

The Substance Module: The Representation, Storage, and Searching of Complex Structures[†]

ALAN J. GUSHURST, JAMES G. NOURSE,* W. DOUGLAS HOUNSHELL, BURTON A. LELAND,
and DAVID G. RAICH

Molecular Design Limited, 2132 Farallon Drive, San Leandro, California 94577

Received June 1, 1991

Chemical structures are typically represented in computer programs as simple graphs, where atoms are represented by a list of nodes and the bonds by a list of nondirectional edges. While this convention allows for the representation of a large variety of chemical structures, it does not lend itself toward the representation of many common substances such as polymers, non-stoichiometric mixtures, and formulations. An extension to this convention has been developed which allows properties to be identified with a defined subgraph in a structure, an *Sgroup*. This extension has been implemented in the Substance Module, a new MACCS-II module, and is used to represent and search a much broader class of chemical substances. A description of the new representation and searching capabilities is given as well as examples of its use.

I. INTRODUCTION

The use of computers to manage chemical information is steadily increasing and has already gained wide acceptance in the pharmaceutical and chemical industry. Chemical information management systems, such as MACCS-II, OSAC, DARC, and CAS, have proven to be valuable tools for storing and searching proprietary and publicly accessible online data.^{1,2} These systems can provide answers to a wide variety of questions simply and effectively with a minimum of time and effort. Typical questions include: "Has this molecule been synthesized? If so, what are its properties? If not, have any similar compounds been synthesized? What compounds have been made that share a common substructure?" While these basic questions can be answered by a traditional library search, using a software system to house and search chemical data is far more efficient.

Despite recent advances in chemical structure searching, including extensions to representing and searching three-dimensional structures,³⁻⁵ most chemical information management software systems have limited capabilities for storing and searching complex substances with complex data relationships; namely, polymers, biopolymers, mixtures, and formulations. These substances are problematic in that they require an association between data and arbitrary parts of the structure. For example, in order to represent a polymer, it is necessary to be able to identify the repeat unit and tie various pieces of data to it, e.g., polymer type, connectivity (head-to-tail or head-to-head), average molecular weight, repeat distance, etc. Currently, many chemical information systems have limited capabilities for associating data at the atom, bond, and molecule level and completely lack the ability to associate data with an arbitrary substructure.

In this paper, we will discuss how the MACCS-II program has been extended to store and search complex substances such as polymers, biopolymers, mixtures, and formulations.⁶ These extensions are part of a new MACCS-II module called the Substance Module. Technically, these extensions have been accomplished by introducing a new construct, the *Sgroup*.

II. SGROUP REPRESENTATION

The notion of an "Sgroup" has been introduced as an extension to chemical structure representation in the Substance Module. The "S" stands for substructure due to the Sgroup's general applicability to partial or complete structures. We define an Sgroup as a *persistent collection of atoms and bonds*, i.e., as an identified substructure that is maintained as an

integral part of the connection table. Because Sgroups are maintained in connection tables, they can be stored in transfer files and databases and can be searched. We have delineated two general types of Sgroups, "chemical Sgroups" and "data Sgroups". The overall organization of Sgroups is presented in Figure 1 as a guide for the following discussion.

A. Chemical Sgroups. Since there are certain chemical uses of these Sgroups which are general, we have implemented three broad classes of chemical Sgroups: (1) polymers, (2) components, mixtures and formulations, and (3) drawing and display shortcuts. All chemical Sgroups are subject to two restrictions. First, they must be composed of intact fragments or substructures delimited by a set of crossing bonds. A crossing bond has one atom contained in the Sgroup and one outside the Sgroup and is not itself considered to be in the Sgroup. Second, they must be hierarchical. That is, if any two chemical Sgroups include any atoms or bonds in common, then one definition must be equal to or a subset of the other.

1. Polymers. Polymers are large molecules built up by the repetition of small, simple chemical units.⁷ The compounds that react to form these chemical units are referred to as monomers. For some polymers, the repetition is linear, while for others it is branched or interconnected to form networks. The manner in which the chemical units connect to each other is described as *polymer configuration*.⁸ Two common types of polymer configuration are head-to-head and head-to-tail. Polymers that are built up from a mixture of two or more monomers are referred to as copolymers: a polymer that contains two or more kinds of monomeric units in the same molecule. Copolymers are further classified based on how the different monomeric units connect to each other. The most common types of copolymers are alternating, block, graft, and random. Polymers can also be classified based on how they are modified, e.g., when two or more separate polymers are linked together somewhere in the middle of their chains, by a bridging agent or by reactive sites on the side chains, the resulting material is referred to as a cross-linked polymer. The special units that start and terminate polymers are called end groups.

Polymers are typically represented via either a *structure-based* or a *source-based* system. A structure-based representation assumes that the structural identity of the chemical units, as well as their sequential arrangements within the polymer, is known. Examples are shown in Figure 2. Often, this connectivity information is not known. In these cases, a source-based representation is used where the polymer is depicted in terms of the monomer(s) that formed it. Consequently, if the connectivity were not known for the block copolymer shown in Figure 2, it could instead be represented

[†] Presented at the Division of Chemical Information, 201st National Meeting of the American Chemical Society, April 16, 1991, in Atlanta.

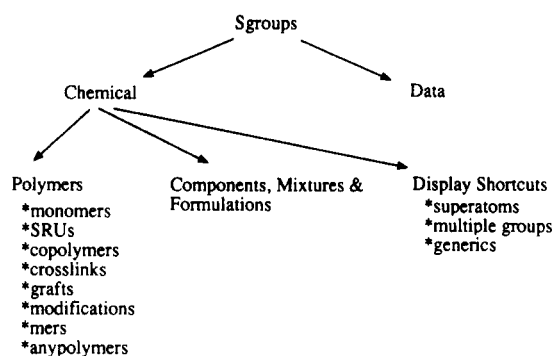


Figure 1. Sgroup organization in the Substance Module.

