The Molecular Information File (MIF): Core Specifications of a New Standard Format for Chemical Data

Frank H. Allen*

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, England

John M. Barnard

BCI Ltd, 46 Uppergate Road, Stannington, Sheffield S6 6BX, England

Anthony P. F. Cook

Synopsys Scientific Systems, 175 Woodhouse Lane, Leeds LS2 3AR, England

Sydney R. Hall

Crystallography Centre, University of Western Australia, Nedlands 6009, Australia

The initial core data definitions of a universal data exchange format for chemical information is proposed. The Molecular Information File (MIF) represents a coalescence of two major format developments: the Standard Molecular Data Format (SMD) and the Crystallographic Information File that conforms to the STAR (Self-Defining Text Archive and Retrieval) syntax. Essentially, the MIF is a re-expression of the major SMD concepts using STAR syntax. The core data items, defined here, cover chemical connectivity representations, stereochemistry, and associated data for complete 2D and 3D molecules and for 2D substructural queries.

INTRODUCTION

Each advance in computer technology, database design, and network capability contributes to the need for a universal method of exchanging data electronically. Nevertheless, specialized formats continue to proliferate for specific applications. This diversity of formats represents a serious obstacle to the global exchange of data and to the development of general access methods for databases.

The need for flexible, extensible, and universal methods for exchanging scientific data is recognized.[1-4] The most important attributes for chemical data files are machine independence and portability; the suitability of a format for long term archiving; the ability to add new data without corrupting existing archives; and a simple robust mechanism for moving data between software packages. In short, a common data file structure is needed to provide a general communication link that can be used by researchers, journals, and databases.

Chemical software packages offer a wide variety of tools for acquiring structural knowledge and for applying this knowledge to problems in the structural domain. Machine-readable representations of two-dimensional (2D) molecular structures are central to these applications. The dissemination, searching, and manipulation of large collections of these representations are basic requirements of modern computational chemistry. Further, molecular modeling, via force field, semiempirical and *ab initio* techniques, is used routinely to predict and categorize the three-dimensional (3D) aspects of molecular structure from this basic 2D information. Other more rapid processes now employ rule-based and artificial intelligence methodologies to generate 3D structural models from 2D representations.[5,6] Many of these ap-

proaches may be used to create in-house archives and databases of 3D structural models as an adjunct to existing databases of experimental determinations of 3D structure.

Most packages have the potential to link the relevant databases of molecular structure, and, as such, they form a vital part of the research and development activities of industry and academia. The exchange of chemical data is, however, often thwarted by highly specific data formats and by the myriad methods used to represent 2D structures, stereochemical descriptors, and 3D structural attributes. These data interchange bottlenecks have detracted from an effective use of the large financial and intellectual investment in proprietary software and database systems. They have also contributed to the need for in-house software, which must be continually upgraded and maintained to accommodate format changes within imported systems.

In this paper, we describe the initial core data definitions of a universal data exchange format for chemistry. The Molecular Information File (MIF) format conforms to the Self-defining Text Archive and Retrieval (STAR) file syntax.[7,8] The definition of the MIF data items is based on the draft revised version of the Standard Molecular Data (SMD) format.[1b] The MIF philosophy is intended to be complementary to the Crystallographic Information File (CIF),[9] which has been adopted by the International Union of Crystallography as the universal standard for the exchange and archiving of crystallographic data.

The MIF represents the coalescence of two major developments: SMD and STAR/CIF, and it is appropriate to preface this paper with a brief history of these activities.

HISTORICAL BACKGROUND

The Standard Molecular Data (SMD) format was initially developed by a group of European pharmaceutical companies

Chart 1

Dictionary name: **mif_core.dic**

_atom_attach_all
_atom_attach_ring
_atom_attach_nh
_atom_attach_h   *(numb)*

The number of atom sites considered to be attached (*i.e.* chemically bonded) to this site. The extensions are

    all all sites
    ring all sites forming rings
    nh all sites excl. hydrogens and unshared electron pairs
    h hydrogen sites

Appears in list containing _atom_id. The permitted range is $0 \rightarrow \infty$.   [atom]

_atom_charge   *(numb)*

Specifies the formal electronic charge on the atom for the different atomic representation conventions. The convention for charge is specified by _define_bonding_convention.

Appears in list containing _atom_id. The permitted range is $-99 \rightarrow 99$.   [atom]

_atom_cip   *(char)*

Specifies the Cahn–Ingold–Prelog designation for the atom. The designators are by Prelog and Helmchen (*Angew. Chem. Int. Ed. Engl.* 1982, **21**, 567–583).

Appears in list containing _atom_id.   [atom]

_atom_coord_x
_atom_coord_y
_atom_coord_z   *(numb)*

Specifies the Cartesian coordinates for the atom at an arbitrary origin and arbitrary orthogonal axes.

Appears in list containing _atom_id. The units extensions are: ' ' (Angstroms *1.0) '_pm' (picometres /100.) '_nm' (nanometres *10.).   [atom]

_atom_id   *(char)*

This specifies a unique code for an 'atom site' in a molecule or fragment. A designated atom site may be occupied by a 'dummy' atom (see _atom_type). A special syntax exists for this code which permits a template molecular fragment to be referred to as an 'atom site', and for the atom within that template to be identified. The syntax is $m > n$ where m is the code identifying the template fragment (contained within a save frame) and n is the code matching an _atom_id value stored within the save frame.

Appears in list as essential element of loop structure. Uniqueness of loop packet tested on _atom_id. May match subsidiary data name(s): _bond_id_1, _bond_id_2, _stereo_vertex_id.   [atom]

_atom_mass_number   *(numb)*

Specifies the mass number or the isotopic state of the atom. The default is the 'most abundant naturally occurring' value.

Appears in list containing _atom_id. The permitted range is $1 \rightarrow \infty$.   [atom]

_atom_label   *(char)*

The code providing unique identification of the atom site for display purposes. See also _atom_type.

Appears in list containing _atom_id.   [atom]

_atom_radical_count   *(numb)*

Specifies the formal radical state of the atom site identified by _atom_id. This is the number of unpaired electrons associated with the atom, *e.g.* the value of 1 specifies one singly occupied orbital on the atom that is not involved in bonding.

Appears in list containing _atom_id. The permitted range is $0 \rightarrow \infty$.   [atom]

_atom_spin_multiplicity   *(numb)*

The spin multiplicity applies only if the _atom_radical_count value is non-zero, and is derived from the formula $2S+1$, where S is the absolute value of the sum of the electron spins ($+\frac{1}{2}$ or $-\frac{1}{2}$) on the atom. Thus, a biradical atom may have the values 1 or 3 [$2(\frac{1}{2}-\frac{1}{2})+1$ or $2(\frac{1}{2}+\frac{1}{2})+1$] to represent a singlet or a triplet biradical.

Appears in list containing _atom_id. The permitted range is $1 \rightarrow \infty$.   [atom]

_atom_type   *(char)*

Specifies the type of atom site identified by _atom_id. Hydrogen atoms will usually not be assigned to sites except where they are connected to more than one other atom, or where it is necessary to reference them, for example, in describing stereochemistry.

H 1	V 23	Rh 45	Ho 67	Ac 89
He 2	Cr 24	Pd 46	Er 68	Th 90
Li 3	Mn 25	Ag 47	Tm 69	Pa 91
Be 4	Fe 26	Cd 48	Yb 70	U 92
B 5	Co 27	In 49	Lu 71	Np 93
C 6	Ni 28	Sn 50	Hf 72	Pu 94
N 7	Cu 29	Sb 51	Ta 73	Am 95
O 8	Zn 30	Te 52	W 74	Cm 96
F 9	Ga 31	I 53	Re 75	Bk 97
Ne 10	Ge 32	Xe 54	Os 76	Cf 98
Na 11	As 33	Cs 55	Ir 77	Es 99
Mg 12	Se 34	Ba 56	Pt 78	Fm 100
Al 13	Br 35	La 57	Au 79	Md 101
Si 14	Kr 36	Ce 58	Hg 80	No 102
P 15	Rb 37	Pr 59	Tl 81	Lr 103
S 16	Sr 38	Nd 60	Pb 82	usp 'un-shared
Cl 17	Y 39	Pm 61	Bi 83	electron pair'
Ar 18	Zr 40	Sm 62	Po 84	dum 'dummy atom'
K 19	Nb 41	Eu 63	At 85	
Ca 20	Mo 42	Gd 64	Rn 86	
Sc 21	Tc 43	Tb 65	Fr 87	
Ti 22	Ru 44	Dy 66	Ra 88	

Appears in list containing _atom_id.   [atom]

_atom_valency   *(numb)*

Specifies the valency state of the atom site identified by _atom_id.

Appears in list containing _atom_id. The permitted range is $-99 \rightarrow 99$.   [atom]

_bond_cip   *(char)*

Specifies the Cahn–Ingold–Prelog designation for the bond. The designators are by Prelog and Helmchen (*Angew. Chem. Int. Ed. Engl.* 1982, **21**, 567–583).

Appears in list containing _bond_id_1, _bond_id_2.   [bond]

_bond_environment   *(char)*

Specifies the connection environment of the bond.

    ring   in a ring
    chain   in a chain

Appears in list containing _bond_id_1, _bond_id_2.   [bond]

_bond_id_1
_bond_id_2   *(char)*

Specify the atom site codes of chemically 'connected sites'. The atom id codes may appear in either order except for asymmetric bonds. Each 'bond' pair may be represented only once. The special syntax for representing atom sites within molecular templates is described in the _atom_id definition.

Appears in list as essential element of loop structure. Uniqueness of loop packet tested on _bond_id_1, _bond_id_2. **Must** match data name _atom_id.   [bond]

_bond_type_casreg3   *(char)*

Code indicating the nature of the bond according to the CASREG3 bonding convention. The convention is specified with _define_bonding_convention. See the reference: Mockus, J.; Stobaugh, R.E. *J. Chem. Inf. Comput. Sci.*, 1980, **20**, 18–22.

