International Tables for Crystallography (2006). Vol. G, Chapter 2.4, pp. 44–52.

2.4. Specification of the Molecular Information File (MIF)

By F. H. Allen, J. M. Barnard, A. P. F. Cook and S. R. Hall

2.4.1. Introduction

This volume is primarily concerned with methods for the exchange of crystallographic information, such as experimental conditions and measurements, computational procedures and results, and the geometrical description of three-dimensional (3D) chemical structures. Such information, now available for some 400 000 compounds (Allen & Glusker, 2002) is, of course, vitally important in chemistry and in many other branches of science. However, it must be appreciated that two-dimensional (2D) chemical structural diagrams are available for over eight million compounds, and are fundamental components of the language of chemistry at all levels. Two-dimensional graphical representations indicate atomic connectivities, formal bond types and residual atomic charges, and provide the universal formalism through which chemists communicate with each other on a daily basis and document their results.

In common with scientists in other disciplines, chemists were early users of computer technology. They solved their information needs through the creation of major databases of chemical compounds and the development of methods for searching these databases for complete structures or substructural fragments. From these data, software can compute the 3D structures and properties of molecules, and the resulting molecular images can be displayed and manipulated. Consequently, 2D chemical diagrams are at the heart of many computerized documentation systems, and are the basis for computational chemistry applications that form part of the routine armoury of the modern chemist.

Computationally, the 2D diagram is treated as a mathematical graph (Harary, 1972). The nodes of the graph represent atoms and the edges of the graph represent bonds. Each of these primary components can have additional attributes: element type, valency, charge *etc.* in the case of the atomic nodes, and bond type, cyclicity indicators *etc.* in the case of the bonded edges. Within this formalism, 2D display coordinates and 3D crystallographic or computed coordinates are additional atom attributes, while interatomic distances can further qualify the bonded edges. Using the concepts of graph theory, it is then possible to write algorithms for the analysis of chemical graphs, *e.g.* for the detection of chemical rings and ring systems (*e.g.* Wippke & Dyott, 1975), for the analysis of functional groups and their relationships *etc.* Most importantly, procedures have also been developed for the matching of complete chemical graphs (full graph isomorphism), and for the location of chemical substructures within a complete chemical graph (sub-graph isomorphism) (*e.g.* Feldmann *et al.*, 1977). In this way it is possible to achieve graphical substructure searches of very large collections of 2D chemical diagrams.

As with early crystallographic data exchange, structural chemistry applications use their own specialized formats for input, manipulation and output. The ready exchange of chemical data is often inhibited by specific data formats and by the enormous variation in methods used to represent 2D structures, stereochemical descriptors and certain 3D structural attributes. These are computational 'bottlenecks' that detract from an effective use of the large financial and intellectual investment in proprietary software and database systems. They have also contributed to the major need for in-house format conversion software, which must be continually upgraded and maintained to accommodate developmental changes within imported systems.

The need for a universal interchange format for chemical information became apparent in the late 1980s, at almost exactly the same time as crystallographers recognized a similar need. Data standards in structural chemistry involve many international organizations and individuals, and consequently a number of proposals for exchanging data were initially put forward. From these the Standard Molecular Data (SMD) format (Bebak *et al.*, 1989; Barnard, 1990) emerged as the leading contender. In the early 1990s, discussions between the SMD and CIF developers led to a re-expression of the SMD data items within the Self-defining Text Archive and Retrieval (STAR) File syntax (Hall, 1991; Hall & Spadaccini, 1994).

This chapter describes the initial core data definitions of a universal exchange format for chemistry, the Molecular Information File (MIF: Allen *et al.*, 1995), that arose from this coalescence of concepts and ideas. MIF is a complementary approach to CIF (Hall *et al.*, 1991). Because SMD was fundamental to the development of MIF, we begin this chapter with a brief history of this project.

2.4.2. Historical background

The Standard Molecular Data (SMD) format was initially developed by a group of European pharmaceutical companies in the mid-1980s. Draft documents were made available from 1987 and the specification was published (Bebak *et al.*, 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 form of a revised format, described as SMD Version 5.0, was published in February 1990 (Barnard, 1990). A document describing the core format, *i.e.* those data items regarded as essential in any exchange file, was prepared by one of us (JMB) for consideration by Subcommittee E49.51 of the American Society for Testing and Materials (ASTM).