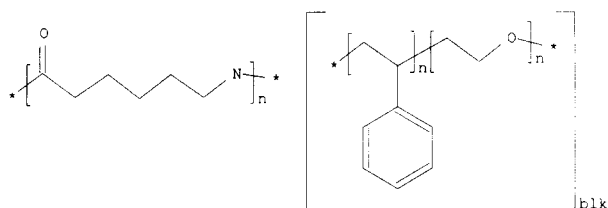


Figure 2. Structure-based representations for nylon-6 and a block copolymer of polystyrene and ethylene oxide.

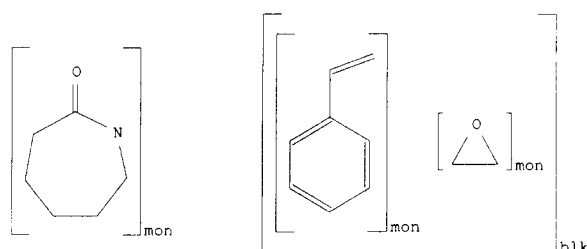


Figure 3. Source-based representations for nylon-6 and a block copolymer of polystyrene and ethylene oxide.

by the styrene and ethylene oxide monomers shown in Figure 3.

The Substance Module allows for either structure-based or source-based representation. For structure-based representation, the workhorse is the *SRU* (structural repeating unit) polymer Sgroup type. The *monomer* polymer Sgroup type, on the other hand, is the workhorse for source-based representation. It is used to allow the registration of monomers for which the structure of the resulting polymer is unknown or unimportant. All polymer Sgroups are graphically delimited by brackets and are distinguished from one another by the labels on the brackets. The labels for SRU and monomer polymer Sgroup types are "n" and "mon", respectively (Figures 2 and 3).

End groups can be specified explicitly but are not required. If the user does not know their composition, or does not choose to show them for some reason, he or she can use **atoms* ("star atoms") to represent them (see Figure 2). The **atom* is a new atom type that acts as a place holder and has many of the same attributes as a generic "any" atom. And, it is registrable.

Additional polymer Sgroup types are supported including the following: *copolymer*(unknown), *alternating*(copolymer), *block*(copolymer), *random*(copolymer), *crosslink*, *graft*, *modification*, and *mer* (see Figures 4 and 5 for examples). There is also a special Sgroup type, *anypolymer*, that is used for posing more general polymer search queries. The crosslink, graft, and modification polymer Sgroup types are similar in that they all describe a structural modification to the backbone or side chain of a polymer. For instance, one could synthesize poly(vinyl acetate) and then modify it by hydrolyzing away some of the acyl groups. This modified homopolymer could then be represented in the Substance Module by using an SRU

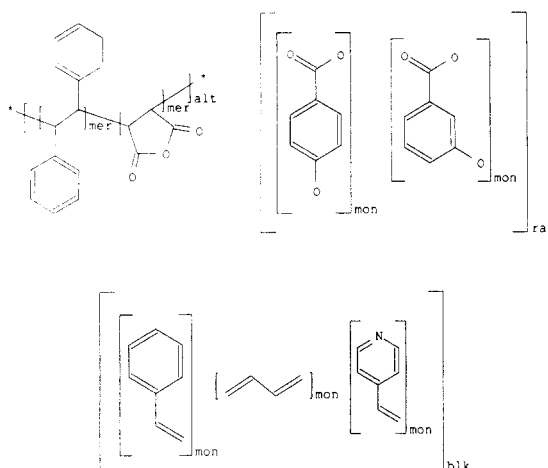


Figure 4. Examples of alternating (alt), block (blk), and random (ran) copolymers.

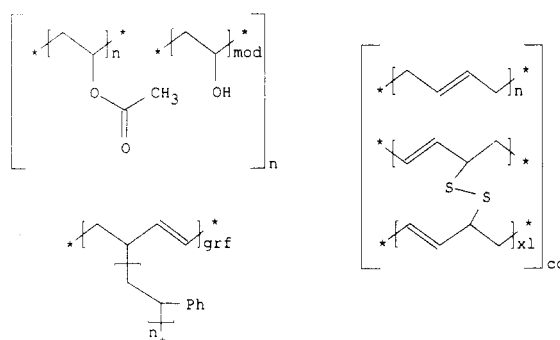


Figure 5. Examples of modification (mod), graft (grf), and crosslink (xl) polymers.

and a modification polymer Sgroup type to identify the original and modified backbone sections, respectively (Figure 5). Note, data Sgroups could be used to indicate to what extent the original backbone is modified (vide infra). The crosslink polymer Sgroup type is reserved for situations where a bridge has been made between two polymers, whereas the graft polymer Sgroup type is used to show where a polymer has been attached to the middle of another polymer (Figure 5). The *mer* polymer Sgroup type is used to describe a unit that does not hook up to another copy of itself, i.e., the repeat count for a *mer* is one. The main usage of the *mer* polymer Sgroup type is in the representation of alternating copolymers, e.g., see Figure 4. Finally, the *anypolymer* polymer Sgroup type is solely reserved for searching. In a search, it will hit all polymer Sgroup types.

Polymer configuration can be assigned to all polymer Sgroup types except for *mer*. The supported values include head-to-tail, head-to-head, and either/unknown. The either/unknown designation is reserved for identifying cases where a mixture of head-to-tail and head-to-head configurations are present or for cases where the polymer configuration is not known. The labels "ht" and "hh" are used to identify polymer Sgroups that have head-to-tail and head-to-head configurations, respectively. The absence of a label indicates that the polymer configuration is either/unknown (Figure 6).

The number of crossing bonds permitted for each polymer Sgroup type largely influences the number of polymers one can represent. A vast majority of polymers can be represented with 2–4 crossing bonds. The number of allowed crossing bonds for each polymer Sgroup type in the Substance Module is given in Table I.