S	single exact bond
D	double exact bond
T	triple exact bond
A	ring alternating normalised bond
U	tautomer normalised bond

Appears in list containing _bond_id_1, _bond_id_2. Related item(s): _bond_type_mif (convention), _bond_type_ccdc (convention).   [bond]

414 *J. Chem. Inf. Comput. Sci., Vol. 35, No. 3, 1995*

ALLEN ET AL.

_bond_type_ccdc (*char*)

Code indicating the nature of the bond according to the Cambridge Crystallographic Data Centre (CCDC) bonding convention. The convention is specified with _define_bonding_convention. See the reference: Allen, F.H. *et al. J. Chem. Inf. Comput. Sci.*, 1991, **31**, 187–204.

S	single (2-electron) bond or sigma bond to metal
D	double (4-electron) bond
T	triple (6-electron) bond
Q	quadruple (8-electron, metal-metal) bond
A	alternating normalized ring bond (aromatic)
C	catena-forming bond in crystal structure
E	equivalent (delocalized double) bond
P	pi-bond (metal-ligand pi interaction)

Appears in list containing _bond_id_1, _bond_id_2. Related item(s): _bond_type_casreg3 (convention), _bond_type_mif (convention). [bond]

_bond_type_mif (*char*)

Code indicating the nature of the bond according to the MIF bonding convention. Aromatic and normalised tautomeric bonds cannot be shown. The convention is specified with _define_bonding_convention.

S	single (2-electron) bond
D	double (4-electron) bond
T	triple (6-electron) bond
O	other (*e.g.* coordination) bond

Appears in list containing _bond_id_1, _bond_id_2. Related item(s): _bond_type_casreg3 (convention), _bond_type_ccdc (convention). [bond]

_define_bonding_convention (*char*)

Specifies the convention used for the values of the items:
_atom_charge
_atom_radical_count
_atom_spin_multiplicity
_atom_valency

casreg3	Chemical Abstracts Reg 3
ccdc	Cambridge Cryst. Data Centre
mif	basic MIF

Where no value is given, the assumed value is 'mif'. [define]

_define_stereo_relationship (*char*)

Defines the enantiomorphic relationship of the stereo geometry in the data cell (data block or save frame). The descriptions of independently defined stereo regions, whose centres all have the same inter-centre relationship, may be grouped together. For each such group the inter-centre relationship is specified by _define_stereo_relationship. Where there is only one such group, all the relevant data may be included in the data block or save frame for the molecule as a whole. Where there are several different such groups, each should be shown in a separate save frame, and _reference_stereo_group used to reference the different save frames. For each stereo group, a 2 level loop structure will be used to define the stereo-centres it contains. Each stereogenic atom site is defined by _stereo_atom_id and _stereo_geometry at the first loop level, and by _stereo_vertex_id at the second. Each stereogenic bond is defined by _stereo_bond_id_1, _stereo_atom_id_2 and _stereo_geometry at the first level, and by _stereo_vertex_id at the second level.

absolute	configuration is as shown
relative	configuration is relative
unknown	configuration is unknown
racemic	2 stereoisomers: equal mix of d & l
absolute-excess	2 stereoisomers; excess as shown
relative-excess	2 stereoisomers; excess unknown

Where no value is given, the assumed value is 'unknown'. [define]

_display_colour (*char*)

The colour code specifying the colour of the object identified by _display_symbol. The permitted colour codes are stored, with the RGB ratios, as a separate validation file 'mif_core_colours.val'.

black	000:000:000_RGB	yellow_gold	255:215:000_RGB
white	255:255:255_RGB		
		brown	165:042:042_RGB
grey	192:192:192_RGB	brown_sienna	160:082:045_RGB
grey_light	211:211:211_RGB	brown_beige	245:245:220_RGB
grey_slate	112:128:144_RGB	brown_tan	210:180:140_RGB
blue	000:000:255_RGB	salmon	250:128:114_RGB
blue_light	176:224:230_RGB	salmon_light	255:160:122_RGB

blue_medium	000:000:205_RGB	salmon_dark	233:150:122_RGB
blue_dark	025:025:112_RGB		
blue_navy	000:000:128_RGB	orange	255:165:000_RGB
blue_royal	065:105:225_RGB	orange_dark	255:140:000_RGB
blue_sky	135:206:235_RGB		
blue_steel	070:130:180_RGB	red	255:000:000_RGB
		red_coral	255:127:080_RGB
turquoise	064:224:208_RGB	red_tomato	255:099:071_RGB
		red_orange	255:069:000_RGB
cyan	000:255:255_RGB	red_violet	219:112:147_RGB
cyan_light	224:255:255_RGB	red_maroon	176:048:096_RGB
green	000:255:000_RGB	pink	255:192:203_RGB
green_light	152:251:152_RGB	pink_light	255:182:193_RGB
green_dark	000:100:000_RGB	pink_deep	255:020:147_RGB
green_sea	046:139:087_RGB	pink_hot	255:105:180_RGB
green_lime	050:205:050_RGB		
green_olive	107:142:035_RGB	violet	238:130:238_RGB
green_khaki	240:230:140_RGB	violet_red	208:032:144_RGB
		violet_magenta	255:000:255_RGB
yellow	255:255:000_RGB	violet_dark	148:000:211_RGB
yellow_light	255:255:224_RGB	violet_blue	138:043:226_RGB

Appears in list containing _display_id. Where no value is given, the assumed value is 'black'. [display]

_display_conn_colour (*char*)

The colour code specifying the colour of the object identified by _display_conn_symbol. The permitted colour codes are stored, with the RGB ratios, as a separate validation file 'mif_core_colours.val'.

<<< **include file** mif_core_colours.val, as above >>>
Appears in list at level 2 containing _display_conn_id. Where no value is given, the assumed value is 'black'. [display_conn]

_display_conn_id (*numb*)

The identifying number of a display object which is connected to another display object in level 1 of the loop packet which is designated as _display_id. The id number appearing as the _display_conn_id *must* appear elsewhere in the loop structure as a value for _display_id.

Appears in list at level 2 as essential element of loop structure. **Must** match data name _display_id. [display_conn]

_display_conn_symbol (*char*)

The symbol code for the bond connection to the displayed objects identified by _display_id and _display_conn_id. The symbol codes and descriptions are stored in the validation file 'mif_core_bonds.val'.

.c	dotted line to object centres
-c	dashed line to object centres
+c	dashed+solid line to object centres
1c	solid line to object centres
2c	double solid line to object centres
3c	triple solid line to object centres
4c	quadruple solid line to object centres
.b	dotted line to object boundary
-b	dashed line to object boundary
+b	dashed+solid line to object boundary
1b	solid line to object boundary
2b	double solid line to object boundary
3b	triple solid line to object boundary
4b	quadruple solid line to object boundary
>b	wedge with apex at _display_conn_id
<b	wedge with apex at _display_id

Appears in list at level 2 containing _display_conn_id. [display_conn]

_display_coord_x
_display_coord_y (*numb*)

The projected coordinates of the display object identified in _display_object and _display_symbol. A display diagram has the y axis from bottom to top; the x axis bottom left to right. The display origin is defined with _display_origin. Coordinate values can be scaled using _display_scale and _display_span_.

Appears in list containing _display_id. [display]

_display_id   *(numb)*

The identifying number of the display object described by the combination of _display_object and _display_symbol.

Appears in list as essential element of loop structure. Uniqueness of loop packet tested on _display_id. May match subsidiary data name(s): _display_conn_id.   [display]

_display_span_x
_display_span_y   *(numb)*

The width of the display diagram in _display_coord_ units.

The permitted range is $0. \rightarrow \infty$.   [display_define]

_display_object   *(char)*

Signals how the object identified by _display_id is to be displayed. The value 'text' signals that _display_symbol contains a character string for display; the value 'icon' signals that _display_symbol is a code which identifies the molecular symbol to be displayed; and 'null' signals that no object will be shown at this site.

text	char string specified as _display_symbol
icon	icon specified as a _display_symbol code
.	no object is displayed

Appears in list containing _display_id. Where no value is given, the assumed value is ' . '.   [display]

_display_origin   *(char)*

A code signaling where the origin of the _display_coord_ values should be placed.

centre	at centre of diagram
bottom	at the bottom left corner
top	at the top left corner

Where no value is given, the assumed value is 'bottom'.   [display_define]

_display_scale   *(numb)*

The number of _display_coord_ units per centimetre.

The permitted range is $0. \rightarrow \infty$.   [display_define]

_display_size   *(numb)*

Ths value specifies the maximum y dimension of the object (in _display_coord_ units) identified by _display_symbol.

Appears in list containing _display_id. The permitted range is $0. \rightarrow \infty$.   [display]

_display_symbol   *(char)*

The interpretation of this item is dependent on the value of the item _display_object. If _display_object is 'text' then the _display_symbol is assumed to be a literal character string that should be displayed. If _display_object is 'icon' then the _display_symbol is a code that specifies the molecular symbol or icon to be displayed. The permitted code enumerations for _display_symbol, and their descriptions, are stored in the separate validation file 'mif_core_molecules.val'.

3s	unsaturated 3-membered ring with circle
4s	unsaturated 4-membered ring with circle
5s	unsaturated 5-membered ring with circle
6s	unsaturated 6-membered ring with circle
7s	unsaturated 7-membered ring with circle
8s	unsaturated 8-membered ring with circle
lp	lone pair symbol
*	any text string (when _display_object is 'text')

Appears in list containing _display_id, _display_object.   [display]

_molecule_name_common
_molecule_name_iupac
_molecule_name_cas_8ci
_molecule_name_cas_9ci
_molecule_name_cas_10ci   *(char)*

Specifies the short and systematic names for the molecule described within the data block or save frame. Data names can have the extensions:

iupac IUPAC name (*Pure Appl. Chem.* 1971, **28**, 1–110 & *The Notation of Organic Chemistry*, Oxford, Perg. 1979)
    cas_8ci *Chemical Abstracts 8th Collective Index*
    cas_9ci *Chemical Abstracts 9th Collective Index*
    cas_10ci *Chemical Abstracts 10th Collective Index*

       [molecule]

_reference_conformation
_reference_stereo_group   *(char)*

The framecode pointer strings (starting with a $ character) which identify save frames containing specific categories of data which have been specified with a _define_ data value.