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 (ASTM, 1994), although the subcommittee did not publish any proposals for a format specification. Later, the *Chemical Abstracts* Service (CAS) circulated a draft proposal for a connection-table-based exchange format for chemical substances and queries. It used some ideas that are similar to the 1990 SMD proposal and is expressed within the framework of the Abstract Syntax Notation 1 (ISO, 2002*a*,*b*). MDL Information Systems Inc. has also published a description of their proprietary formats (Dalby *et al.*, 1992) and a number of other software systems now provide interfaces to these formats.

Affiliations: Frank H. Allen, Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, England; John M. Barnard and Anthony P. F. Cook, BCI Ltd, 46 Uppergate Road, Stannington, Sheffield S6 6BX, England; Sydney R. Hall, School of Biomedical and Chemical Sciences, University of Western Australia, Crawley, Perth, WA 6009, Australia.

During this period, the IUCr Working Party on Crystallographic Information had commissioned one of us (SRH) to coordinate the development of a universal file to replace the existing fixed-format Standard Crystallographic File Structure (SCFS: Brown, 1988). As documented in Chapter 1.1, the CIF approach was adopted as the international standard in 1990 and published by Hall *et al.* (1991). Although the small-molecule CIF is able to store a representation of 2D chemical topology, its data definitions do not meet all the needs of the chemical community. In 1991, the IUCr became interested in further extending CIF into the chemical arena and discussions took place between representatives of the CIF project and of the SMD Technical Working Group. 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 (Barnard & Cook, 1992). In particular, SMD/STAR provides for a clearer separation of the data structure and the data content 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 integrating with other applications of the STAR File.

2.4.3. MIF objectives

Molecular information embraces a 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 items represent the data descriptors of molecular chemistry and it is intended that all of these will eventually be accommodated in the MIF approach. However, the initial MIF implementation (Allen *et al.*, 1995), summarized in this chapter, treated only the important core information: the data items needed to specify the connectivity and stereochemistry of molecules 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 Chapter 4.8. This is the abbreviated text version of the definition attributes contained in the electronic dictionary file. The core MIF data items provide descriptors for representing the 2D connectivity of a molecule or substructure, the conventions for relative or absolute stereochemical relationships, and the coordinates and conventions used for the generation of 2D and 3D graphical depictions. These data items apply to 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 this chapter.

2.4.4. MIF concepts and syntax

The syntax of the Molecular Information File is based on that of the STAR File (Hall, 1991; Hall & Spadaccini, 1994). A MIF is an ASCII text file that can be read or amended using a standard text editor, and that can be processed computationally without conversion to another format. The organization and expression of MIF data is summarized in Table 2.4.4.1. Each file consists of a series of data blocks and each block consists of a series of 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.

Table 2.4.4.1. *Brief overview of the MIF syntax*

A text string is a string of characters bounded by white space, single or double quotes, or semicolons in column 1.
A data name is a text string bounded by white space starting with an underline.
A data value 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 values.
A save frame is a collection of data within a data block, preceded by '`save_`*`framecode`*' and closed with '`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.

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 compare this chapter with Chapter 2.2 or refer to the published details of the STAR syntax (Hall & Spadaccini, 1994), the specification of the CIF core data items (Hall *et al.*, 1991) and the Dictionary Definition Language (Hall & Cook, 1995) used to define data items in the electronic version of a STAR dictionary.

CIF data, described by over a thousand items in the current dictionaries (see Part 4), encompass the fields of crystallographic structure and diffraction techniques, and these data items could readily be incorporated into a MIF. It should be noted, however, that currently the reverse is not possible because the current CIF syntax does not support nested loops or save frames.

2.4.4.1. Data identification

The fundamental principle that underpins MIF is exactly as for CIF: every data item is represented by a unique data tag followed by its associated data value. These combinations are referred to as tag–value pairs or tuples. Data names must start with an underscore (*i.e.* underline) character and data values may be any type of string, ranging from a single character to many lines of text. Here are some simple examples of MIF data items:

```
_atom_mass_number       79
_atom_type              Se
_display_colour         blue_medium
```

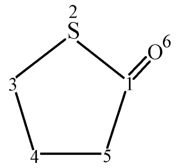
The complete list of MIF core data items is given in Chapter 4.8.