2. Components, Mixtures, and Formulations. Mixtures and formulations are substances that are made up of two or more individual components. A mixture is distinguished from a formulation in that the order of mixing is not important to

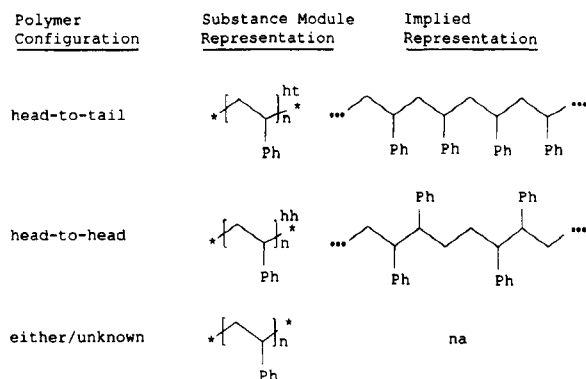


Figure 6. Examples of the Substance Module representation and implied representation for each value of polymer configuration.

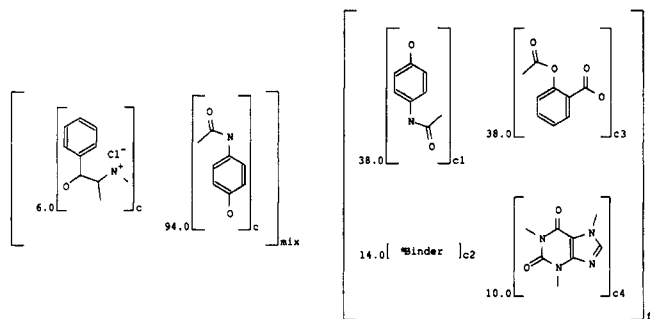


Figure 7. Examples of a mixture (mix) and formulation (f). The percent composition of each component (c) in the overall mixture or formulation is designated via a data Sgroup, e.g., 38.0 on the first component in the formulation.

Table I. Allowed Number of Crossing Bonds for Each Polymer Sgroup Type

polymer Sgroup type	allowed no. of crossing bonds in query structure	allowed no. of crossing bonds in database structure ^a
SRU	0-4	0, 2-4
monomer	0	0
mer	0-8	0-8
copolymer (& all subtypes)	0-4	0, 2-4
graft	0-4	0, 2-4
modification	0-4	0, 2-4
crosslink	0-8	0-8
anypolymer	0-8	na

^a Zero crossing bonds are for source-based representation, while higher values are for structure-based representation.

the overall behavior of the mixture. A formulation, however, is like a recipe in which the order of mixing is important in determining the properties of the final substance. In Figure 7, the decongestant and over-the-counter painkiller are examples of a mixture and formulation, respectively.

In the Substance Module, *mixture* and *formulation* Sgroups are defined by a set of unordered and ordered *component* Sgroups, respectively (Figure 7). Each component must have 0 crossing bonds, and all pieces of the structure within a mixture or formulation Sgroup must belong to a component Sgroup. Nested (hierarchical) mixtures and formulations are allowed, but there must be intervening component Sgroups, i.e., a mixture can be part of another mixture so long as it is specified as a component of the top-level mixture. Mixture and formulation Sgroups are also restricted to 0 crossing bonds.

3. *Drawing and Display Shortcuts.* The drawing and display shortcuts are a special class of Sgroups designed to make structure display more manageable and aesthetically pleasing. Biopolymers, for example, are difficult to visualize at the atomic level. Consequently, they are often presented in terms of well-defined building blocks such as amino acids and nucleotides (Figure 8). Structures with long repeating

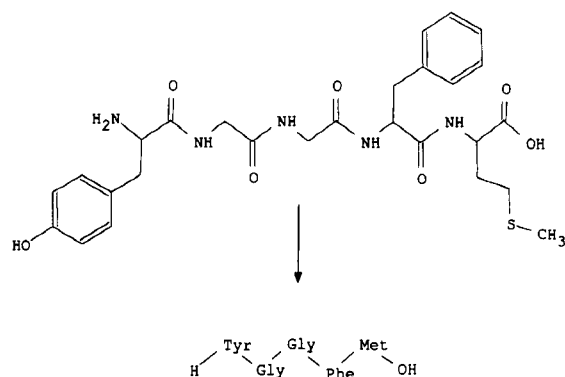


Figure 8. Two different but equivalent representations of the pentapeptide methionine enkephalin.

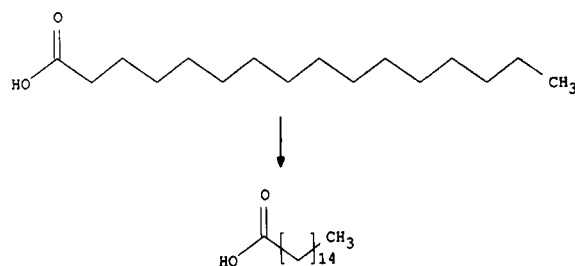


Figure 9. Example showing how palmitic acid can be represented with a multiple group.

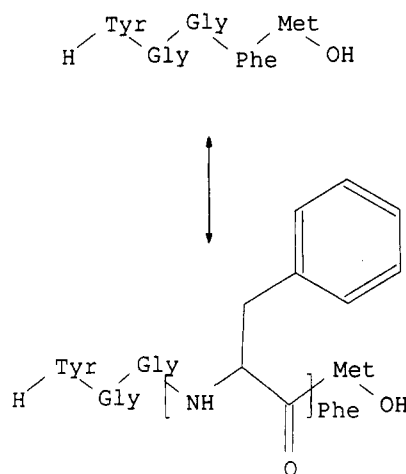


Figure 10. Example showing how a superatom (Phe) can be expanded and contracted.

sequences can also be made more manageable by bracketing the repeating unit and associating a repeat count (Figure 9).

In the Substance Module, we introduced three types of drawing and display shortcuts: *superatoms*, *multiple groups*, and *generics*. Superatoms are chemical Sgroups with 0-8 crossing bonds that can be "contracted" and "expanded". Contracting a superatom effectively removes the superatom definition (set of atoms and bonds) from the structure and replaces it with a user-specified label. Expanding a superatom is simply the reverse process (see Figure 10). Superatoms may be used to simplify the display of peptides, structures with standard protecting groups, polymer formulations with common additives, etc.