Appears in list as essential element of loop structure.   [reference]

_stereo_atom_id   *(char)*

Specifies the identity of the stereogenic centre to which the sites defined under _stereo_vertex_id are attached.

Appears in list as essential element of loop structure. **Must** match data name _atom_id.   [stereo]

_stereo_bond_id   *(char)*

Specifies the atom site identifiers of the connected atoms that form the stereogenic bond. These atom sites are connected to the atom sites defined under _stereo_vertex_id and which define the stereochemistry of the cited bond.

Appears in list as essential element of loop structure. **Must** match data name _atom_id.   [stereo]

_stereo_geometry   *(char)*

Specifies the geometry of the stereo-centre being described. Note that the enumeration values are not descriptions of the chemical entities, just the stereo geometry. Thus, the stereochemistry of hindered biphenyls may conform to the *allene* geometry, though they are not chemically allenes. The order of the associated _stereo_vertex_id values is determined by the geometry. The first site is selected as that closest to the axis of highest symmetry for the described geometry. Subsequent sites are selected sequentially by RH rotations about this axis; otherwise by selecting a site closest to the initial site.

square	4 sites
olefin	4 sites in rectangle
allene	4 sites in distort tetr.
tetrahedron	4 sites
square_pyramid	5 sites
trigonal_bipyramid	5 sites
octahedron	6 sites
cube	8 sites

Appears in list as essential element of loop structure.   [stereo]

_stereo_vertex_id   *(char)*

Specifies the identity of atom sites at the vertices of the stereogenic centre. The order of the selected vertex sites for each geometry is described in _stereo_geometry. Where one or more of the vertices in the specified geometry corresponds to an implicit hydrogen atom or unshared electron pair, the value '.' may be used. The special syntax for representing atom sites within molecular templates is described in the _atom_id definition.

Appears in list at level 2 as essential element of loop structure. **Must** match data name _atom_id.   [stereo]

**Table 1.** Mif Attributes And Syntax

• a *text string* is string of characters bounded by a [ ]['][″][;]
• a *data name* is a text string starting with an underline
• a *data item* is a text string not starting with underline, preceded by an identifying data name
• a *list* is a sequence of data names, preceded by 'loop' and followed by a list of data items
• a *save frame* is a collection of data within a data block, preceded by 'save*framecode*' and closed with a 'save'.
• a *data block* is a collection of data, preceded by 'data *blockcode*'
• a *global block* is a collection of data, preceded by global, that is common to all subsequent data blocks
• a *file* may contain any number of data blocks or global blocks
• a *data name* must *be unique within a data block*

in the mid-1980s. Draft documents were made available from 1987, and the specification was published[1a] in 1989. A meeting in Frankfurt in 1988 established a series of technical working groups under the auspices of the Chemical Structure Association (CSA) to examine the format specifications in detail and to make recommendations for any revision. As a result, a draft revised format, described as *SMD Version 5.0* was published in February 1990.[1b] A document describing the *core* format, i.e., those data items regarded as essential in any exchange file, was prepared by one of us (J.M.B.) for consideration by Sub-Committee E49.51 of the American Society for Testing Materials (ASTM).

At approximately the same time (1987), the International Union of Crystallography (IUCr) established a Working Party on Crystallographic Information to investigate the electronic storage and transmission of text and data. In 1988 it commissioned one of us (S.R.H.) to coordinate the development of a universal file to replace the existing fixed format Standard Crystallographic File Structure (SCFS).[10] This led to the development of the Crystallographic Information File (CIF) based on the STAR syntax.[7] The final specification of the CIF, which included definitions for data items used in small-molecule and inorganic crystal structure studies, was adopted by the Executive Committee of the IUCr in 1990 and was published[9] in 1991.

The CIF is now employed by crystallographers worldwide for data applications and is a format generated by all major crystal structure packages. CIFs are used for laboratory archives, to transfer data between laboratories and to databases and, most importantly, for the submission of machine readable manuscripts to journals. Extensions to the initial small-molecule and inorganic core definitions are currently being prepared to cover the needs of macromolecular crystallography and of powder diffraction data. The CIF has been adopted by the Cambridge Crystallographic Data Centre (UK), the Protein Data Bank (Brookhaven, U.S.A.), and by the International Centre for Diffraction Data (Newtown Square, PA, U.S.A.), as a data entry format.

Although a CIF is able to store a representation of the topology of small molecules, its data definitions do not fulfill all of the needs of the chemical community. In 1991, the IUCr became interested in further extending the CIF into the chemical area, and discussions took place between representatives of the SMD Technical Working Groups and of the IUCr. These meetings decided that an integration of the SMD format and the STAR syntax was desirable because it provided a number of advantages over the existing SMD specifications.[2] In particular, SMD/STAR provides for a clearer separation of the data structure and the data content
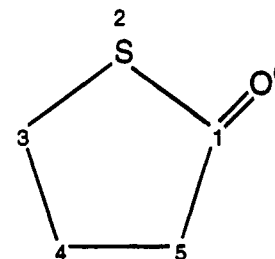

```
# atom attributes list

    loop_
        _atom_id
        _atom_type
        _atom_attach_h
                            1  C  0      2  S  0      3  C  2
                            4  C  2      5  C  2      6  O  0

# bond attributes list

    loop_
        _bond_id_1
        _bond_id_2
        _bond_type_mif
                            1  2  S      2  3  S      3  4  S
                            4  5  S      5  1  S      1  6  D
```

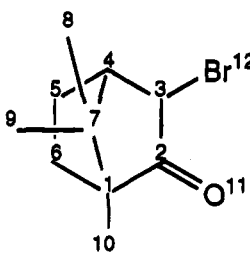**Figure 1.** MIF coding of atom and bond properties for thiabutyrolactone.

roles, together with more flexible data extensibility in future versions. In addition, automated data validation of STAR/SMD files is possible using electronic data dictionaries. In a wider context, there were obvious opportunities for an integration with other applications of the STAR File.

In December 1993, the ASTM subcommittee E49.51 approved a standard specification for the content (i.e., recommended data items) of computerized chemical structural files,[11] although the subcommittee has not so far published any proposals for a format specification. Recently, the Chemical Abstracts Service (CAS) has circulated a draft proposal for a connection-table based exchange format for chemical substances and queries. It uses some ideas that are similar to the 1990 SMD proposal[1b] and is expressed within the framework of the ASN.1 notation.[3] CAS intends to publish this format when it is finalized. MDL Information Systems Inc. have also published a description of their proprietary formats,[4] and a number of other software systems now provide interfaces to these formats.

MIF OBJECTIVES

Molecular information embraces the broad spectrum of data related to chemical and molecular structure. It includes both individual and linked data items, *inter alia*: spectroscopic measurements, thermochemical data, electrochemical properties, crystal structure information, and so on. These represent the object-oriented data descriptors of molecular chemistry, and it is intended that all of these will eventually be accommodated in the MIF approach. This paper describes only the initial MIF core information involving data items needed to specify molecular connectivity and stereochemistry and their 2D and 3D spatial representations. The MIF data items needed for more extensive applications must, in the future, involve the collaborative efforts of informatics and database experts from chemical industry and academia.

A dictionary of the initial MIF core data items described in this paper is given in Chart 1. This is the abbreviated text version of the definition attributes contained in the electronic dictionary file **mif_core.dic** (which is available

```
data_bromocamphor

loop_
    _atom_id
    _atom_label
    _atom_type
    _atom_attach_h
    _atom_coord_x
    _atom_coord_y
    _atom_coord_z
              1  C1   C   0   4.69027   2.57756   2.10705
              2  C2   C   0   3.61112   2.44777   3.16754
              3  C3   C   1   4.16317   2.26258   0.69475
              4  C4   C   1   5.39943   1.87018  -0.13167
              5  C5   C   2   6.48959   2.15784   0.79703
              6  C6   C   2   6.57842   3.69178   1.11990
              7  C7   C   0   5.27609   3.93542   1.94650
              8  C8   C   3   5.93386   1.67891   2.14924
              9  C9   C   3   5.57194   0.17837   2.14326
             10  C10  C   3   6.85298   2.00528   3.35651
             11  O2   O   0   3.03484   2.36201   0.29318
             12  BR3  Br  0   5.41337   2.68686  -1.87368

loop_
    _bond_id_1
    _bond_id_2
    _bond_type_mif
              1   2  S     2   3  S     3   4  S     4   5  S
              5   6  S     6   1  S     1   7  S     7   8  S
              7   9  S     1  10  S     2  11  D     3  12  S
              4   7  S

_display_scale      50
_display_span_x    500
_display_span_y    500

loop_
    _display_id
    _display_object
    _display_symbol
    _display_colour
    _display_size
    _display_coord_x
    _display_coord_y
    loop_
        _display_conn_id
        _display_conn_symbol
        _display_conn_colour
        1    .    .         .    .    251  195  2  1b  black  stop_
        2    .    .         .    .    334  244  3  1b  black  stop_
        3    .    .         .    .    334  339  4  1b  black  stop_
        4    .    .         .    .    251  387  5  1b  black  stop_
        5    .    .         .    .    168  339  6  1b  black  stop_
        6    .    .         .    .    168  244  1  1b  black  stop_
        7    .    .         .    .    217  292  1  1b  black      4  1b  black  stop_
        8    .    .         .    .    191  426  7  1b  black  stop_
        9    .    .         .    .    100  292  7  1b  black  stop_
       10    .    .         .    .    251  100  1  1b  black  stop_
       11  text    O   blue   10    401  206  2  2b  black  stop_
       12  text   Br  yellow  10    417  387  1  1b  black  stop_
```

Figure 2. MIF coding of atom properties (including 3D coordinates), bond properties, and display information for (+)-3-bromocamphor.

via anonymous FTP from the Internet address 130.95.232.12). The core MIF data items provide descriptors for representing the 2D connectivity of a molecule or substructure, the conventions for absolute or relative stereochemical relationships, and the coordinates and conventions used for the generation of 2D and/or 3D graphical depictions. These data items apply to either complete molecules or to substructures with incomplete or variable attributes. As a consequence they are well suited for query definitions in substructure search systems, a feature that will be discussed later in the paper.