2.4.4.2. Looped lists

Repetitive data are stored in a MIF as lists of values, as they are in a CIF. Each list is prefaced by a `loop_` statement and a sequence of data names that identify the data values that follow in 'packets' of equal length. The values in each packet match the order and number of the data names. Any number of packets may appear in a looped list.

Atom and bond properties are typical of the information to appear in a looped list. The atoms and bonds of thiabutyrolactone in MIF format are shown in Fig. 2.4.4.1. The description of each data item in this example is given in Chapter 4.8, although the meanings are clear from the self-descriptive data names. 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 by any other data name, data block or end of file. Comments may be included in a MIF and are preceded by a `#` character, as illustrated in Fig. 2.4.4.1.

Hierarchical data may require the use of nested loop structures (see the `_display_*` loop in Fig. 2.4.4.2). Note that the packet for `_display_id` of 7 has two sets of `_display_conn_` values giving



```
# 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
```

Fig. 2.4.4.1. MIF coding of atom and bond properties for thiabutyrolactone.

connections to atom sites 1 and 4 (the other connections to site 7 appear in the next two packets). Data items that appear in looped lists are identified in the MIF dictionary (see Chapter 4.8) as having the attribute `_list` set to either ‘yes’ or ‘both’. Other relationships between looped data items are also specified in the dictionary.

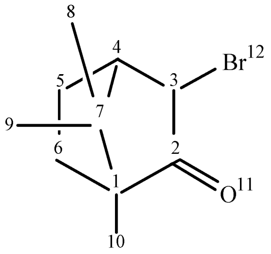
2.4.4.3. Save frames

Save frames are employed in a MIF to encapsulate grouped data for efficient cross-referencing. If a set of data needs to appear repeatedly in a data application, it is efficient to place this data into an addressable save frame. Molecular fragments, such as amino-acid units, are a case in point. A save frame is bounded by the statement `save_`***`framecode`*** and terminated by a `save_` statement. It can be referenced within the parent data block using the value `$`***`framecode`*** where the *framecode* matches the string in the `save_`***`framecode`***. Note that all data names must be unique within a save frame, but the same data names may appear in other save frames or in the parent data block. Save frames may not contain other save frames but save-frame references (`$`***`framecode`***) may appear in other save frames.

Save frames can be used in a MIF for many purposes. A simple application, the storage of alternative 3D conformational representations describing cyclohexane, is illustrated in Fig. 2.4.4.3. Within the STAR syntax, save-frame references (`$`***`framecode`***) may occur before or after the save-frame definition within any data block. MIF preserves this basic STAR syntax. Save frames are particularly useful for defining commonly referenced structural templates and examples of this facility are discussed and illustrated (Figs. 2.4.7.1 and 2.4.7.2) in Section 2.4.7.

2.4.4.4. Data blocks

A data block is a sequence of unique data items or save frames. It is opened with a `data_`***`blockcode`*** statement and closed by another data-block statement or a `global_` statement (see below). The *blockcode* string identifies the block within the file. Examples of data blocks are shown in Figs. 2.4.4.2, 2.4.4.3 and 2.4.6.1. Each data block in a file must have a unique *blockcode*.



```
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_define_scale       50
_display_define_span_x     500
_display_define_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_
```

Fig. 2.4.4.2. MIF coding of atom properties (including 3D coordinates), bond properties and display information for (+)-3-bromocamphor.

```
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_id
        _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_
```

Fig. 2.4.4.3. Atom and bond properties for cyclohexane, together with 3D coordinate representations of three alternative conformations: chair, boat and twisted boat.

2.4.4.5. Global blocks

A global block is similar to a data block except that it is opened with a `global_` statement and contains data that are common or 'default' to all subsequent data blocks in a file. Global data items remain active until re-specified in a subsequent data block or global block.

In some applications it may be efficient to place data that are 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 corrects an error in Hall & Spadaccini (1994)]. Examples of global data are shown in Figs. 2.4.7.1 and 2.4.7.2, in which a variety of frequently referenced structural units are encapsulated within save frames specified in global blocks.

2.4.5. Atoms, bonds and molecular representations

The MIF dictionary (see Chapter 4.8) contains definitions of the principal data items needed to specify molecular connectivity and spatial representations. These definitions are grouped according to purpose or, as referred to in the DDL dictionary language (Hall & Cook, 1995), by category. Categories are formally specified in the MIF dictionary using the data attribute `_category` but they may also be identified from the data-name construction '*_<category>_<subcategory>_<descriptor>*'. Note that data items appearing in the same looped list must belong to the same category.