In some situations, a structure may have more than one occurrence of a given superatom definition. A replicate function was implemented as a drawing aid for these situations. The user can invoke the replicate function to find all groups identical to an existing superatom and have the superatom definitions automatically generated for those groups.

Multiple groups are chemical Sgroups which are to be repeated a specified integral number of times. They are defined

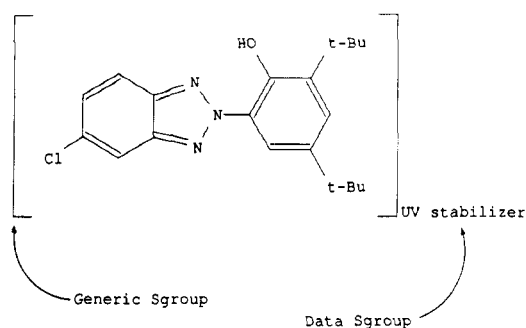


Figure 11. Example showing how a generic Sgroup and a data Sgroup can be used to create a "customized chemical Sgroup".

in their contracted state but may be expanded similarly to superatoms. They may have 0 or 2 crossing bonds and can be used both as drawing shortcuts and for display simplification. As a drawing shortcut, the multiple group definition can be input, expanded, and then deleted (by removing only the brackets) as a quick means of entering a "backbone" for later modification. As a display simplification, structures such as palmitic acid can be represented by using a multiple group (Figure 9).

The generic Sgroup is another type of chemical Sgroup that is provided to add clarity to the structural display. Generic Sgroups are simply displayed as a group of brackets, without any identifying label, encompassing a chemical substance. When used in conjunction with data Sgroups, generic Sgroups give the user a way of defining "customizable chemical Sgroups". For instance, UV stabilizers are common additives to polymers, yet they are not offered in the MACCS-II system as one of the standard chemical Sgroups. However, the user can still attain the desired representation and display by bracketing the chemical substance with a generic Sgroup and then applying a data Sgroup to the bracket (Figure 11). Generic Sgroups may have 0–8 crossing bonds.

B. Data Sgroups. There are many situations where one may want to tie data to a part of the structure in order to more fully describe the structure as a whole. For instance, one may have a terpolymer composed of acrylonitrile, butadiene, and styrene. Yet to represent this terpolymer simply in terms of these building blocks is insufficient. A complete representation must also take into account the percent composition and distribution of these building blocks in the terpolymer in order to fully describe the substance. Note, there are many different materials that are composed of acrylonitrile, butadiene, and styrene, many of which have different properties.

Data Sgroups can be used to add such structural detail by allowing the user to tie data to an arbitrary substructure. Data is associated with this substructural piece only and not with the entire structure. While data Sgroups can also be defined in terms of crossing bonds, it is convenient to think of them as arbitrary substructures carrying their own data. They may be entered such that their definitions are hierarchical, but there is no requirement that they be hierarchical.

A data Sgroup (set of atoms and bonds) can either be linked directly to a chemical Sgroup, such as in the UV stabilizer example in Figure 11, or to an arbitrary set of atoms and bonds (except that once a bond is included, the two atoms connected by the bond are automatically included). If the data Sgroup is tied to a chemical Sgroup, the definition of the data Sgroup will be modified when the definition of the chemical Sgroup is modified, otherwise the data Sgroup definition will only be modified when explicitly redefined. Each data Sgroup has associated with it a data field, display characteristics, and data values. The fields and default display states allowed are defined by a database administrator as a part of the data dictionary. The current implementation allows for the same data

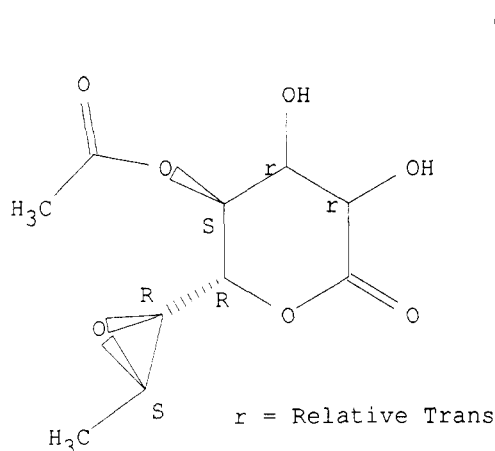


Figure 12. Example of how data Sgroups can be used to identify sites with absolute (*R* and *S*) and relative (*r*) stereo configuration. The data "Relative Trans" additionally indicates that the hydroxyl groups are trans to each other.

field control as for data which are attached to the structure (text, numeric, and formatted).

Cases where one may want to link data directly to a chemical Sgroup include percent composition (e.g., of a component in a mixture or formulation as shown in Figure 7 or of an SRU in a copolymer), structure identifiers (e.g., UV stabilizer as illustrated in Figure 11, plasticizer, pigment, or antioxidant), or average molecular weight of a polymer. There are also many cases where one might want to attach data to an arbitrary set of atoms and bonds. For example, data Sgroups can be used to: (1) annotate a particular set of atoms, e.g., as part of a pharmacophore or as "reactive" sites; (2) annotate fragments, e.g., with stoichiometric multipliers of a salt/solvate, as major/minor, or as active/inert; (3) describe an unknown portion of a structure by attaching data to a *atom, e.g., the "binder" used in the formulation shown in Figure 7; (4) explain more fully the nature of a particular site, e.g., stereochemical purity or isotopic purity. For example, in Figure 12 data Sgroups are used to identify sites that have absolute and relative stereo configuration. The sites that have absolute stereo configuration are labeled "*R*" and "*S*", while the sites that have relative stereo configuration are labeled "*r*". The data "Relative Trans" additionally indicates that the hydroxyl groups are trans to each other. In short, the use of data Sgroups permits a user-extensible chemical structure representation which is fully integrated with the connection table and which can be stored in a database or transfer file.