MIF CONCEPTS AND SYNTAX

The syntax of the Molecular Information File is based on that of the STAR File.[7,8] A MIF is an ASCII text file that can be read or amended using a standard text editor and can be processed computationally without conversion to another format. The organization and expression of MIF data is summarized in Table 1. Each file consists of a series of *data blocks*, and each block consists of a series individual *data items*. There may be any number of items within a block and any number of blocks within a file. A data block represents a logical grouping of data items, and in most MIF applications a data block will usually specify a complete chemical entity, i.e., a fully-defined molecule or a query substructure.

The MIF syntax, unlike that of a CIF, places no restrictions on line lengths or nested loop levels. For a detailed understanding of the differences between a MIF and a CIF, the reader should refer to the published details of the STAR

```
data_cyclohexane

_molecule_name_common          cyclohexane

     loop_
          _atom_id
          _atom_type
          _atom_attach_h   1  C  2    2  C  2    3  C  2    4  C  2    5  C  2    6  C  2
     loop_
          _bond_id_1
          _bond_id_2
          _bond_type_mif
                    1 2 S    2 3 S    3 4 S    4 5 S    5 6 S    6 1 S

     loop_
          _reference_conformation      $chair    $boat    $twisted_boat

save_chair
     loop_
          _atom_
          _atom_coord_x
          _atom_coord_y
          _atom_coord_z   1    1.579   0.159   0.263
                          2    0.756   0.507  -0.986
                          3    0.825   0.493   1.541
                          4   -0.549  -0.131   1.590
                          5   -1.377   0.222   0.347
                          6   -0.626  -0.158  -0.937
save_

save_boat
     loop_
          _atom_id
          _atom_coord_x
          _atom_coord_y
          _atom_coord_z   1    1.657  -0.426   0.356
                          2    1.031   0.133  -0.927
                          3    0.960   0.133   1.602
                          4   -0.568  -0.040   1.558
                          5   -1.051  -0.738   0.279
                          6   -0.499  -0.028  -0.964
save_

save_twisted_boat
     loop_
          _atom_id
          _atom_coord_x
          _atom_coord_y
          _atom_coord_z   1    0.933   0.922   0.971
                          2    1.186   0.220  -0.368
                          3   -0.119   0.161   1.796
                          4   -1.135  -0.581   0.911
                          5   -1.371   0.181  -0.397
                          6   -0.083   0.236  -1.238
save_
```

Figure 3. Atom and bond properties for cyclohexane, together with 3D coordinate representations of three alternative conformations: chair, boat, and twisted boat.

syntax,[8] the specification of the CIF core data items,[9] and the Dictionary Definition Language[12] used to define data items in the electronic version of a STAR dictionary. CIF data (which numbers close to a thousand items in the 1994 dictionaries) encompass the fields of crystallographic structure and diffraction techniques, e.g., chemical formulae, molecular geometry and crystal structure. These data items can readily be incorporated into a MIF. It should be noted, however, that the reverse may not be true (i.e., MIF data in a CIF) because of the more restrictive CIF syntax.

DATA IDENTIFICATION

The underpinning principle of MIF data is that every data item is represented by a unique *data name* followed by the associated data value. These are referred to as *tag/value* pairs or *tuples*. Data names must start with an underscore (i.e., underline) character. Data values may be any type of string

ranging from a single character to many lines of text. Here are several simple examples of MIF data items.

_atom_mass_number	79
atom_type	Se
_display_colour	blue_medium

The complete list of MIF core data items is given in Chart 1.

LOOPED LISTS

Repetitive data are stored in a MIF as lists of values. Each list is prefaced by a **loop_** statement and a sequence of data names identifying the items which follow in the list as "packets" of data values. The values in each packet match the order and number of data names. Any number of packets may appear in a looped list.
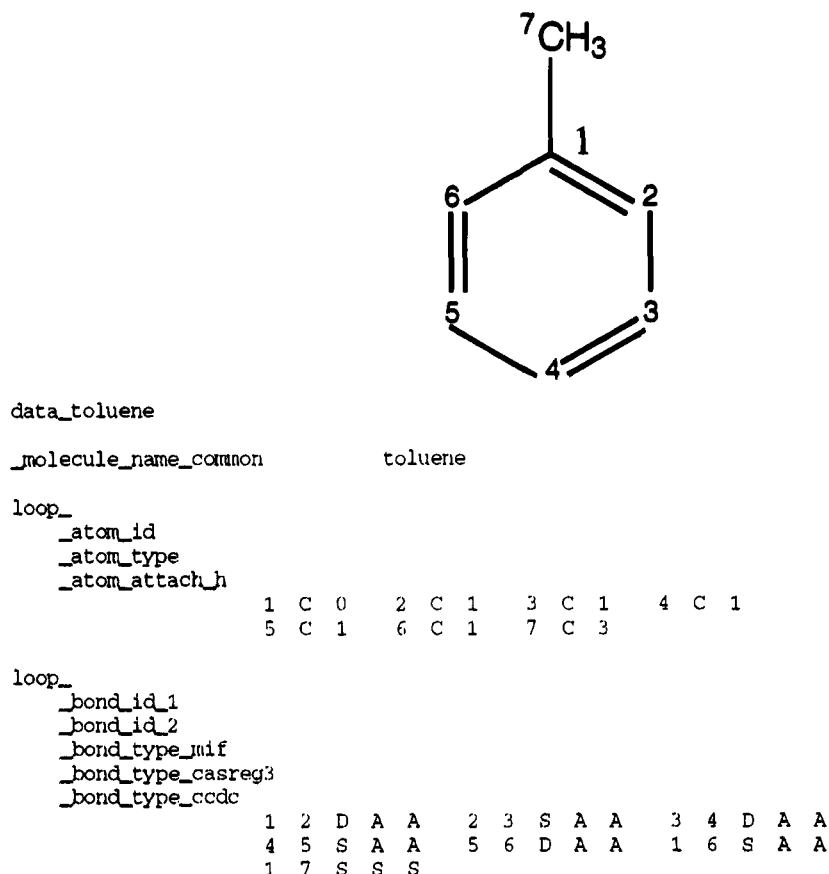
```
       data_toluene

       _molecule_name_common            toluene

       loop_
           _atom_id
           _atom_type
           _atom_attach_h
                            1  C  0    2  C  1    3  C  1    4  C  1
                            5  C  1    6  C  1    7  C  3

       loop_
           _bond_id_1
           _bond_id_2
           _bond_type_mif
           _bond_type_casreg3
           _bond_type_ccdc
                            1  2  D  A  A    2  3  S  A  A    3  4  D  A  A
                            4  5  S  A  A    5  6  D  A  A    1  6  S  A  A
                            1  7  S  S  S
```

Figure 4. Three alternative bonding conventions for toluene stored in the same MIF data block.

Atom and bond properties are typical of the information stored in a looped list. The atoms and bonds of *thiabutyrolactone* in MIF format are shown in Figure 1 (the description of each data item in this example is given in Chart 1). The number of data values in each list is an exact multiple of the number of data names at the start of each loop structure. Looped lists are terminated by the next list or any other data name, data block, or end of file. As shown in Figure 1, comments may be included in a MIF and are preceded by a # character.

Complex lists of data may require the use of *nested* loop structures (see the _display_loop in Figure 2). Data items which commonly appear in looped lists are identified in the MIF dictionary (see Chart 1). In the electronic dictionary these items have the attribute _list set to either "yes" or "both". Other interdependencies of looped data items are also specified in the dictionary.

SAVE FRAMES

Save frames are employed in MIFs to encapsulate grouped data for convenient cross referencing. For applications which require the same group of data to be referenced repeatedly, it is efficient to place this data into an addressable data cell. Molecular fragments, such as amino acid units, are a case in point. A save frame is bounded by the statement **save_***framename* and terminated by a **save_** statement. This data cell can be referenced within the parent data block using the **$framename** keyword. Note that all data names must be unique within the save frame, but the same data names may appear in other save frames or in the parent data block. Save frames may not contain other save frames, but save frame references (*$framename*) may appear in other save frames.

Save frames can be used in a MIF for many purposes, and a simple application, the storage of alternative 3D conformational representations within a data block describing cyclohexane, is illustrated in Figure 3. Within the STAR syntax, save frame references (*$framename*) may occur before or after the save frame definition within any data block. The MIF preserves this basic STAR syntax, and a two-pass parsing mechanism may be necessary to interpret certain data blocks. Save frames are particularly useful for defining commonly referenced structural templates; and examples of this facility are discussed and illustrated (Figures 5—7) in a later section of this paper.

DATA BLOCKS

A *data block* is a sequence of unique data items or save frames. It is opened with a **data_***blockname* statement, and it is closed by another data block statement or a **global_** statement (see below). The *blockname* string identifies the block within the file. Examples of data blocks are shown in Figures 2—4. Each data block in a file must have a unique *blockname*.

GLOBAL BLOCKS

A *global block* is similar to a data block except that it is opened with a **global_** statement and contains data which is common or "default" to all subsequent data blocks in a file. Global data items remain active until respecified in a subsequent data block or global block.