The values of some data items are restricted, by definition in the MIF dictionary, to standard codes or states. For example, the item `_bond_type_mif` can only have values `S`, `D`, `T` or `O` as in its dictionary definitions:

`S`: single (two-electron) bond;
`D`: double (four-electron) bond;
`T`: triple (six-electron) bond;
`O`: other (*e.g.* coordination) bond.

The MIF dictionary plays the important additional role of validating and standardizing data values. This is illustrated with the data item `_display_colour`, which identifies the colours of 'atom' and 'bond' graphical objects. The colour codes or states for this item are specified in its dictionary definitions as a set of permitted red/green/blue (RGB) ratios, and no other colours may be used in a MIF. This has the technical advantage of making colour states searchable for chemical applications.

Fig. 2.4.4.2 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 in Cartesian coordinates [these are taken from diffraction results (Allen & Rogers, 1970)]. 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 above. The data names needed to depict stereochemistry are discussed with examples (Figs. 2.4.8.1, 2.4.8.2 and 2.4.8.3) in Section 2.4.8.

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. The MIF data required to generate a 2D chemical diagram are shown in Fig. 2.4.4.2. The diagram generated from this data will be in a display area of 500 × 500 coordinate units at a scale of 50 units per cm (the 2D chemical diagram shown in Fig. 2.4.4.2 is not to this scale). The default origin (the bottom left corner of the display area) can be specified with the item `_display_define_origin`. The data used to depict a 2D structure form 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 colour of the atom site are 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 definition, as is the colour of the icon if specified by `_display_conn_colour`.

2.4.6. Bonding conventions

Chemical information systems use a variety of conventions to specify 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-atom 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 hydrogen atoms and the bond types associated with a given atom are clearly related. Operational systems make use of these relationships to perform internal checks and to provide flexibility in substructure search processes.

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, while the 'casreg3' convention (Mockus & Stobaugh, 1980) extends these to include aromaticity in terms of 'ring alternating normalized bonds' and tautomerism *via* a 'tautomer normalized bond'. The 'ccdc' convention is that employed in the Cambridge Structural Database System (Allen *et al.*, 1991; Allen, 2002) to categorize bond types encountered in both organic and metal-organic molecules.

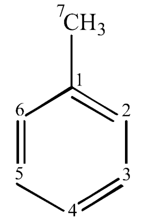
An important advantage of the MIF approach is that a molecule can be represented using all three bonding conventions within the same data block. An example of alternative bonding conventions encoded for toluene is shown in Fig. 2.4.6.1.

2.4.7. 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, molecular templates can be encapsulated as save frames, either within a data block for a specific molecule, or within a global block that is accessible to many data blocks. A simple application of a MIF template is shown in Fig. 2.4.7.1, where a 4-methylcyclohexyl ligand is used to encode the molecule tris(methylcyclohexyl)phosphine. In this example a molecular fragment is constructed in the save frame mechex, where the 'atom' sites and 'bond' connections appear in `_atom_*` and `_bond_*` loops. The molecule (2-methylcyclohexyl)(3-methylcyclohexyl)(4-methylcyclohexyl)phosphine is encoded by referencing the template fragment as the save frame $mechex. In the 'atom' loop, the item `_atom_environment` identifies the components of the target molecule as an 'atom' or 'frag' (fragment). If the component is a fragment, the items `_atom_frag_key` and `_atom_frag_id` are used to specify the frame code and the ID of the attached atom in the fragment, respectively. In the 'bond' loop, the connections from the atom P(1) to the template are encoded simply in terms of the `_atom_id` values. The necessary redefinition of the hydrogen and non-hydrogen counts of the template atoms is accomplished using the `_atom_attach_h` and `_atom_attach_nh` items, respectively. 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 in Fig. 2.4.7.2. This employs the template peptide units described by the atoms and bonds in the save frames $alanyl and $seryl. The complete dipeptide is specified in its 'atom' list as the template peptides (identified by their save-frame names) and an additional carboxylate O atom. Note that only the atom sites affected by molecule formation are identified explicitly in this list, which gives the values of `_atom_attach_nh`, `_atom_attach_h` and `_atom_charge` for the modified sites in the zwitterionic form of alanylserine.



```
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
```