III. SGROUP USER INTERFACE

In the MACCS-II program, structure input and manipulation can be driven through a series of menus or by a command language. The menus are referred to as the "Draw Menus" and allow the user to build up structures for registration in a database or for use as a query in a search. The command language, referred to as the Molecular Editor or MEDIT,⁹ supports nearly all the functionality offered in the Draw Menus. The Draw Menus use interactive graphics, while the Molecular Editor may be interactive or batch-oriented. The Sgroup features mentioned above have been implemented in both of these areas.

A. Draw Menus. A new side menu for drawing and manipulating structures with Sgroups has been added to the Draw Menus (Figure 13). Chemical Sgroups are defined through a mechanism where the user identifies the crossing bonds and a "seed" atom in the group. The seed atom is propagated outward until a crossing bond boundary is hit or until the entire fragment is encompassed in the Sgroup definition. Once the group is defined, brackets are automatically calculated and

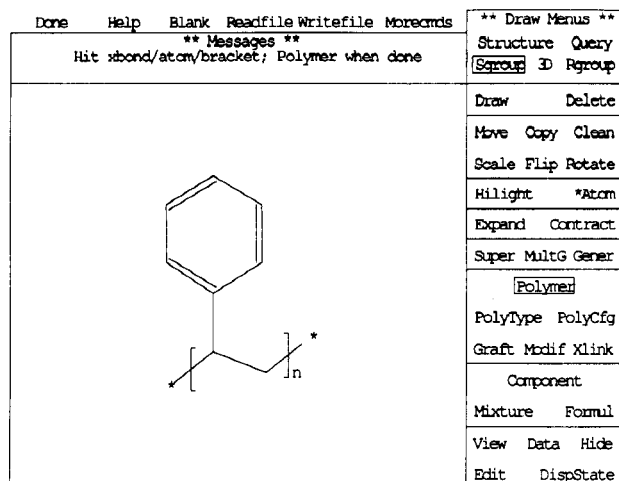


Figure 13. Sgroup Draw Menu. The command options are located at the top and the right, while the drawing area and message box are positioned center and left.

Table II. Chemical Sgroup Types, Subtypes, and Corresponding Display Labels

Sgroup type, subtype	display label
SRU	n
monomer	mon
mer	mer
copolymer	co
copolymer, alternating	alt
copolymer, block	blk
copolymer, random	ran
graft	grf
modification	mod
crosslink	xl
anypolymer	anyp
component	c ^a
mixture	mix
formulation	f
superatom	user defined
multiple group	an integral number, e.g., 5
generic	no label

^a Components in mixture Sgroups are unnumbered, whereas components in formulation Sgroups may be numbered, e.g., c2.

displayed by the program. The user then has the option to graphically modify the display of the brackets. Data Sgroups are defined by specifying the chemical Sgroup to which it is linked or by specifying each atom and bond to be included. On definition, their display is also automatically determined and can be thereafter modified by the user.

Chemical Sgroups are depicted graphically by brackets, whereas data Sgroups are displayed as character strings. These brackets and character strings serve as the "handles" to these objects. For example, a chemical Sgroup can be moved, re-defined, deleted, etc. by selecting the appropriate option on the side menu and then touching a bracket associated with the chemical Sgroup of interest. Chemical Sgroups have an additional label in the lower right corner that serves as an identifier for the Sgroup. Generic Sgroups are the only exception, they are generated with no labels. A list of the identifiers used in the Substance Module is provided in Table II. The brackets' positions are determined automatically when a chemical Sgroup is defined, but may be altered by the user. The position of a data Sgroup is also determined at the time of definition using defaults, but the user has a variety of methods which afford a great deal of flexibility in modifying how the data Sgroup is to be displayed.

The user interface has utilities to enforce conventions and to simplify certain operations. For example, in order to define a simple modification polymer, the user first draws the un-

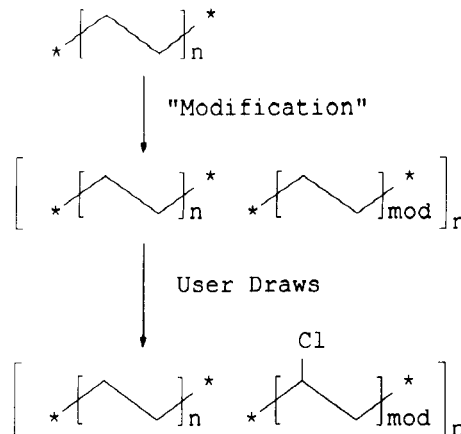


Figure 14. Operation of the modification button (Modif) in the Sgroup Draw Menu to simplify the entry of a modified homopolymer.

modified SRU, then picks the "Modif" button, and indicates the SRU which was modified. The program then makes a copy of that SRU, changes the type of the copy to "Modification", and encloses both in another polymer Sgroup. Only three penhits are required for the user to carry out this sequence of events. The user can then make the chemical modification on the copy of the SRU (Figure 14).

The hierarchy of components in mixture/formulation Sgroups is enforced by the user interface. The user interface also includes simplifications such that component Sgroups are automatically defined when a mixture or formulation Sgroup is defined; but the user has the option of defining components, e.g., for the purpose of collecting multiple fragments in one component, as in the case of a salt.

B. Molecular Editor. The Molecular Editor has been extended to allow for the definition, identification, modification, and deletion of Sgroups. The Molecular Editor is a powerful tool that can be used to automate and simplify many operations. The existing functionality is ideally suited for custom applications, e.g., input of complete peptide structures via superatom templates.

IV. SGROUP SEARCHING

The two most significant searching protocols are full-structure and substructure searching. A brief description of each searching protocol is given below followed by a discussion on how Sgroups have been addressed in these areas.

Full-structure searching is the most diverse class of searching in MACCS-II. It can range from an exact match search where the structure in the database must match the query exactly to a fuzzier search where certain attributes of the structure can differ but still find a match, e.g., bond types need not match exactly in a tautomer search. Several searches of this type have been identified as having general utility: current (or exact), tautomer, isomer, formula, and salt. Consequently, each of these searches have been packaged into a single command that can be executed by the touch of a button.