In some applications it may be efficient to place data that is common to all data blocks within a global block. In particular, save frames may be defined within global blocks and then referenced in subsequent data blocks [this statement
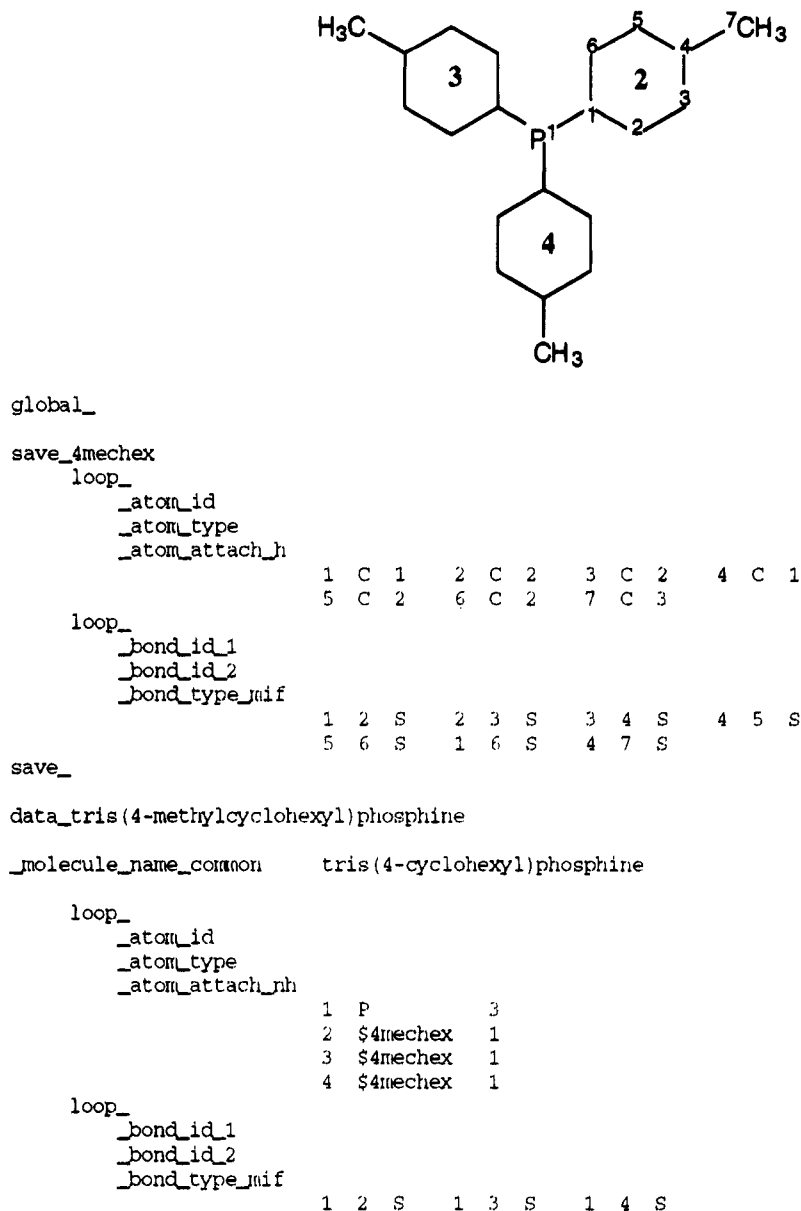
```
global_

save_4mechex
        loop_
            _atom_id
            _atom_type
            _atom_attach_h
                            1   C   1       2   C   2       3   C   2       4   C   1
                            5   C   2       6   C   2       7   C   3
        loop_
            _bond_id_1
            _bond_id_2
            _bond_type_mif
                            1   2   S       2   3   S       3   4   S       4   5   S
                            5   6   S       1   6   S       4   7   S
save_

data_tris(4-methylcyclohexyl)phosphine

_molecule_name_common      tris(4-cyclohexyl)phosphine

        loop_
            _atom_id
            _atom_type
            _atom_attach_nh
                            1   P               3
                            2   $4mechex        1
                            3   $4mechex        1
                            4   $4mechex        1
        loop_
            _bond_id_1
            _bond_id_2
            _bond_type_mif
                            1   2   S       1   3   S       1   4   S
```

Figure 5. MIF representation of tris(4-methylcyclohexyl) phosphine using a "pre-prepared" global save frame for the 4-methylcyclohexyl ligand (see text).

corrects an error in ref 8]. Examples of global data are shown in Figures 5—7. Here, a variety of frequently referenced structural units are encapsulated within save frames specified in global blocks.

ATOMS, BONDS, AND MOLECULAR REPRESENTATIONS

The MIF Core dictionary (Chart 1) describes the principal data items needed to specify molecular connectivity and spatial representations. MIF data items are grouped according to purpose or, as referred to in the DDL dictionary language,[12] by *category*. Categories are formally specified in the electronic version of the dictionary using the data attribute _category. They may also be recognized from the data names which usually have the form "_<category>_<subcategory>_<descriptor>". Data items in a looped list (i.e., each item has _list set as "yes") must belong to the same category.

The allowed values for some data items are restricted to standard codes defined in the MIF dictionary. The items _bond_type_mif and _define_stereo_relationship are two

data items where only the allowed codes may be used, otherwise the data, and the MIF, are considered to be invalid. The role of dictionaries in determining the validity of data is an important aspect of MIF applications. For example, standard codes must also be used for the data items _display_colour and _display_conn_colour, which are used to specify colors for the "atom" and "bond" graphical objects. Only the color codes specified in the MIF dictionary are recognized as valid. There are important practical reasons for this. Each code has associated with it a red/green/blue (RGB) ratio which can be translated by the MIF application into a standard color. [Technical note: because of the large number of possible color codes, these are stored as a separate dictionary validation file **mif_core_colours.val** which is automatically opened when accessing the dictionary file **mif_core.dic**.]

Figure 2 shows MIF data for the molecule *(+)-3-bromocamphor*. The "atom" list contains the items _atom_id, _atom_type, and _atom_attach_h which identify the chemical properties of the atoms, plus the items _atom_coord_x,*_y, and * z which specify the 3D molecular structure
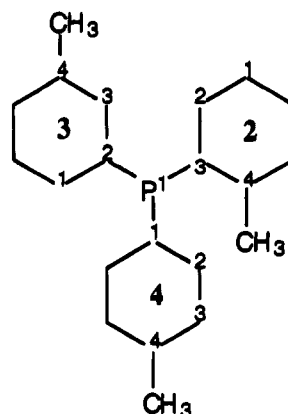
```
global_

save_mechex
        loop_
            _atom_id
            _atom_type
            _atom_attach_h    1  C  2    2  C  2    3  C  2    4  C  1
                              5  C  2    6  C  2    7  C  3
        loop_
            _bond_id_1
            _bond_id_2
            _bond_type_mif    1  2  S    2  3  S    3  4  S    4  5  S
                              5  6  S    1  6  S    4  7  S
save_

data_tris(methylcyclohexyl)phosphine

_molecule_name_common
    (2-methylcyclohexyl)(3-methylcyclohexyl)(4-methylcyclohexyl)phosphine

        loop_
            _atom_id
            _atom_type
            _atom_attach_nh
            _atom_attach_h
                              1  P            3  0
                              2  $mechex   .  .        2>3  .  3  1
                              3  $mechex   .  .        3>2  .  3  1
                              4  $mechex   .  .        4>1  .  3  1
        loop_
            _bond_id_1
            _bond_id_2
            _bond_type_mif
                              1  2>3  S       1  3>2  S       1  4>1  S
```

Figure 6. MIF representation of (2-methylcyclohexyl)(3-methylcyclohexyl)(4-methylcyclohexyl)phosphine using a single global save frame that encapsulates the structure of methylcyclohexane, together with "external" referencing of save frame atoms in _atom_ and _bond_ loops.

in Cartesian coordinates (these are taken from diffraction results[13]). The item **_atom_label** is also used with any graphical depiction of the 3D model. The "bond" loop in this example uses the simple **_bond_type_mif** conventions described in Chart 1. The data names needed to depict stereochemistry are discussed with examples (Figures 8—10) in a later section.

Thus, the MIF approach to representing 2D chemical structure separates the specification of chemical atom and bond properties. This provides additional flexibility in the description of the graphical objects, such as atomic nodes and bonded connections. This is illustrated in the example in Figure 2. The 2D chemical diagram is drawn in a display area of 500 × 500 coordinate units at a scale of 50 units per cm. The default origin [the bottom left corner of the display area] can be specified with the item **_display_origin**. The data used to depict a 2D structure forms a two-level loop with the "atomic" graphical objects at level 1, and the "bond" graphical objects at level 2. The item **_display_object** has the "values" (null or no object), "text" (an element or number

string), or "icon". The size and color of the atom site is specified with **_display_size** and **_display_colour**. The bonds connected to each atom site are specified as a sequence of **_display_conn_**id numbers (in loop level 2). These numbers must match one of the **display_id** numbers at level 1. The connection object is specified with a **_display_conn__symbol** code, which must be a standard value in the dictionary validation file **mif_core_bonds.val**. The color of the icon is specified as a **_display_conn_colour** code.

