, AcyHets = 0 when qualifying the fragment

C 2 C 3 C 2 N 2 C

the nitrogen can be deemed cyclic since no extra heteroatoms are permissible, giving extra values of MinExtraRings = 1, MinExtraAts = 3, and MinExtraHets = 1. If the values of the acyclic counts for atoms or heteroatoms exceeds those of AcyAts or AcyHets, respectively, then the fragment is eliminated; otherwise the cyclic information is now processed.

Ring system parameters are tested for the inclusion of any sum of the minimum inferred parameters for all cyclic components in the fragment within the parameters of the HSI. The parameters are totalized for all components of the fragment, including those derived from processing the non-cyclic information (i.e., MinExtraRings, MinExtraAts, and MinExtraHets) and tested against the HighPars parameters of the HSI. In many cases, when using broadest conceivable extension rules for derivation, there may be more than one choice of minimum inferred parameters for each cyclic component. The possible permutations of choices of totalized parameters are then enumerated until one permutation gives total parameter values which do not exceed those of the HSI, in which case the fragment is valid. If no such permutation is found, then the fragment is eliminated.

Once identified, the fragment is written to a local fragment list, the operations on which have been described in the preceding paper in this series.[1]

6. DOUBLY-CONNECTED GENERIC PARTIAL STRUCTURES

In many cases a partial structure (PS) may be doubly connected to its parent PS. Such occurrences are common in patents, and the connection is represented in the ECTR by inter-PS connection records which describe pairs of connected atoms and bonds.

The double connection may occur in a chain or it may form part of an incomplete ring as in the "alkylene" of Figure 14. This latter case causes problems for both specific and generic PSs since perception of the cyclic bonds in the PS is required. The problem has been resolved for specific PSs using a ring perception algorithm developed by Carruthers.[8]

The parameters used to represent generic PSs do not take account of the environment in which the PS occurs and therefore do not change if the PS forms part of a ring or a chain. The "alkylene" group would be represented by the same parameter values if it were contained within a chain as those representing the same group in Figure 14.

The generation of fragments from such parameter values would therefore be incorrect since many cyclic fragments would be missing. The problem has been solved by systematic replacement of acyclic bonds by cyclic bonds of the same type. This is performed using pairs of atoms in the original

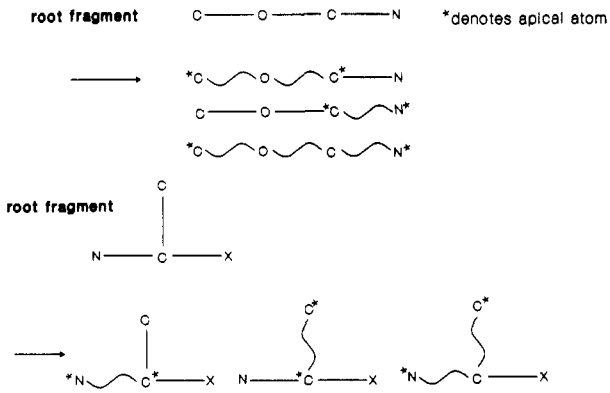


Figure 15. Bond replacement in doubly-connected fragments.

(or root) fragment as the possible externally-connected (or 'apically-connected') atoms to the parent and replacing all the acyclic bonds between them. The path between these two atoms then becomes the "half-ring" path which completes the ring in the parent. Since the choice of atoms for apical connection depends on their valency, they are tested accordingly. Figure 15 shows examples of the replacement of bonds in both augmented atoms and sequence fragments.

The root fragments generated from the HSI, which may originally have contained cyclic bonds, may produce many possible half-ring fragments by this replacement. One problem with this process is that the half-ring fragments so produced do not relate to the fragment dictionary since it is the generation of the root fragment which is based on the dictionary tree and not the resulting half-ring fragment. A result of this is that there may be some half-ring fragments which should be generated but which do not have an associated root fragment in the dictionary tree.

Consider, for example, the root sequence fragment in Figure 15. If this fragment is identified in a generic PS which forms a half-ring (due to double connection to its parent within a ring), then the three new fragments shown result by the process of replacement, any of which may be found in the screen dictionary. If, however, the original root fragment does not exist in the screen dictionary, then it would not be identified in the generic PS and the three new fragments could not be generated. But it is necessary to generate these three fragments if they appear in the screen dictionary.

This problem is resolved by a reverse replacement operation during the generation of the dictionary tree. This is carried out by systematic replacement of cyclic bonds by acyclic bonds when the tree is generated. This extends the dictionary tree to include all root fragments which may be later replaced to produce half-ring fragments. The three half-ring, sequence fragments of Figure 15 therefore produce their root fragment representation in the screen dictionary. All fragments in the tree which are generated in this way are flagged to indicate that they need be examined for HSI generation only if that HSI describes a doubly-connected generic PS.

7. CONCLUSION

The result of the treatment described above on a generic partial structure is a list of incomplete fragments, which may be continued in other partial structures to produce inter-PS fragments, and a bit screen which represents the intra-PS fragments contained wholly within one of the structure choices denoted by the HSI. A standard representation is used for fragments and bit screens eminating from specific partial structures, as described in the preceding paper in this series.[1] The linkage of incomplete fragments and the "bubble-up" of bit screens is also described.

Examination of the default parameter list for the term "radical" reveals infinite structural variation in the extension determined by the parameter values. Clearly, since the extension is infinite in variation, the number and variation of fragments which can be contained within the extension are also infinite. The result is qualification of the complete screen set whenever this term is used *with its default parameters*. In many cases, "radical" is further qualified by parameter values, in which case fragments may be eliminated by processes described above. The POSS bit screen (described in the preceding paper in this series[1]) only will appear black since all fragments generated from HSIs are assigned POSS.

ACKNOWLEDGMENT

We gratefully acknowledge funding from International Documentation in Chemistry mbH, Derwent Publications Ltd., and Questel SA, in support of the research described here, and thank Dr. W. Dethlefsen of BASF, Professor P. Willett, and Dr. J. M. Barnard of Barnard Chemical Information Ltd. for generous advice. We also thank Chemical Abstracts Service for provision of documentation on the screen sets.

REFERENCES AND NOTES

(1) Holliday, J. D.; Downs, G. M.; Gillet, V. J.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 14. Fragment Generation from Generic Structures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 453–462.

(2) Dethlefsen, W.; Lynch, M. F.; Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 11. Theoretical Aspects of the Use of Structure Languages in a Retrieval System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 233–253.

(3) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 10. The Generation and Logical Bubble-Up of Ring-Screens for Structurally Explicit Generics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 215–224.

(4) 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. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 160–164.

(5) Stiegler, G.; Maier, B.; Lenz, H. Automatic Translation of GENSAL representations of Markush structures into GREMAS fragment codes at IDC. Paper presented at fall ACS meeting, Washington D.C., 1990.

(6) Turing, A. M. On Computable Numbers with an Application to the Entscheidungs-Problem. *Proceedings of the London Mathematical Society*; London Mathematical Society: London, 1936; Vol. 2, pp 230–265.

(7) Dethlefsen, W.; Lynch, M. F.; Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 12. Principles of Search Operations Involving Parameter Lists: Matching-Relations, User-Defined Match Levels, and Transition from the Reduced Graph Search to the Refined Search. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 253–260.

(8) Carruthers, L. Automatic Ring Perception in Generic Chemical Structures; Master's thesis, University of Sheffield, 1983.