A new, customizable, full-structure search has recently been implemented in the latest release of MACCS-II. The "flexmatch" search, as it is called, is based on a number of switches that control which structural attributes must match exactly. The switches act like filters and give the user the ability to fine-tune his search criteria to a level that has never been possible before. Table III lists the switches that are available through a flexmatch search. A typical flexmatch search might have the CHA (charge) and MAS (mass) switches set. Such a search would ensure that the charge and isotopic labels match on each atom, while at the same time, all the other structural attributes not specified by switches are allowed to vary.

Table III. Flexmatch Switches in MACCS-II

switch code	structural attribute
BON	bond order
CHA	charge on atoms
DAT ^a	data Sgroups
END ^a	end groups
FRA	all fragments
HYD	hydrogen count on fragments
ION	total charge on fragments
MAS	isotopic labels on atoms
MET	connectivity of metal bond
MIX ^a	component, mixture, and formulation Sgroups
MSU ^a	monomer/SRU uniqueness
POL ^a	polymer Sgroups
RAD	radical values on atoms
SAL	exact salt (counterions must match)
STE	configuration of stereocenters
TAU	tautomer match
TYP ^a	polymer Sgroup types
VAL	valence on atoms

^aThese codes are specific to the Substance Module.

Substructure searches retrieve molecules that contain the query structure embedded within them. An example of a substructure search might be "find all structures that contain a phenyl ring". This search would hit toluene and naphthalene provided they were present in the database being searched.

A. Chemical Sgroup Searching. 1. Polymers. Polymer Sgroups are searchable at the full-structure and substructure level. In a full-structure search, the user can choose to consider or ignore (the corresponding flexmatch codes are given in parentheses) the following: (1) all polymer Sgroups (POL); (2) the polymer type, subtype, and connectivity (TYP); and (3) the end groups (END). It is also possible to choose whether monomers and SRU's related by simple condensation or addition reactions will find each other (MSU). The flexibility of registering either source-based or structure-based polymer representations necessitates having the capability to recognize the equivalence of either form. For example, proprietary database administrators often prefer to minimize redundant database entries and want to determine if a duplicate structure is on file for new registrations. An algorithm has been developed which can selectively ignore information in a candidate structure for certain condensation and addition polymerizations and permit matches in which the target structure is a subset of the query.

Full-structure searching is also capable of taking into account the "phase shift" possibility of SRU's. The choice of the crossing bonds for any SRU is arbitrary. While these are usually chosen to correspond to the bonds formed in the polymerization reaction, this is not required. The searching algorithm used in the Substance Module has been developed in a way that does not require the same choice of crossing bonds in the query and database structure. As for polymer configuration, all values, head-to-tail, head-to-head, and either/unknown, will hit themselves. With these features, it is possible for the structures in Figure 15 to find each other in a full-structure search.

In a substructure search, the query must be a subgraph of the chemical structure in the database in order to be considered a hit. The polymer types, subtypes, and configuration must match as well. There are three main facets to the matching of polymer type and subtype. First, all polymer types will hit polymer Sgroups with the same type. Second, a copolymer query with no specified subtype will hit copolymer Sgroups with any subtype. Third, the anypolymer query type will hit all polymer types. These points are summarized in Table IV. The substructure search mechanism also supports the "phase shift" capabilities described above for full-structure searching. Finally, all values of polymer configuration will hit themselves

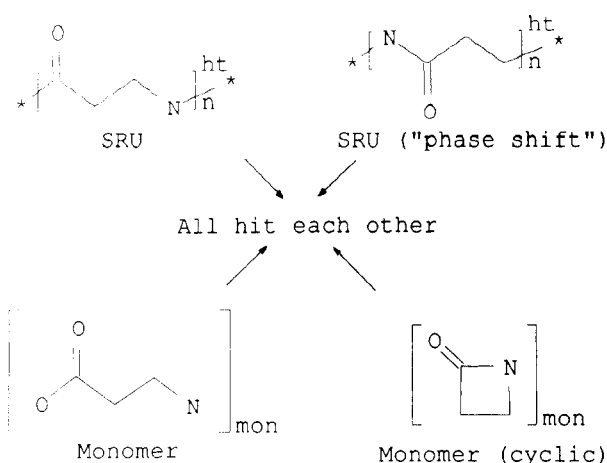


Figure 15. Examples of different source-based and structure-based representations that can be made to hit each other in a flexmatch search.

Table IV. Polymer Sgroup Matching in Substructure Searching

query Sgroup label	database Sgroup label									
	n	mon	mer	co	alt	blk	ran	mod	xl	grf
n	X									
mon		X								
mer			X							
co				X	X	X	X			
alt					X					
blk						X				
ran							X			
mod								X		
xl									X	
grf										X
anyp	X	X	X	X	X	X	X	X	X	X

while the either/unknown value will additionally hit head-to-tail and head-to-head.

2. Components, Mixtures, and Formulations. Component, mixture, and formulation Sgroups are searchable at the full-structure and substructure level. The MIX switch in a flexmatch search allows the user to specify whether the component, mixture, or formulation Sgroups are required to match as well as the chemical structure. For substructure searching, the component ordering in formulations is considered "relative", i.e., the query hits a database structure if the components appear in the same order in both the query and the database structure. However, the user has the option to ignore component ordering by specifying the query with unnumbered components.

3. Drawing and Display Shortcuts. The drawing and display shortcuts (superatoms, multiple groups, and generics) are not searchable Sgroups. However, the underlying chemical structures are still searchable. For example, a phenyl ring in a search query will hit a phenyl ring defined as a superatom in the database and vice versa.

B. Data Sgroup Searching. Data Sgroups are searchable at the full-structure and substructure level. In a search query, the user defines each data Sgroup that is required; each of these has a set of atoms/bonds, a field, and a value. The DAT switch in a flexmatch search specifies whether data Sgroups should match exactly or be ignored. In a substructure search, a candidate is considered a hit if it contains a data Sgroup that is defined by the same atom/bond set (or superset) in the query data Sgroup, for the particular field specified by the query, and with a value matching the query's value. In the Substance Module implementation, the language for matching data Sgroup values is identical to that used in searching other conventional MACCS-II data fields.