BONDING CONVENTIONS

Chemical information systems use a variety of conventions for specifying attributes such as aromaticity, bond order alternation, tautomerism, etc. These system-dependent conventions decide the values that are permitted for quantities such as bond order, electronic charge, and hydrogen count. Most systems also provide for redundancy between chemical attributes. For example, the valency, the number of connected non-hydrogen atoms, the number of terminal hydrogens, and the bond types associated with a given atom are
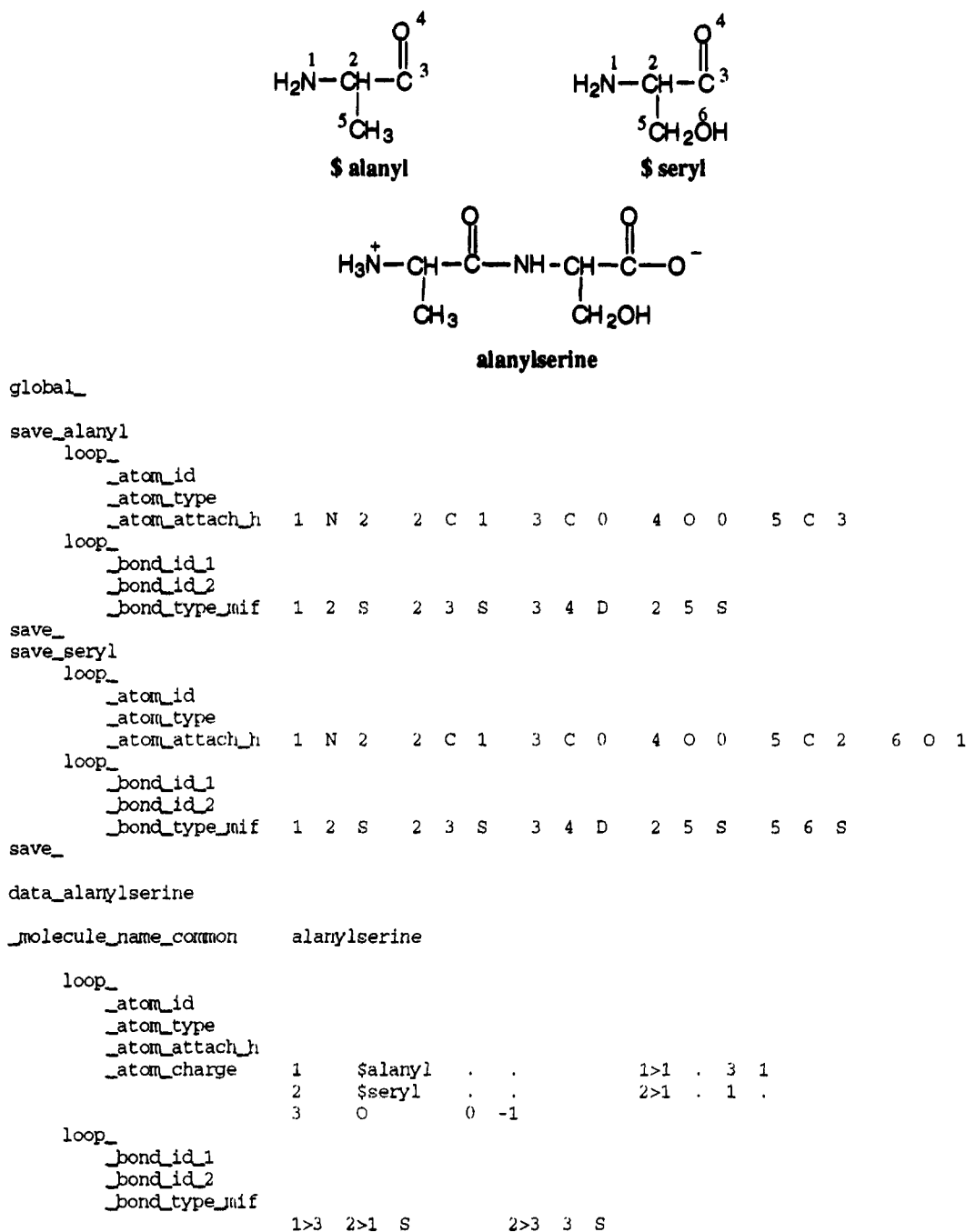
$ alanyl $ seryl

alanylserine

```
global_

save_alanyl
        loop_
            _atom_id
            _atom_type
            _atom_attach_h   1 N 2    2 C 1    3 C 0    4 O 0    5 C 3
        loop_
            _bond_id_1
            _bond_id_2
            _bond_type_mif   1 2 S    2 3 S    3 4 D    2 5 S
save_
save_seryl
        loop_
            _atom_id
            _atom_type
            _atom_attach_h   1 N 2    2 C 1    3 C 0    4 O 0    5 C 2    6 O 1
        loop_
            _bond_id_1
            _bond_id_2
            _bond_type_mif   1 2 S    2 3 S    3 4 D    2 5 S    5 6 S
save_

data_alanylserine

_molecule_name_common     alanylserine

        loop_
            _atom_id
            _atom_type
            _atom_attach_h
            _atom_charge    1    $alanyl    .  .      1>1  .  3  1
                            2    $seryl     .  .      2>1  .  1  .
                            3    O          0  -1
        loop_
            _bond_id_1
            _bond_id_2
            _bond_type_mif
                            1>3  2>1  S        2>3  3  S
```

Figure 7. MIF representation of the dipeptide alanylserine constructed using alanyl and seryl templates encapsulated in global save frames.

related. Systems try to make use of these relationships to perform internal checks and to provide flexibility in the substructure search process. The MIF data definitions provide for three bonding conventions. These are the data items **bond_type_mif, bond_type_casreg3**, and **bond_type_ccdc**. The *mif* convention defines only single, double, triple, and "other" bonds; the _casreg3 convention extends these[14] to include aromaticity in terms of "ring alternating normalized bonds" and tautomerism via a "tautomer normalized bond"; and the _ccdc convention is that employed in the Cambridge Structural Database System[15] to categorize bond types encountered in both organic and metalloorganic molecules.

An important advantage of the MIF approach is that a molecule can be represented using all three bonding conventions within the same data block. An example of alternative bonding conventions encoded for *toluene* is shown in Figure 4.

STRUCTURAL TEMPLATES

In many chemical information systems it is standard practice to build complete 2D molecular representations through the use of a library of commonly referenced structural templates, e.g., ligands, functional groups, amino acid units, etc. In a MIF, templates can be encapsulated as save frames, either within a data block for a specific molecule, or within a global block which is accessible to many data blocks. A simple application of a MIF template is shown in Figure 5 where a 4-methylcyclohexyl ligand is used to encode the molecule *tris(4-methylcyclohexyl)phosphine*. In this example the template is constructed (as the save frame *4mechex*) so that the first carbon can be connected via a single bond to an external site. This is done by setting _atom_attach_h to 1. The molecular connections shown in the _bond_loop which involve the *4mechex* template (_atom_id values 2, 3, and 4 in Figure 5) imply
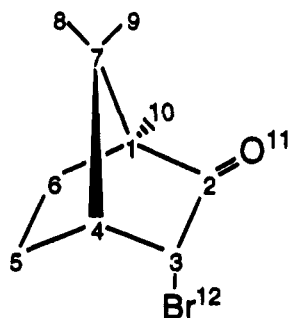
```
data_bromocamphor_2

molecule_name_common        (+)-3-bromocamphor

molecule_name_iupac         3R-bromo-1R,7,7-trimethyl-4S-bicyclo[2.2.1]heptan-2-one

loop_
    _atom_id
    _atom_type
    _atom_attach_h
    _atom_cip
                  1  C  0  R      2  C  0  .      3  C  1  R
                  4  C  1  S      5  C  2  .      6  C  2  .
                  7  C  0  .      8  C  3  .      9  C  3  .
                 10  C  3  .     11  O  0  .     12  Br 0  .
```

Figure 8. CIP stereochemical descriptors for (+)-3-bromocamphor.

connections to the first _atom_id of the save frame, i.e., to C(1), in each case.

"Connection-specific" templates, such as that shown in Figure 5, are useful for many applications, but they are not a general approach to molecular connectivity for the following reasons. (a) Single substitution points can occur at more than one site in a template structure, e.g., at the 1-, 2-, 3- or 4-positions of methylcyclohexane. (b) Connections to the template structure may occur through multiple, rather than single, bonds, e.g., an external carbon atom could connect to C(1) of 4-methylcyclohexane to form a methylenic linkage with loss of both H-atoms from C(1). (c) Two or more atoms in a template structure may be connected simultaneously to external atoms, e.g., in building polypeptide chains from amino acid units. Requirements (a) and (b) could be met by building a very large number of "connection-specific" templates and, indeed, requirement (c) could be accommodated by expanding the syntactical rules to allow alternative connection sites in a template to be tagged and externally referenced.

However, a more general approach to structural templates is to allow the values of specific data items in a template to be substituted in an external reference. More specifically in a MIF, data values in external _atom_ and _bond_ loops may override those present in save frames. This is achieved by providing for a special syntax for the atom site id's in save frames. The syntax *n>m* refers to the _atom_id value *n* assigned externally to the template, and *m* is the _atom_id value within the template. When the save frame contains references to further save frames, this syntax is extended to allow nested atom references to any depth, as in *n>q*. Examples of this type of syntax are shown in Figures 6 and 7.

In the first example (Figure 6), the molecule *(2-methyl-cyclohexyl)(3-methylcyclohexyl)(4-methylcyclohexyl)phosphine* is encoded using a single template (*$mechex*) of the methylcyclohexane molecule. Connections from P(1) to the template atoms 2>3, 3>2, and 4>1 are encoded in the external _bond_ loop. The necessary redefinition of the

hydrogen and non-hydrogen counts of the template atoms is accomplished in the external _atom_ loop. The external values override any values that are contained in, or derived from, the data in the template.

The same approach is used to construct the dipeptide *alanylserine* (see Figure 7). This employs the templates *$alanyl* and *$seryl*. The external _bond_ loop shows that the peptide bond is between the alanyl carboxylic carbon (site 1>3) and the seryl nitrogen (site 2>1) and between the O⁻ (site 3) and the seryl carboxylic carbon (site 2>3). The external _atom_ loop shows how the hydrogen counts of the alanyl and seryl nitrogen atoms are reduced to 1, and the formal positive charge on the alanyl nitrogen is assigned to form the zwitterionic dipeptide structure. Periods in the atom loop indicate null fields, i.e., fields which are not relevant, or for which existing save frame data remains valid.

Note that many of the _atom_id data values in the save frames of Figures 5—7 are the same as those in the main molecules. It is the responsibility of applications software to resolve any ambiguities in expanding up to the full atomic enumeration of each complete main molecule.