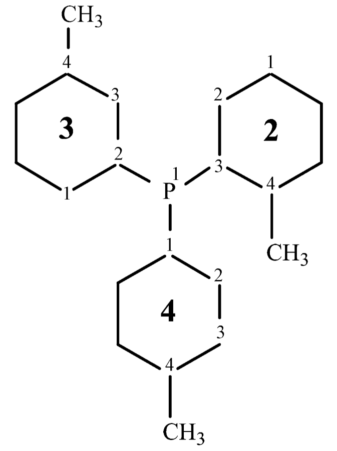
Fig. 2.4.6.1. Three alternative bonding conventions for toluene stored in the same MIF data block.

2.4.8. Stereochemistry and geometry at stereogenic centres

The Cahn–Ingold–Prelog (CIP) notation (Cahn *et al.*, 1966; Prelog & Helmchen, 1982) is available in the MIF definitions to specify the stereochemistry of a molecule. The CIP notation is restricted to tetrahedral atomic centres and to olefinic type stereogenic bonds, and the CIF approach is unsuitable for describing molecules with partially known stereochemistry, molecules containing more complex geometries or substructural queries. The MIF data items representing stereochemical quantities are as follows:

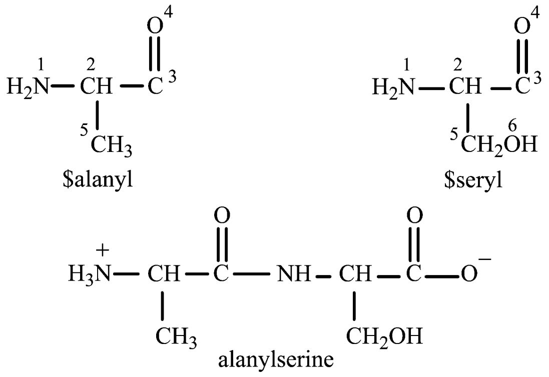
```
_define_stereo_relationship
_atom_cip
_bond_cip
_stereo_atom_id
_stereo_bond_id_1
_stereo_bond_id_2
_stereo_geometry
_stereo_vertex_id
```

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`. The MIF atom-property data for the molecule (+)-3-bromocamphor are shown in Fig. 2.4.8.1. In this the absolute configuration is expressed as the atom CIP values R, R and S for nodes 1, 3 and 4. The period in this example is used to indicate a null field.



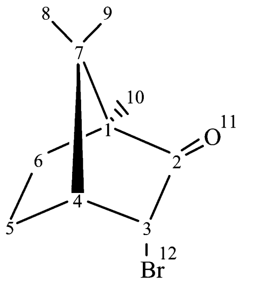
```
global_

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_environment
         _atom_frag_key
         _atom_frag_id
         _atom_type
         _atom_attach_nh
         _atom_attach_h
            1   atom   .         .  P  3  0
            2   frag   $mechex  3  C  3  1
            3   frag   $mechex  2  C  3  1
            4   frag   $mechex  1  C  3  1
     loop_
         _bond_id_1
         _bond_id_2
         _bond_type_mif
            1  2  S      1  3  S      1  4  S
```

Fig. 2.4.7.1. 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.



```
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_environment
         _atom_frag_key
         _atom_frag_id
         _atom_type
         _atom_attach_nh
         _atom_attach_h
         _atom_charge
               1  frag   $alanyl   1   N   1   3  +1
               2  frag   $alanyl   3   C   3   0   0
               3  frag   $seryl    1   N   2   1   0
               4  frag   $seryl    3   C   3   0   0
               5  atom   .         .   O   1   0  -1
     loop_
         _bond_id_1
         _bond_id_2
         _bond_type_mif
               2 3 S    4 5 S
```

Fig. 2.4.7.2. MIF representation of the dipeptide alanylserine constructed using alanyl and seryl templates encapsulated in global save frames.



```
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  .
```

Fig. 2.4.8.1. CIP stereochemical descriptors for (+)-3-bromocamphor.

The stereogenic centre of a stereo group in a molecule has a relationship within that group that is specified by **`_define_stereo_relationship`**. Descriptions of the standard codes for **`_define_stereo_relationship`** are as follows.

