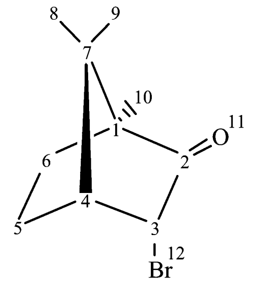
International Tables for Crystallography (2006). Vol. G, Section 2.4.9, pp. 50–51.



```
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.

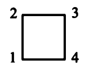
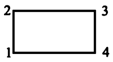
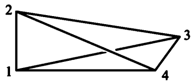
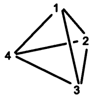
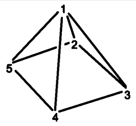
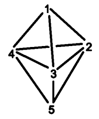
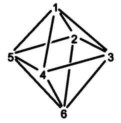
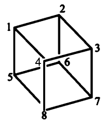
Geometry	Proper rotations	Reflection (chiral)
	square $\begin{bmatrix}1&2&3&4\\1&2&3&4\end{bmatrix}_{E}\begin{bmatrix}1&2&3&4\\2&1&4&3\end{bmatrix}_{C2}\begin{bmatrix}1&2&3&4\\4&1&2&3\end{bmatrix}_{C4}$	
	olefin $\begin{bmatrix}1&2&3&4\\1&2&3&4\end{bmatrix}_{E}\begin{bmatrix}1&2&3&4\\2&1&4&3\end{bmatrix}_{C2}\begin{bmatrix}1&2&3&4\\4&3&2&1\end{bmatrix}_{C2}$	
	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 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.

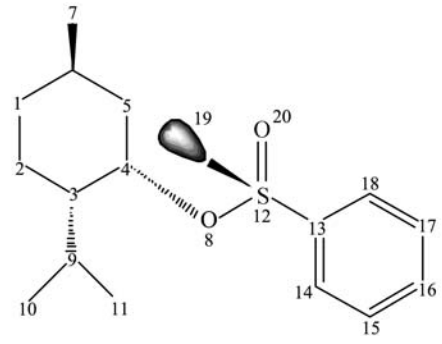
The geometry of each stereogenic centre is described individually in terms of a prototype geometrical model. The basic principles of this approach have been described elsewhere (Barnard *et al.*, 1990). The eight geometries currently defined for the MIF data item `_stereo_geometry` are given in Fig. 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.

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

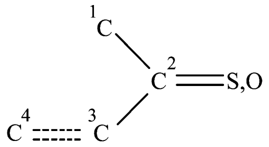


```
data_menthyl_p_toluenesulfinate

_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 *v*1, *v*2, *v*3 signals that a data item must have the value *v*1 or *v*2 or *v*3, and (*b*) the value string *v*1:*v*2 signals that a data item must have a value in the range *v*1 to *v*2. 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.