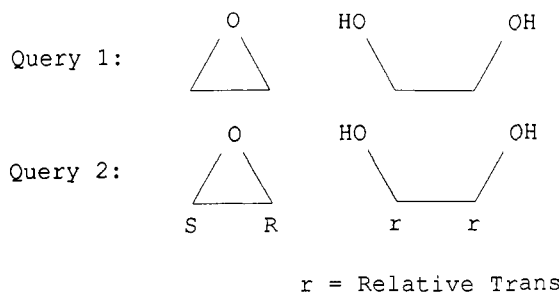


Figure 16. Two substructure search queries illustrating data Sgroup searching. See text for description.

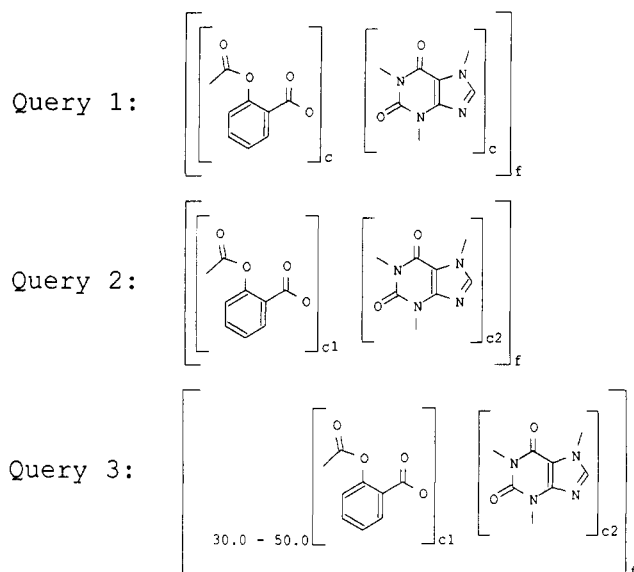


Figure 17. Three substructure search queries illustrating formulation Sgroup searching and numeric data Sgroup searching. See text for description.

V. SELECTED EXAMPLES

The following examples illustrate some of the capabilities that are available in the Substance Module. The first two examples highlight data and formulation Sgroup searching. The last example demonstrates how the Substance Module can be used in concert with the MACCS-II form-making capabilities to organize chemical data and generate reports.

As noted above, data Sgroups provide a means of more fully describing structural features. In Figure 12, data Sgroups are used to specify absolute and relative stereo configurations. Applying such data to a structure permits greater selectivity in searching by allowing more precise queries. For instance, the two substructure search queries shown in Figure 16 essentially ask the following questions:

- query 1 "Find all structures that contain an epoxide and a 1,2-diol."
- query 2 "Find all structures that contain an epoxide and a 1,2-diol, where the epoxide has absolute stereo configuration and the diol has relative stereo configurations."

The first query is the least precise and will hit the structure in Figure 12 as well as a similar structure that lacks the data Sgroups. The second query, however, will only hit structures like the one in Figure 12 that contain an epoxide and a 1,2-diol as well as the pertinent data Sgroups.

The three substructure search queries shown in Figure 17 illustrate formulation Sgroup searching and range searching for numeric data Sgroups. The queries translate to the following questions:

- query 1 "Find all formulations that contain aspirin and caffeine."

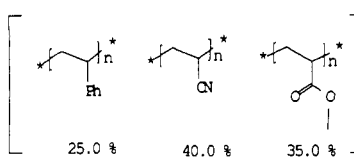
				Specific Gravity
				1.06
				Water Absorption - 24 h @ 73 F
				0.16 %
				Melt Viscosity: Poise @ 446 F - @ 100 sec-1
Run 2 -- Batch A				14,000
				Deflection Temperature @ 66 psi, 1/4"
				205 F
				Mold Shrinkage
				5 - 7 10E-3 in/in
RUN	Batch	Tensile Strength	Flexural Strength	Flexural Modulus
1	A	7,495 psi	9,750 psi	370,000 psi
2	A	7,450 "	9,950 "	350,000 "
3	A	7,380 "	9,900 "	350,000 "
1	B	6,020 "	7,850 "	300,000 "
2	B	6,475 "	7,900 "	300,000 "
3	B	6,400 "	8,250 "	310,000 "
1	C	5,390 "	6,200 "	240,000 "
2	C	5,285 "	6,400 "	230,000 "
				Tensile Elongation
				20 %
				20 %
				25 %
				60 %
				55 %
				50 %
				65 %
				60 %

Figure 18. Example report generated in MACCS-II summarizing the data collected for three different batches of the terpolymer acrylonitrile, styrene, and methylacrylate.

query 2 "Find all formulations that contain aspirin and caffeine, where aspirin was added before caffeine."

query 3 "Find all formulations that contain aspirin and caffeine, where aspirin was added before caffeine, and where aspirin makes up 30-50% of the formulation by, e.g., weight."

Note that the second query has numbered components while the first does not. By numbering the components, the user is indicating that the relative order of the components in the formulation is important. Consequently, query 2 will hit the formulation given in Figure 7, but it will not hit a similar formulation where aspirin was added after caffeine. Since the components are not numbered in query 1, this query will hit all formulations that simply contain aspirin and caffeine components. Query 3 similarly requires that aspirin be added before caffeine, but it also requires that aspirin must make up 30-50% of the formulation by weight. This additional criteria is specified through a numeric data Sgroup.

The last example (Figure 18) illustrates how one can organize research data and generate reports by using the Substance Module in concert with the MACCS-II form-making capabilities. The data displayed in Figure 18 summarizes the results for a research project aimed at developing a new liner for a hot tub. The liner itself is a random terpolymer of acrylonitrile, styrene, and methylacrylate. The liner must be durable and capable of handling a variety of mechanical stresses in order to be successful. Several common measures of durability are tensile strength, flexural strength, flexural modulus, and tensile elongation. These properties were used to gauge the mechanical performance of several different batches of the targeted terpolymer. The batches A, B, and C differ from each other only in the relative percent composition of the respective monomeric units that make up the terpolymer. Several runs were taken for each batch, and the results are shown in Figure 18. Batch A demonstrates the best overall performance with regard to requirements for the intended use.