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

`relative`: The configuration of the stereogenic centres is only relative and the mirror reflection of the centres 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.

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

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

The geometry of each stereogenic centre is described individually in terms of a prototype geometrical model. The basic principles of this approach have been described elsewhere (Barnard *et al.*, 1990). The eight geometries currently defined for the MIF data item **`_stereo_geometry`** are given in Fig. 2.4.8.2. They include the organic stereogenic geometries (the tetrahedron, the rectangular description of olefin-related compounds and the anti-rectangle used to describe allene-related systems) and the common archetypal metal coordination geometries (square planar, tetrahedral, trigonal bipyramidal, square pyramidal, octahedral and cubic). This list is non-exclusive and can be extended as required in later versions of the MIF dictionary.

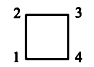
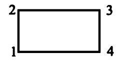
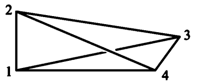
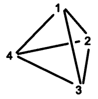
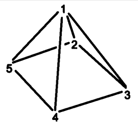
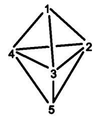
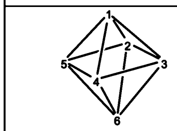
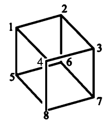
Geometry	Proper rotations	Reflection (chiral)
	square $\begin{bmatrix}1&2&3&4\\1&2&3&4\end{bmatrix}_{E}\begin{bmatrix}1&2&3&4\\2&1&4&3\end{bmatrix}_{C2}\begin{bmatrix}1&2&3&4\\4&1&2&3\end{bmatrix}_{C4}$	
	olefin $\begin{bmatrix}1&2&3&4\\1&2&3&4\end{bmatrix}_{E}\begin{bmatrix}1&2&3&4\\2&1&4&3\end{bmatrix}_{C2}\begin{bmatrix}1&2&3&4\\4&3&2&1\end{bmatrix}_{C2}$	
	allene $\begin{bmatrix}1&2&3&4\\1&2&3&4\end{bmatrix}_{E}\begin{bmatrix}1&2&3&4\\2&1&4&3\end{bmatrix}_{C2}$ $\begin{bmatrix}1&2&3&4\\4&3&2&1\end{bmatrix}_{C2}$	$\begin{bmatrix}1&2&3&4\\2&1&3&4\end{bmatrix}_{\sigma}$
	tetrahedron $\begin{bmatrix}1&2&3&4\\1&2&3&4\end{bmatrix}_{E}\begin{bmatrix}1&2&3&4\\2&1&4&3\end{bmatrix}_{C2}$ $\begin{bmatrix}1&2&3&4\\1&3&4&2\end{bmatrix}_{C3}$	$\begin{bmatrix}1&2&3&4\\2&1&3&4\end{bmatrix}_{\sigma}$
	square_pyramid $\begin{bmatrix}1&2&3&4&5\\1&2&3&4&5\end{bmatrix}_{E}\begin{bmatrix}1&2&3&4&5\\1&3&4&5&2\end{bmatrix}_{C4}$	$\begin{bmatrix}1&2&3&4&5\\1&3&2&5&4\end{bmatrix}_{\sigma}$
	trigonal_bipyramid $\begin{bmatrix}1&2&3&4&5\\1&2&3&4&5\end{bmatrix}_{E}\begin{bmatrix}1&2&3&4&5\\1&3&4&2&5\end{bmatrix}_{C3}$ $\begin{bmatrix}1&2&3&4&5\\5&3&2&4&1\end{bmatrix}_{C2}$	$\begin{bmatrix}1&2&3&4&5\\1&3&2&4&5\end{bmatrix}_{\sigma}$
	octahedron $\begin{bmatrix}1&2&3&4&5&6\\1&2&3&4&5&6\end{bmatrix}_{E}\begin{bmatrix}1&2&3&4&5&6\\1&3&4&5&2&6\end{bmatrix}_{C4}$ $\begin{bmatrix}1&2&3&4&5&6\\6&3&2&5&4&1\end{bmatrix}_{C2}\begin{bmatrix}1&2&3&4&5&6\\5&6&2&1&4&3\end{bmatrix}_{C3}$	$\begin{bmatrix}1&2&3&4&5&6\\1&3&2&5&4&6\end{bmatrix}_{\sigma}$
	cube $\begin{bmatrix}1&2&3&4&5&6&7&8\\1&2&3&4&5&6&7&8\end{bmatrix}_{E}\begin{bmatrix}1&2&3&4&5&6&7&8\\4&1&2&3&8&5&6&7\end{bmatrix}_{C4}$ $\begin{bmatrix}1&2&3&4&5&6&7&8\\8&5&1&4&7&6&2&3\end{bmatrix}_{C3}\begin{bmatrix}1&2&3&4&5&6&7&8\\4&8&5&1&3&7&6&2\end{bmatrix}_{C2}$	$\begin{bmatrix}1&2&3&4&5&6&7&8\\2&1&4&3&6&5&8&7\end{bmatrix}_{\sigma}$