STEREOCHEMISTRY AND THE GEOMETRY AT STEREOGENIC CENTERS

The Cahn—Ingold—Prelog (CIP) notation[16] is available in the MIF definitions to specify the stereochemistry of a molecule. The CIP notation is restricted to tetrahedral atomic centers and to olefinic type stereogenic bonds; it is not suitable for describing molecules with partially known stereochemistry, molecules containing more complex geometries, or substructural queries. The MIF data items representing stereochemical quantities are:

    _define_stereo_relationship
    _atom_cip
    _bond_cip
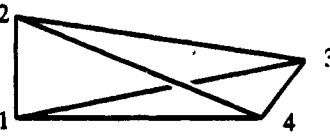   _stereo_atom_id
    _stereo_bond_id_1
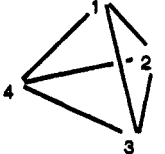   _stereo_bond_id_2
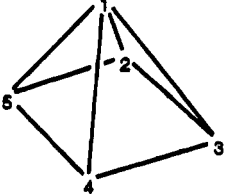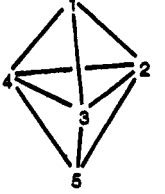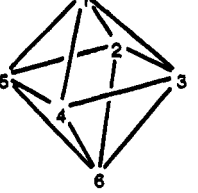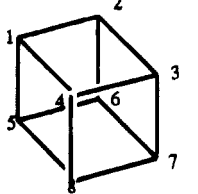
Geometry	Proper rotations	Reflection (chiral)
square (diagram: 2, 3, 1, 4)	$\begin{bmatrix}1\,2\,3\,4\\1\,2\,3\,4\end{bmatrix}_E\ \begin{bmatrix}1\,2\,3\,4\\2\,1\,4\,3\end{bmatrix}_{C2}\ \begin{bmatrix}1\,2\,3\,4\\4\,1\,2\,3\end{bmatrix}_{C4}$	
olefin (diagram: 2, 3, 1, 4)	$\begin{bmatrix}1\,2\,3\,4\\1\,2\,3\,4\end{bmatrix}_E\ \begin{bmatrix}1\,2\,3\,4\\2\,1\,4\,3\end{bmatrix}_{C2}\ \begin{bmatrix}1\,2\,3\,4\\4\,3\,2\,1\end{bmatrix}_{C2}$	
allene (diagram: 2, 3, 1, 4)	$\begin{bmatrix}1\,2\,3\,4\\1\,2\,3\,4\end{bmatrix}_E\ \begin{bmatrix}1\,2\,3\,4\\2\,1\,4\,3\end{bmatrix}_{C2}$ $\begin{bmatrix}1\,2\,3\,4\\4\,3\,2\,1\end{bmatrix}_{C2}$	$\begin{bmatrix}1\,2\,3\,4\\2\,1\,3\,4\end{bmatrix}_\sigma$
tetrahedron (diagram: 1, 2, 3, 4)	$\begin{bmatrix}1\,2\,3\,4\\1\,2\,3\,4\end{bmatrix}_E\ \begin{bmatrix}1\,2\,3\,4\\2\,1\,4\,3\end{bmatrix}_{C2}$ $\begin{bmatrix}1\,2\,3\,4\\1\,3\,4\,2\end{bmatrix}_{C3}$	$\begin{bmatrix}1\,2\,3\,4\\2\,1\,3\,4\end{bmatrix}_\sigma$
square_pyramid (diagram: 1, 2, 3, 4, 5)	$\begin{bmatrix}1\,2\,3\,4\,5\\1\,2\,3\,4\,5\end{bmatrix}_E\ \begin{bmatrix}1\,2\,3\,4\,5\\1\,3\,4\,5\,2\end{bmatrix}_{C4}$	$\begin{bmatrix}1\,2\,3\,4\,5\\1\,3\,2\,5\,4\end{bmatrix}_\sigma$
trigonal_bipyramid (diagram: 1, 2, 3, 4, 5)	$\begin{bmatrix}1\,2\,3\,4\,5\\1\,2\,3\,4\,5\end{bmatrix}_E\ \begin{bmatrix}1\,2\,3\,4\,5\\1\,3\,4\,2\,5\end{bmatrix}_{C3}$ $\begin{bmatrix}1\,2\,3\,4\,5\\5\,3\,2\,4\,1\end{bmatrix}_{C2}$	$\begin{bmatrix}1\,2\,3\,4\,5\\1\,3\,2\,4\,5\end{bmatrix}_\sigma$
octahedron (diagram: 1, 2, 3, 4, 5, 6)	$\begin{bmatrix}1\,2\,3\,4\,5\,6\\1\,2\,3\,4\,5\,6\end{bmatrix}_E\ \begin{bmatrix}1\,2\,3\,4\,5\,6\\1\,3\,4\,5\,2\,6\end{bmatrix}_{C4}$ $\begin{bmatrix}1\,2\,3\,4\,5\,6\\6\,3\,2\,5\,4\,1\end{bmatrix}_{C2}\ \begin{bmatrix}1\,2\,3\,4\,5\,6\\5\,6\,2\,1\,4\,3\end{bmatrix}_{C3}$	$\begin{bmatrix}1\,2\,3\,4\,5\,6\\1\,3\,2\,5\,4\,6\end{bmatrix}_\sigma$
cube (diagram: 1, 2, 3, 4, 5, 6, 7, 8)	$\begin{bmatrix}1\,2\,3\,4\,5\,6\,7\,8\\1\,2\,3\,4\,5\,6\,7\,8\end{bmatrix}_E\ \begin{bmatrix}1\,2\,3\,4\,5\,6\,7\,8\\4\,1\,2\,3\,8\,5\,6\,7\end{bmatrix}_{C4}$ $\begin{bmatrix}1\,2\,3\,4\,5\,6\,7\,8\\8\,5\,1\,4\,7\,6\,2\,3\end{bmatrix}_{C3}\ \begin{bmatrix}1\,2\,3\,4\,5\,6\,7\,8\\4\,8\,5\,1\,3\,7\,6\,2\end{bmatrix}_{C2}$	$\begin{bmatrix}1\,2\,3\,4\,5\,6\,7\,8\\2\,1\,4\,3\,6\,5\,8\,7\end{bmatrix}_\sigma$

Figure 9. Archetypal coordination geometries used in stereochemical definition of the MIF data item _stereo_geometry.

_stereo_geometry
_stereo_vertex_id

The CIP stereochemical designators (R,S,E,Z,r,s,e,z, etc.) are specified with the MIF data items _atom_cip and _bond_cip. For (+)-3-bromocamphor, the absolute configuration is expressed as the atom CIP values R, R, and S for nodes 1, 3, and 4. The MIF atom property data for this molecule are shown in Figure 8. The period in this example is again used to indicate a null field.

The stereogenic center of each stereo group in a molecule has a relationship within that group defined by _define_

stereo_relationship. A description of the standard codes for _define_stereo_relationship are as follows:

absolute: The configuration of all stereogenic centers is exactly as described. This represents an enantiomerically pure compound with a known absolute configuration.

relative: The configuration of the stereogenic centers is only relative, and the mirror reflection of the centers will also describe the same molecule. Only the configuration described in the MIF, or its mirror image, will be present in the molecule. This represents an enantiomerically pure compound with the described relative configuration.
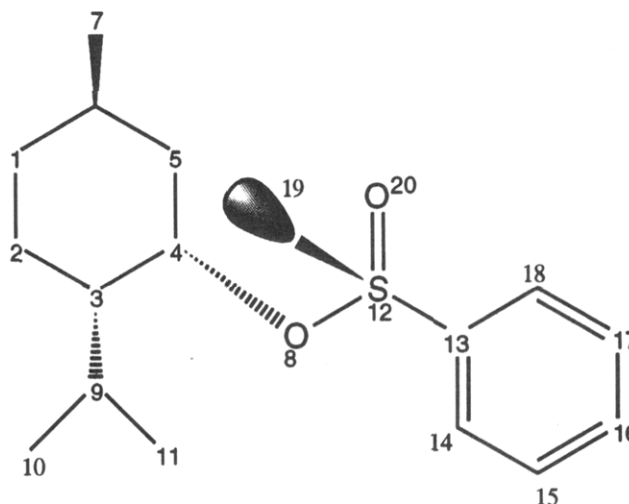
```
data_menthyl_p_toluenesulphinate

_molecule_name_common      menthyl-p-toluenesulphinate

_molecule_name_iupac       "(1R,2S,5R)-(-)-menthyl (S)-p-toluenesulphinate"

loop_
    _atom_id
    _atom_type
    _atom_attach_h
    _atom_cip
           1 C 2 .      2 C 2 .      3 C 1 S      4 C 1 R      5 C 2 .
           6 C 1 R      7 C 3 .      8 O 0 .      9 C 1 .     10 C 3 .
          11 C 3 .     12 S 0 S     13 C 0 .     14 C 1 .     15 C 1 .
          16 C 1 .     17 C 1 .     18 C 1 .     19 usp 0     20 O 0 .
loop_
    _bond_id_1
    _bond_id_2
    _bond_type_mif
           1  2 S   2  3 S   3  4 S   4  5 S   5  6 S   6  7 S   4  8 S
           3  9 S   9 10 S   9 11 S   8 12 S  12 13 S  12 19 S  12 20 D
          13 14 S  14 15 D  15 16 S  16 17 D  17 18 S  13 18 D

_define_stereo_relationship           absolute

loop_
    _stereo_atom_id
    _stereo_geometry
    loop_
        _stereo_vertex_id
                6  tetrahedron     7  5  1  .  stop_
                3  tetrahedron     .  2  4  9  stop_
                4  tetrahedron     8  .  3  5  stop_
               12  tetrahedron    19 20 13  8  stop_
```

Figure 10. Stereochemical data for menthyl-*p*-toluenesulfinate.

racemic: The configuration of the stereogenic centers is only relative, and the mirror reflection of the centers will also describe the same molecule. Both this configuration and its mirror image are present in a 1:1 ratio. This represents a racemic mixture of the molecule with the described relative configuration.

absolute_excess: The configuration of the stereogenic centers describe the absolute configuration of the excess component of a mixture of this configuration and its mirror reflection. This describes an enantiomeric excess in which the excess component has the described absolute configuration.

relative_excess: The configuration of the stereogenic centers is only relative. A mixture of this configuration and its mirror image is present, with one or the other of the components in excess. This describes an enantiomeric excess mixture.

unknown: The configurational relationship between the stereogenic centers is not known.

The geometry of each stereogenic center is described individually in terms of a prototype geometrical model; the basic principles of this approach have been described elsewhere.[17] The eight geometries currently defined for the MIF data item **_stereo_geometry** are illustrated in Figure 9. They include the organic stereogenic geometries: the tetrahedron, the rectangular description of olefin-related compounds, and the antirectangle used to describe allene related systems as well as the common archetypal metal coordination geometries: square planar, tetrahedral, trigonal bipyramidal, square pyramidal, octahedral, and cubic. This list is nonexclusive and can be extended in later versions of the MIF dictionary.