The form, i.e., the boxes, in Figure 18 was generated in MACCS-II, while the polymer representation and percent composition data were afforded by the Substance Module. By storing this information in a database, it not only makes report generation possible, but it also makes the information more accessible to a broader audience. For example, although the properties for batches B and C are not optimal for usage in a hot tub liner, they may be for some other products such as truck tops or RV parts. Thus, a researcher interested in making a new truck top might benefit from work done on an

unrelated project provided the chemical information is made easily accessible.

VI. CONCLUSIONS

In this paper, we have described how Sgroups can be used to dramatically extend representation of chemical structures in computer programs. In the Substance Module, chemical Sgroups are used to represent, store, and search polymers, biopolymers, mixtures, and formulations. A more general class of data Sgroups was also implemented which permits the user to add properties to the connection table and have them searchable as an integral part of the structure.

ACKNOWLEDGMENT

We gratefully acknowledge helpful discussions with many of our colleagues in the chemical industry, and the contributions of many other employees of Molecular Design Limited, particularly those of James Dill, David Hughes, and Jorge Manrique.

REFERENCES AND NOTES

- (1) For general discussions of chemical information management systems, see: (a) *Chemical Structures: The International Language of Chemistry*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1988, and references therein. (b) *Communication, Storage and Retrieval of Chemical Information*; Ash, J. E., Chubb, P., Ward, S. E., Welford, S. M., Willett, P., Eds.; Ellis Horwood: Chichester, 1985, and references therein.
- (2) For general discussions of the MACCS-II system, see: Ahrens, E. K. F. Customisation for Chemical Database Applications. In *Chemical Structures: The International Language of Chemistry*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1988; pp 97-111, and references therein.
- (3) Martin, Y. C.; Bures, M. G.; Willett, P. Searching Databases of Three-Dimensional Structures. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1990; Chapter 6, pp 213-263.
- (4) Christie, B. D.; Henry, D. R.; Güner, O. F.; Moock, T. E. MACCS-3D: A Tool for Three-Dimensional Drug Design. In *Online Information 90-14th International Online Information Meeting Proceedings*; Raitt, D. I., Ed.; Learned Information: Oxford, 1990; pp 137-161.
- (5) Güner, O. F.; Henry, D. R.; Pearlman, R. S. Use of Flexible Queries for Searching Conformationally Flexible Molecules in Databases of Three-Dimensional Structures. *J. Chem. Inf. Comput. Sci.*, submitted for publication.
- (6) The developed capabilities are quite general and can be used to represent a wider range of substances than discussed in this paper, e.g., derivatized surfaces, monolayer assemblies, etc.
- (7) For a review of polymer chemistry, see: (a) *Textbook of Polymer Science*; Billmeyer, F. W. Jr., Ed.; Wiley: New York, 1984. (b) *Principles of Polymerization*; Odian, G., Ed.; Wiley: New York, 1981.
- (8) *Textbook of Polymer Science*; Billmeyer, F. W. Jr., Ed.; Wiley: New York, 1984; pp 6, 55.
- (9) The Molecular Editor is a set-oriented structure manipulation language. This language will be the subject of a future paper.

Improvements in Derwent Plasdac System[†]

JULIE A. BRIGGS,* EDGAR A. FERNS, and KATHLEEN E. SHENTON

Derwent Publications Ltd., 128 Theobalds Road, London WC1X 8RP, Great Britain

Received June 1, 1991

The Plasdac Code has undergone many changes, culminating in the introduction of Key Serial Numbers in 1978. Further development along the same path is not possible, so with subscriber agreement Derwent began a major revision. Plasdac subscribers were involved from the start in the design. Initially a questionnaire was sent out to ascertain the problem areas; later an Advisory Group was formed consisting of nine Plasdac users and representation by the Japanese Plasdac Association. The main requirements were continuity with the old code, extendability, no significant change in coverage, retention of the hierarchical nature, and linking between related terms to make searching more specific. The enhanced code comprises greatly extended nonstructural codes; Specific Compound Numbers for chemicals, which for polymer formers are embedded in a hierarchy of generic terms based on the current system; Chemical Aspects terms for generic structure searches; and levels of linking via proximity operators.

BACKGROUND

It was 25 years ago that Derwent introduced the Plasdac Code in order to handle polymer information in the patent literature. In the beginning the Plasdac Punch Code, as it was known in those days, was based on a punch card containing 80 columns and 12 rows which represented a maximum of 960 possible punch positions for encoding data. Due to this limit to the number of such positions available, groups of punch positions were used to represent concepts, and this multiple usage led to poor relevance in searching.

Take, for example, the concept "acetate", which is represented by the single punch position 067. In order to create a code for the concept vinyl acetate, 067 was combined with the punch position vinyl carboxylic esters (066); for the concept cellulose acetate, 067 was combined with three other punch positions cellulose (252), modified polymer (231), and esterification (239).

In 1976 the punch-card system was loaded on-line with each punch position being represented by an alphanumeric code, derived from its coordinates on the card. Several modifications and improvements took place over the years, including in 1977 the addition of codes for all the elements, but this was achieved by the combination of punch positions to create the extra codes required. Such a system is unavoidably prone to false drops if separate indexing ideas are expressed on the same "card records", since spurious combinations of codes can be recovered.

1978 saw the most dramatic of these modifications with the creation of Key Serial numbers (KS). These were assigned to precoordinated groups of punch codes. Thus, a key serial was created for propylene homopolymer, ethylene binary copolymer, phenol monomer, and so on. Key serial numbers were also assigned to some concepts such as "carbon black light stabiliser" and "glass fibre filler". The obvious benefit of these key serials was the ability to search specifically for those combinations of concepts to which they had been applied. However, since key serials were unique to each concept, they

[†] Presented at ACS Spring Meeting Polymer Symposium in Atlanta on April 16, 1991.