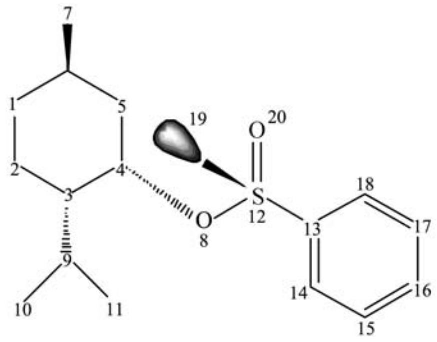
Fig. 2.4.8.2. Archetypal coordination geometries used in stereochemical definition of the MIF data item **`_stereo_geometry`**.

The vertex site of the geometrical model must be occupied by either an atom, an explicit or implicit hydrogen atom, or by an explicitly declared electron pair. In each case, there exist permutations of the enumerated vertices that, 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 and the comparison of two geometries requires the use of the permutation operators. These permutations are also indicated in Fig. 2.4.8.2.

For each stereogenic centre (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 centres, is given in Fig. 2.4.8.3. In this example, the null symbol (a period) is used to indicate an implicit hydrogen atom or an unshared electron pair.

2.4.9. MIF query applications

A MIF is suitable for interrogating databases 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

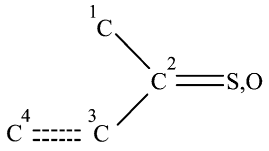


```
data_menthyl_p_toluenesulfinate

_molecule_name_common     menthyl-p-toluenesulfinate

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

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_
```

Fig. 2.4.8.3. Stereochemical data for menthyl-*p*-toluenesulfinate.