The vertex site of the geometrical model must be occupied by either an atom, an explicit or implicit hydrogen, or by an
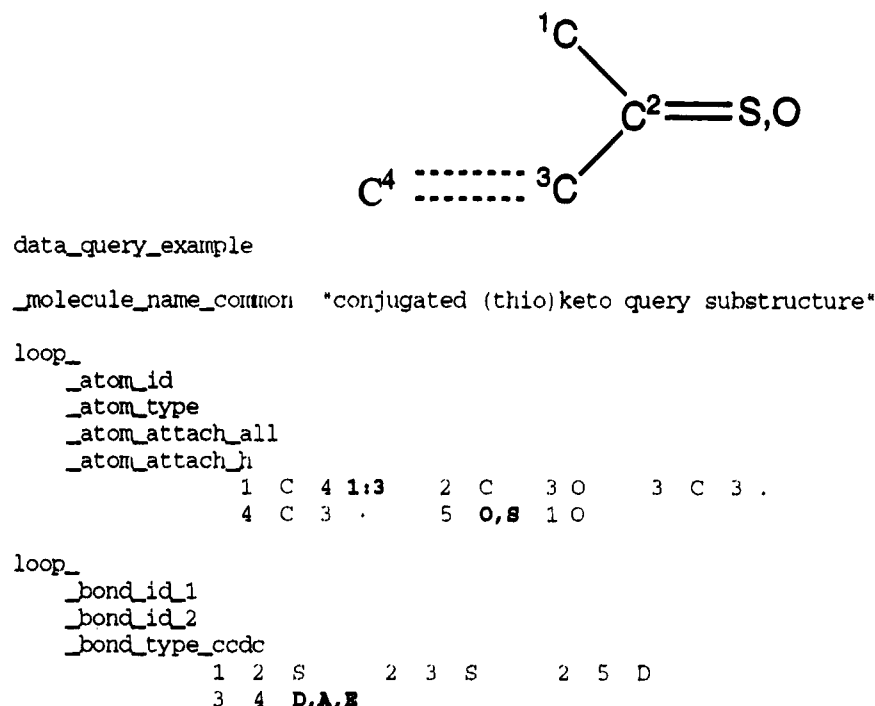
```
data_query_example

_molecule_name_common   "conjugated (thio)keto query substructure"

loop_
    _atom_id
    _atom_type
    _atom_attach_all
    _atom_attach_h
                    1  C  4  1:3     2  C   3 0      3  C  3 .
                    4  C  3  .       5  O,S  1 0

loop_
    _bond_id_1
    _bond_id_2
    _bond_type_ccdc
                1  2  S        2  3  S        2  5  D
                3  4  D,A,E
```

Figure 11. Query substructure for conjugated ketones or thioketones. Atom C_1 is sp^3 hybridized (total number of attached hydrogen and non-hydrogen atoms $= 4$) and carries at least one H atom. Bond $C_3=C_4$ may be localized double (D), aromatic (A), or delocalized double (E) in CCDC conventions.

explicitly declared electron pair. In each case, there exist permutations of the enumerated vertices which, if applied, do not change the meaning of the description of the relevant stereo element. Thus, the MIF does not define a canonical ordering for citing geometric vertices: the comparison of two geometries should be performed by applying the permutation operators. These permutations are also indicated in Figure 9. For each stereogenic center (defined by a **stereo_atom_id**, or by **stereo_bond_id_1** and **_2**), the atom sites forming the stereochemical element specified by a **_stereo_geometry** code are stored as a sequence of **_stereo_vertex_id** values. An example of the specification of absolute stereochemistry, including the ordered enumeration of the tetrahedral vertices for the four stereogenic centers, is given in Figure 10. In this example, the *null* symbol (a period) is used to indicate an implicit hydrogen atom or an unshared electron pair.

MIF QUERY APPLICATIONS

A MIF is ideally suited to interrogating data bases because data items are permitted to have a single value, or a "sequence" of alternative values. This latter option is designated by the dictionary attribute **_type_conditions** which, for MIF applications, is set to "sequenced data" (via the code "seq") . This permits a value string to contain alternative "values" satisfying the following constructs: (a) the value string *v1,v2,v3* signals that a data item must have the value *v1* or *v2* or *v3* and (b) the value string *v1:v2* signals that a data item must have a value in the *range v1* to *v2*. Combinations of these constructions are permitted. All values must comply with the requirements defined by the attributes **_enumeration** and **_enumeration_range**.

An example of a substructural query in a MIF is shown in Figure 11 for a *conjugated ketone* or *thioketone* fragment. Points of permitted variability of atom properties occur at atom 1, an sp^3 carbon that must have at least one attached

H atom, and at atom 5, which can be S or O. The conjugated multiple $C-C$ bond $(3-4)$ is defined to be either localized double, delocalized double, or aromatic using CCDC bonding conventions. Query coding of this type should be readily generable from most graphical 2D search interfaces or be readable directly by a variety of 2D substructure search programs.

CONCLUSION

The current deluge of machine-readable information and of chemical applications software has generated a critical need for a universal approach to the storage and exchange of electronic data. Most existing protocols are either fixed format or relational and have a limited lifetime because they are targeted at specific applications and are relatively inflexible. Some formats that may be considered to offer universality[3,4] are, in our opinion, either too complex for routine applications by chemists or lack the ready extensibility that will eventually be required. Human, as well as computing, efficiencies are vitally important for future interchange developments, and the attributes of simplicity, readability, portability, generality, and extensibility effectively summarize the criteria for any universal format.

The Molecular Information File based on the STAR File syntax, described in this paper, fulfills all of the criteria for universality. Also of key importance to future applications is the ability of the MIF, via the DDL dictionary, to specify data attributes at a definition level that is appropriate to object oriented data handling.

In this paper, we outline the basic MIF approach and provide definitions for an initial core of data items that are fundamental for the representation of 2D and 3D chemical structures and 2D substructures. These core data items cover most of the basic data exchange requirements of molecular modeling and database applications. However, this is only a starting point for chemical data exchange. Future MIF

developments on applications software and, particularly, on data definitions are expected to encompass many areas of chemistry. Obvious MIF extensions would include, *inter alia*, data definitions covering reactions, 3D queries, and generic (Markush) structures. Further developments will, of necessity, require the collaborative involvement and support of appropriate subject specialists from both academia and industry.

Researchers in chemical informatics or chemical computing who would like to be involved in the development of MIF data and access software should contact either Syd Hall (*syd@crystal.uwa.edu.au*) or Frank Allen (*fha1@chemcrys.cam.ac.uk*). Copies of the latest electronic dictionaries can be obtained by anonymous FTP from the Internet address 130.95.232.12 or by e-mail from *bm@iucr.ac.uk* or *syd@crystal.uwa.edu.au*.

ACKNOWLEDGMENT

REFERENCES AND NOTES

(1) (a) Bebak, H.; Buse, C.; Donner, W. T.; Hoever, P.; Jacob, H.; Klaus, H.; Pesch, J.; Roemelt, J.; Schilling, P.; Woost, B.; Zirz, C. The Standard Molecular Data Format (SMD Format) as an Integration Tool in Computer Chemistry. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 1−5. (b) Barnard, J. M. Draft Specification for Revised Version of the Standard Molecular Data (SMD) Format. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 81−96.

(2) Barnard, J. M.; Cook, A. P. F. The Molecular Information File (MIF): A Standard Format for Molecular Information. Report to the Chemical Structure Association (U.K.), December 1992.

(3) Abstract Standard Notation 1. ISO Standards, ISO/IEC 8824 and ISO/IEC 8825, 1990.

(4) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats used by Computer Programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.

(5) Dolata, D. P.; Leach, A. R.; Prout, K. WIZARD: AI in Conformational Analysis. *J. Computer-Aided Mol. Des.* **1987**, *1*, 73−85.

(6) Rusinko, A., III; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. CONCORD, A Program for the Rapid Generation of High-Quality Approximate 3-Dimensional Molecular Structures. The University of Texas at Austin and Tripos Associates, St. Louis, MO, 1988.

(7) Hall, S. R. The STAR File: A New Format for Electronic Data Transfer and Archiving. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 326−333.

(8) Hall, S. R.; Spadaccini, N. The STAR File: Detailed Specifications. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 505−508.

(9) Hall, S. R.; Allen, F. H.; Brown, I. D. The Crystallographic Information File (CIF): a New Standard Archive File for Crystallography. *Acta Crystallogr.* **1991**, *A47*, 655−685.

(10) Brown, I. D. The Standard Crystallographic File Structure (SCFS). *Acta Crystallogr.* **1988**, *A44*, 232−233.

(11) Standard Specification for the Content of Computerized Chemical Structural Information Files or Data Sets; ASTM Standard E 1586-93; American Society for Testing Materials: Philadelphia, PA, 1994.

(12) Hall, S. R.; Cook, A. P. F. Data Definition Language for STAR File Dictionaries. *J. Chem. Inf. Comput. Sci.* Submitted for publication.

(13) Allen, F. H.; Rogers, D. X-ray Studies of Terpenoids. Part III. A Redetermination of the Crystal Structure of (+)-3-Bromocamphor. *J. Chem. Soc. B.* **1970**, 632−656.

(14) Mockus, J.; Stobaugh, R. E.; The Chemical Abstracts Registry System VII. Tautomerism and Alternating Bonds. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 18−22.

(15) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C.F.; Mitchell, E. M.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. The Development of Versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187−204.

(16) (a) Cahn, R. S.; Ingold, C. K.; Prelog, V. Specification of Molecular Chirality. *Angew. Chem., Int. Ed. Engl.* **1966**, *5*, 385−415. (b) Prelog, V.; Helmchen, G. Basic Principles of the CIP-System and Proposals for a Revision. *Angew. Chem., Int. Ed. Engl.* **1982**, *21*, 567−583.

(17) Barnard, J. M.; Cook, A. P. F.; Rohde, B. Storage and Searching of Stereochemistry in Substructure Search Systems. In *Beyond the Structural Diagram*; Bawden, D.; Mitchell, E.; Ellis Horwood: Chichester, UK, 1990; pp 29−41.

CI940121G