'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 Fig. 2.4.9.1 for a conjugated ketone or thioketone fragment. Points of permitted variability of atom properties occur at atom 1, an sp^3 carbon atom that must have at least one attached hydrogen 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 generated from most graphical 2D search interfaces or be readable directly by a variety of 2D substructure search programs.



```
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 O     3  C  3 .
          4  C  3 .      5  O,S  1 O

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

Fig. 2.4.9.1. Query substructure for conjugated ketones or thioketones. Atom C1 is sp^3 hybridized (total number of attached hydrogen and non-hydrogen atoms = 4) and carries at least one hydrogen atom. Bond C3—C4 may be localized double (D), aromatic (A) or delocalized double (E) in CCDC conventions.

2.4.10. Conclusion

The present proliferation of formats for chemical applications tends to inhibit and complicate the exchange and use of chemical data. Many widely used chemical formats have a finite half-life because they are inflexible and not readily extensible. Others offer universality [*e.g.* Abstract Syntax Notation 1 (ISO, 2002*a*,*b*); Dalby *et al.* (1992); see http://www.daylight.com/smiles/] but lack visual simplicity, generality or machine readability. Nevertheless, the Molecular Information File approach has these properties but needs significantly more development to be a viable exchange approach for mainstream chemistry. The MIF dictionary enables chemical data items to be defined at high precision and this offers real benefits for the creation of a domain ontology in this field.

Herein we have outlined the basic MIF approach and provided 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 chemical data-exchange requirements of molecular modelling and database applications, but are clearly only a first step towards the level of chemical data exchange needed. Future MIF developments in applications software and, particularly, in data definitions are expected to encompass other aspects of chemistry. These developments will need the collaborative involvement and support of subject specialists from both academia and industry.

References

Allen, F. H. (2002). *The Cambridge Structural Database: a quarter of a million crystal structures and rising. Acta Cryst.* B**58**, 380–388.

Allen, F. H., Barnard, J. M., Cook, A. P. F. & Hall, S. R. (1995). *The Molecular Information File (MIF): core specifications of a new standard format for chemical data. J. Chem. Inf. Comput. Sci.* **35**, 412–427.

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. (1991). *The development of versions 3 and 4 of the Cambridge Structural Database System. J. Chem. Inf. Comput. Sci.* **31**, 187–204.

Allen, F. H. & Glusker, J. P. (2002). *Preface*. *Acta Cryst.* B**58**, Part 3.

Allen, F. H. & Rogers, D. (1970). *X-ray studies of terpenoid derivatives, Part III. A re-determination of the crystal structure of (+)-3-bromocamphor: the absolute configuration of (+)-camphor. J. Chem. Soc. B*, pp. 632–636.

ASTM (1994). *Standard specification for the content of computerized chemical structural information files or data sets.* ASTM Standard E 1586-93. American Society for Testing and Materials, Philadelphia, PA, USA.

Barnard, J. M. (1990). *Draft specification for revised version of the Standard Molecular Data (SMD) format. J. Chem. Inf. Comput. Sci.* **30**, 81–96.

Barnard, J. M. & Cook, A. P. F. (1992). *The Molecular Information File (MIF): a standard format for molecular information.* Report. Chemical Structure Association, London, England.

Barnard, J. M., Cook, A. P. F. & Rohde, B. (1990). *Beyond the structure diagram*, edited by D. Bawden & E. Mitchell, pp. 29–41. Chichester: Ellis Horwood.

Bebak, H., Buse, C., Donner, W. T., Hoever, P., Jacob, H., Klaus, H., Pesch, J., Roemelt, J., Schilling, P., Woost, B. & Zirz, C. (1989). *The Standard Molecular Data format (SMD format) as an integration tool in computer chemistry. J. Chem. Inf. Comput. Sci.* **29**, 1–5.

Brown, I. D. (1988). *Standard Crystallographic File Structure-87. Acta Cryst.* A**44**, 232.

Cahn, R. S., Ingold, C. K. & Prelog, V. (1966). *Specification of molecular chirality. Angew. Chem. Intl Ed. Engl.* **5**, 385–415.

Dalby, A., Nourse, J. G., Hounshell, W. D., Gushurst, A. K. I., Grier, D. L., Leland, B. A. & Laufer, J. (1992). *Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. J. Chem. Inf. Comput. Sci.* **32**, 244–255.

Feldmann, R. J., Milne, G. W. A., Heller, S. R., Fein, A., Miller, J. A. & Koch, B. (1977). *An interactive substructure search system. J. Chem. Inf. Comput. Sci.* **17**, 157–163.

Hall, S. R. (1991). *The STAR File: a new format for electronic data transfer and archiving. J. Chem. Inf. Comput. Sci.* **31**, 326–333.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *The Crystallographic Information File (CIF): a new standard archive file for crystallography. Acta Cryst.* A**47**, 655–685.

Hall, S. R. & Cook, A. P. F. (1995). *STAR Dictionary Definition Language: initial specification. J. Chem. Inf. Comput. Sci.* **35**, 819–825.

Hall, S. R. & Spadaccini, N. (1994). *The STAR File: detailed specifications. J. Chem. Inf. Comput. Sci.* **34**, 505–508.

Harary, F. (1972). *Graph theory*, 3rd ed. London: Addison-Wesley.

ISO (2002*a*). ISO/IEC 8824-1. *Abstract Syntax Notation One (ASN.1). Specification of basic notation.* Geneva: International Organization for Standardization.

ISO (2002*b*). ISO/IEC 8825-1. *ASN.1 encoding rules. Specification of Basic Encoding Rules (BER), Canonical Encoding Rules (CER) and Distinguished Encoding Rules (DER).* Geneva: International Organization for Standardization.

Mockus, J. & Stobaugh, R. E. (1980). *The Chemical Abstracts Service Chemical Registry System. VII. Tautomerism and alternating bonds. J. Chem. Inf. Comput. Sci.* **20**, 18–22.

Prelog, V. & Helmchen, G. (1982). *Basic principles of the CIP-system and proposals for a revision. Angew. Chem. Intl Ed. Engl.* **21**, 567–583.

Wippke, W. T. & Dyott, T. M. (1975). *Use of ring assemblies in ring perception algorithm. J. Chem. Inf. Comput. Sci.* **15**, 140–147.